Bayesian Inference (Contd.)

Rohosen Bandyopadhyay

October 14, 2013

1 Proper/improper prior

- A prior $p(\theta)$ is called a **proper prior** if $\int p(\theta)d\theta < \infty$
- It is called an **improper prior** if $\int p(\theta)d\theta = \infty$
- If we use a proper prior for θ, then the posterior for θ is also proper.
- If we use an improper prior for θ then the posterior may or may not be proper! (So, one needs to prove that the posterior is proper while using an improper prior for θ)
- Small notes:
  1. Since we mainly deal with posterior distribution (e.g. we want to know what is the posterior mean to give a point estimate for θ) and we need prior only to derive the posterior, it is OK to use improper prior, but NOT OK to have an improper posterior.
  2. Prof. Baines' recommendation: Use proper prior. Trying to show that the posterior is proper, while using an improper prior, can be messy!

2 More on priors

- Example: Suppose, $X_i|\mu,\sigma^2 \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$, then

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

- How to specify prior on (μ,σ^2)

  1. **Recipe 1(Independence):** We could assume μ,σ^2 apriori independent so that

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

  2. **Recipe 2(Conditional):** Or we could specify:

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

- For this example, it turns out that the conjugate priors are:

$$\begin{aligned} \mu|\sigma^2 &\sim N(\mu_0, \frac{1}{\kappa_0}\sigma^2) \\ \sigma^2 &\sim Inv-\chi^2(\nu_0,\sigma_0^2) \end{aligned}$$

Then one can show that the posteriors are as follow:

$$\begin{aligned} \mu|\sigma^2,\mathbf{x} &\sim N(\mu_n, \frac{1}{\kappa_n}\sigma^2) \\ \sigma^2|\mathbf{x} &\sim Inv-\chi^2(\nu_n,\sigma_n^2) \\ \text{where, } \nu_n &= \nu_0 + n \\ \kappa_n &= \kappa_0 + n \\ \mu_n &= \frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}} \\ \sigma_n^2 &= \frac{1}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}} \end{aligned}$$

[**Aside - Inverse Chi-square(Inverse Gamma):** If $X \sim \chi^2_{(\nu)}$ then, $\frac{\nu S^2}{X} \sim Inv-\chi^2(\nu,S^2)$ and the pdf is given by:

$$p(x|\nu,S^2) = \frac{(\nu/2)^{(\nu/2)}}{\Gamma(\nu/2)}(\sigma^2)^{1/2}x^{-(\nu/2+1)}e^{\frac{\nu\sigma^2}{2x}} \;\;]$$

- Thus we can have the **joint posterior** of $(\mu,\sigma^2|\mathbf{x})$ as given by:

$$p(\mu, \sigma^2|\mathbf{x}) = p(\sigma^2|\mathbf{x})p(\mu|\sigma^2, \mathbf{x})$$

Joint posterior is particularly useful when we are interested in the joint structure between the set of parameters, e.g. if we want to know the correlation structure of the parameters or want to construct joint credible set for the set of parameters.

- To make inference about μ we use the **marginal posterior**, which can be obtained as follows:

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

  Unless we are particularly interested in the joint structure of the parameters, marginal posterior is what we require frequently.

- **Note:** If $\mathbf{X_i}|\boldsymbol{\mu}, \Sigma \sim N(\boldsymbol{\mu}, \Sigma)$; where, $\mathbf{x_i}, \boldsymbol{\mu} \in \mathbb{R}^p$, Σ is $p \times p$ positive definite symmetric matrix; then the **conjugate prior** for $(\boldsymbol{\mu}, \Sigma)$ is given by:

$$\boldsymbol{\mu}|\Sigma \sim N(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\Sigma)$$
$$\Sigma \sim \text{Inv-Wishart}(\nu_0, \Lambda_0)$$

3 Computational difficulties

- **Problem:** We have:

$$\mu|\sigma^2, \mathbf{x} \sim N(\mu_n, \frac{1}{\kappa_n}\sigma^2)$$
$$\sigma^2|\mathbf{x} \sim Inv - \chi^2(\nu_n, \sigma_n^2)$$

  How to compute the marginal posterior of $\mu|\mathbf{x}$? If we try:

$$p(\mu|\mathbf{x}) = \int p(\sigma^2|\mathbf{x})p(\mu|\sigma^2, \mathbf{x})d\sigma^2 = \text{ too much Algebra}$$

- **Way-out:** We *sample* from $p(\sigma^2|\mathbf{x})$ and $p(\mu|\sigma^2, \mathbf{x})$! Thus, if we sample (μ, σ^2) from joint posterior $p(\mu, \sigma^2|\mathbf{x})$ then from the resulting sample we can have sample of μ which is from marginal posterior $p(\mu|\mathbf{x})$. Now e.g. for the posterior mean of μ, a good approximation would be the sample mean computed from the sample of μ.

- Outside of very simple setting it is usually much easier to sample from a posterior than to compute it analytically. For sampling we need an algorithm of sampling and computing efficiency to carry out the algorithm.

4 Monte Carlo Integration:

- Let X be a random variable with p.d.f. $\pi(x)$. Suppose, we want to compute:

  $$\theta = E_\pi[X] = \int x\pi(x)dx$$

  If we sample $X_1, X_2, ...X_n \overset{i.i.d.}{\sim} \pi$, then we can use:

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

  to approximate θ. It can be shown: $lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} x_i = \theta$

- Example 1: Suppose we sample from a $N(0, 1)$ distribution and we compute the sample mean, using the following R-code:

  $$mean(rnorm(n = m, mean = 0, sd = 1))$$

  then we will see that the value $\to 0$ as $m \to \infty$

- More generally, to compute $E_\pi[g(X)]$ we can use:

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

  Here also it can be shown: $\hat{\theta} \to \theta$ as $m \to \infty$ for all nice functions g.

- Example 2: $Z \sim N(0, 1)$, compute $E(e^{z+\cos(z)})$. R-code to find the estimate:

  $$z = rnorm(10000); mean(exp(z + cos(z)))$$

5 Gibbs Sampling

- **Idea:** Sample from the posterior distribution and then use the sample to compute quantities of interest such as posterior means, posterior s.d., posterior credible intervals etc.

- **Problem:** How to sample from the posterior $p(\theta|\mathbf{x})$

- **Toy-example:**

  - Goal: To sample from $p(x_1, x_2)$.
    Suppose, we can sample from $p(x_1|x_2)$ and from $p(x_2|x_1)$
    (It can be shown that even if we are sampling from the conditional distributions, the resulting sample turns out to be one drawn from the joint distribution.)

- Algorithm [Gibbs Sampler]: To obtain sample from $p(x_1, x_2)$;

  1. Select starting state $(x_1^{(0)}, x_2^{(0)})$, 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. Set $t = t + 1$, go to 2.

- Example: Suppose, $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$ then to sample from the joint distribution we can use the following conditional distributions:
  $x_1|x_2 \sim N(...)$ and $x_2|x_1 \sim N(...)$.

- In general, to sample from $p(x_1, x_2, ...x_p)$

  1. Select starting state $(x_1^{(0)}, x_2^{(0)}, ..., x_p^{(0)})$, set $t = 0$.
  2. Sample $x_1^{(t+1)}$ from $p(x_1|x_2^{(t)}, x_3^{(t)}, ..., x_p^{(t)})$
  3. Sample $x_2^{(t+1)}$ from $p(x_2|x_1^{(t+1)}, x_3^{(t)}, ..., x_p^{(t)})$ ....

  (k+1). Sample $x_k^{(t+1)}$ from $p(x_k|x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{k-1}^{(t+1)}, x_{k+1}^{(t)}, ...x_p^{(t)})$......

  (p+1). Sample $x_p^{(t+1)}$ from $p(x_p|x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{p-1}^{(t+1)})$

  Set $t = t + 1$ and go to 2.

- If $p(x_1, x_2, ...x_p)$ are highly correlated Gibbs sampling takes much more time to converge than when they are nearly independent.

6 Markov Chain:

A Markov Chain is a stochastic process with the property that future states are independent of the past states given the current state.

- Sequence: $(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, ...)$
- Markov Chain: $P(x^{(t+1)}|x^{(1)}......x^{(t)}) = P(x^{(t+1)}|x^{(t)})$
- Markov chains are controlled by their transition kernel/density.
- Suppose x takes k possible values:

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

give the transition probabilities.

- Consider the following Markov chain which has three states. The transition probabilities are described in the diagram.

stoch.jpeg

- For the above example, the transition matrix would be:

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

- Markov chains are described by their transition matrices.
- If x has a continuous state space (e.g. $x \in \mathbb{R}$) then the markov chain is described by the transition kernel/density:

$$P(X^{(t+1)} \in A | x^{(t)} \in U) = \mathbb{P}(U, A) \text{ for all i, j}$$

Important Definitions

- **Irreducibility:** A markov chain is irreducible if it is possible to reach every state from every other state in a finite no. of moves.
- **Aperiodicity:** Starting at state i, you don't have to return to state i at regular period.
- **Transience:** A state i is said to be transient if starting at state i, there is a non-zero probability of never returning to state i.
- **Recurrence:** A state i is said to be recurrent if it is not transient.
- **Positive Recurrence:** A recurrent state i is positive recurrent if it's expected return time is finite.
- **Ergodicity:** Aperiodicity + Positive Recurrence
- For an ergodic markov chain, the long-run average of the chain converges to a stationary distribution.

$$P(X^{(t)} = i) \overset{t \to \infty}{\to} \pi_i$$

- For stationary distribution:

$$\pi = \pi D,\ \pi(y) = \int \pi(x) p(x, y) dx$$