Bayesian Statistical Methods: A Primer

SU WANG

Spring 2016

Contents

1 Basics: Prior, Likelihood, Posterior & Predictive

1.1 Prior, Likelihood, & Posterior

In a nutshell, *Bayesian methods* seek to integrate expert knowledge (or reasonable prior beliefs) and new data in building & updating predictive models. For instance, to write a density function for the probability of getting a head in tossing a chosen coin (i.e. $\pi(\theta)$), we have[1]:

- Prior[2] (Belief/Knowledge):

$$\begin{aligned}\pi(\theta) &= Beta(\theta|\alpha,\beta) \\ &= \frac{1}{Beta(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\end{aligned} \tag{1.1}$$

- Likelihood (New Data): 7 out of 10 tosses came up head.

$$l(x_1,...x_{10}|\theta) = \prod_{i=1}^{10}\theta^i(1-\theta)^{1-i} \quad ,i=1 \text{ for head; } i=0 \text{ for tail} \tag{1.2}$$

With the information, applying *Bayes' Theorem*, we compute the density function – the *posterior density*, where the prior belief and the new data are integrated. By Bayes' Theorem, we know that[3]:

$$\begin{aligned}P(\theta|X) &= \frac{P(\theta,X)}{P(X)} \\ &= \frac{P(X|\theta)P(\theta)}{P(X)} \\ &= \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)d\theta} \\ &\propto P(X|\theta)P(\theta)\end{aligned} \tag{1.3}$$

Therefore, the posterior density, which is notated as $\pi_n(\theta|X)$ by convention, is

[1] θ denotes the parameter of the density function which "regulates" the probability distribution. It is equivalent to the mean and the standard deviation in the density function of the *Gaussian Distribution*.

[2] Set $\alpha = \beta$, for instance, our prior knowledge will be "the coin is fair". The greater the value is, the more strongly we believe in its fairness.

[3] X denotes new data.

computed as follows, using the current example:

$$\begin{aligned}\pi_n(\theta|x_1,...,x_{10}) &\propto l(x_1,...x_{10}|\theta)\pi(\theta)\\ &= \prod_{i=1}^{10}\theta^i(1-\theta)^{1-i}\pi(\theta)\\ &= \theta^7(1-\theta)^3\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &= \theta^{\alpha+7-1}(1-\theta)^{\beta+3-1}\end{aligned}$$

Therefore, for this example, the (posterior) density function after updating with the new data is a $Beta$ distribution with the parameters $\alpha_n = \alpha+7; \beta_n = \beta+3$, which can be interpreted as slightly "moved" by the data in favor of the probability of getting heads.

1.2 Predictive

Suppose we would like to use the newly updated density function to make a prediction as the outcome of the 11th toss of the coin. We do not know the density function of the outcome of the toss, notated as $f(x)$, in prior. To make use of the posterior density we have computed, we compute the *expecatation* of $f(x)$ – $f_n(x)$, which is called the *predictive density* of $f(x)$, as follows:

$$\begin{aligned}f_n(x) &= \int f(x|\theta)\pi_n(\theta|x_1,...,x_{10})d\theta\\ &= \int \theta^x(1-\theta)^{1-x}\frac{1}{Beta(\alpha_n,\beta_n)}\theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}d\theta\\ &= \frac{Beta(\alpha_n+x,\beta_n+1-x)}{Beta(\alpha_n,\beta_n)}\int \frac{1}{Beta(\alpha_n+x,\beta+1-x)}\theta^{\alpha_n+x-1}(1-\theta)^{\beta_n+1-x-1}d\theta\\ &= \frac{\Gamma(\alpha_n+x)\Gamma(\beta_n+1-x)}{\Gamma(\alpha_n+\beta_n+1)}\frac{\Gamma(\alpha_n+\beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)}\end{aligned}$$

It is readily checked that, if we set the prior parameters $\alpha = \beta = 2$, then the probability of the 11th toss coming up head is $f_n(x) = \frac{9}{14} \approx 0.643$, which coincides with intuition: after the updating, the favor is slightly slanted towards getting a head.

2 Hypothesis Testing

To set the stage, suppose we know that some data are normally distributed (i.e. $X \sim N(\mu, \sigma^2)$). For simplicity, assume that μ is unknown and σ^2 is known. From the knowledge of an expert, we know that μ is also normally distributed[4]: $\mu \sim N(\nu, \tau^2)$.

Now suppose we would like to test the hypothesis that $\mu = \tilde{\mu}$. Formally expressed:

$$H_0 = \mu = \tilde{\mu}$$
$$H_1 = \mu \neq \tilde{\mu}$$

Further suppose that we have n pieces of data, with which we are able to compute a posterior density. The Bayesian hypothesis testing then computes the ratio between the probability where H_0 is true and that where H_1 is true. The result is a statistic $\mathcal{B}$. By selecting a confidence level, we have a threshold value λ, such that if $\mathcal{B} > \lambda$, the null hypothesis – H_0 – will be accepted. Otherwise, it is rejected.

From the previous section, we know that the posterior densities for the H_0-true and H_1-true cases are $P(H_0|x_1, ..., x_n)$ and $P(H_1|x_1, ..., x_n)$. The computation for the ratio goes as follows[5]:

$$\begin{aligned}
\mathcal{B} &= \frac{P(H_0|x_1, ..., x_n)}{P(H_1|x_1, ..., x_n)} \\
&= \frac{P(x_1, ..., x_n|H_0)}{P(x_1, ..., x_n|H_1)} \frac{P(H_0)}{P(H_1)} \\
&= \frac{\prod_{i=1}^{n} N(x_i|\tilde{\mu}) N(\tilde{\mu}|\nu)}{\int_{\mu \neq \tilde{\mu}} \prod_{i=1}^{n} N(x_i|\mu) N(\mu|\nu) d\mu} \\
&= \frac{\prod_{i=1}^{n} N(x_i|\tilde{\mu}) N(\tilde{\mu}|\nu)}{\int \prod_{i=1}^{n} N(x_i|\mu) N(\mu|\nu) d\mu}
\end{aligned} \tag{2.1}$$

The computation is simple despite the ostensible complex form, as two normal densities conjugate to produce a posterior normal distribution $N(X|\nu_n, \tau_n^2)$, where $\nu_n = \frac{n\bar{x}/\sigma^2 + \nu/\tau^2}{n/\sigma^2 + 1/\tau^2}$, and $\tau_n^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}$.

[4] We are assuming two conjugatable distrubtions for convenience. The techniques that handle the cases where this is not the case will be covered later on.

[5] Explanation for the last step: In a continuous density, it can be proved that the probability mass at an exact point is equal to 0 (i.e. $P(X = x) = 0$).

3 Monte Carlo Integration

Oftentimes the posterior density[6] $\pi_n(\theta)$ does not conjugate with the density of data $f(\theta)$. Therefore, we will not be able to compute for a predictive density as described in section 1.2. Specifically, the computation for the following integral is intractable:

$$f_n(\theta) = \int f(\theta)\pi_n(\theta)d\theta \tag{3.1}$$

However, if the posterior density is a known distribution (e.g. a normal distribution), we *will* be able to sample from it. Suppose we sampled M data points from $\pi_n(\theta)$: $\{\theta_1, ..., \theta_M\}$, we can use these data to build a *discretized approximation* of the integral $f_n(\theta)$ as follows:

$$\hat{f_n(\theta)} = \frac{1}{M}\sum_{i=1}^{M} f(\theta_i), \quad \text{where } i = 1, ..., M \tag{3.2}$$

This discretization of a continuous distribution is called *Monte Carlo Integration*, which is illustrated graphically as follows:

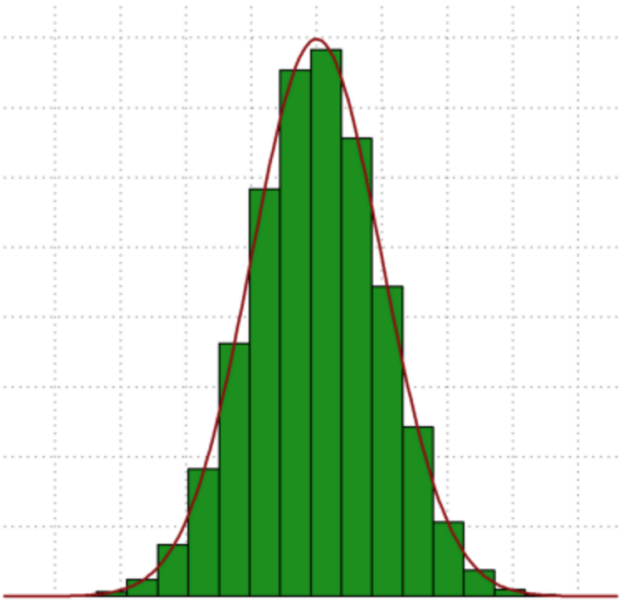

Figure 3.1: Discretized Approximation in MCMC

It can be shown that the expectation of the approximation, and with the number of samples approaches infinity, the variance of the estimate from the true value

[6] Or any density with which we hope to build upon to make predictions with regard to new data.

approaches 0:

$$E[\hat{f_n}(\theta)] = E[\frac{1}{M}\sum_{i=1}^{M} f(\theta_i)] = \frac{1}{M}\sum_{i=1}^{M} E[f(\theta_i)] = \frac{1}{M}\sum_{i=1}^{M} f_n(\theta) = f_n(\theta) \quad (3.3)$$

$$Var[\hat{f_n}(\theta)] = Var[\frac{1}{M}\sum_{i=1}^{M} f(\theta_i)] = \frac{1}{M^2}\sum_{i=1}^{M} Var[f(\theta)] \leq \frac{1}{M^2}\int f(\theta)^2 \pi_n(\theta) d\theta \quad (3.4)$$

$$\text{where } \sum_{i=1}^{M} Var[f(\theta)] = \int f(\theta)^2 \pi_n(\theta) d\theta - \hat{f_n}(\theta)^2 \leq \int f(\theta)^2 \pi_n(\theta) d\theta$$

More specifically, we have $\frac{1}{M^2}\int f(\theta)^2 \pi_n(\theta) d\theta \to 0$, therefore by Comparison Test, $Var[\hat{f_n}(\theta)]$ must converge to 0 too. This shows that, with a large-enough sample size, Monte Carlo Integration produces decent approximation to the true value of the integral in Eq. 3.1. Finally, note that Monte Carlo Integration is under the assumption that the samples $\{\theta_1, ..., \theta_M\}$ are drawn from identical and independently distributed distributions (i.e. i.i.d.)[7].

[7]To make a simple analogy, consider the case of tossing a fair coin. Each time the same coin is tossed – therefore the distribution of the outcome is identical; and the outcome of each toss is independent of the results from other tosses.

4 Sampling

4.1 Rejection Sampling

The distributions we have seen so far have been "named" densities (e.g. normal, beta, gamma, etc.), from which we may directly sample for data to analyze (or run simulation). In practice, however, oftentimes we need to handle "exotic" densities the properties of which we do not know well enough to allow such luxury. The following density is an example[8]:

$$f(\theta) \propto \theta^{a-1} e^{b\theta} (1 - e^{-c\theta})^d \tag{4.1}$$

To sample from $f(\theta)$, our first step is to perform some algebraic manipulation to write it in the form of the product of two density, one of which is a "named" density: $f(\theta) = h(\theta)g(\theta)$. In Eq. 4.1, for instance, we may take $h(\theta) = (1 - e^{-c\theta})^d$, and $g(\theta) = \theta^{a-1} e^{b\theta}$, where $g(\theta)$ is our "named" density – $g(\theta) \sim Ga(\theta|a, b)$.

We then introduce a latent parameter u, which is uniformly distributed[9], to build a join density $f(\theta, u)$. Note if we sample $\tilde{\theta}$ and $\tilde{u}$ from $f(\theta, u)$, $\tilde{\theta}$ will be from $f(\theta)$. Suppose we decide to get M samples, we may set the joint density as follows[10]:

$$f(\theta, u) = M \cdot 1(0 < u < \frac{h(\theta)}{M}) \cdot g(\theta) \tag{4.2}$$

This is an legitimate operation because:

$$\begin{aligned} \int f(\theta, u) du &= M \int 1(0 < u < \frac{h(\theta)}{M}) \cdot g(\theta) du \\ &= M \int_0^{\frac{h(\theta)}{M}} 1 du \cdot g(\theta) \\ &= h(\theta) \cdot (g\theta) = f(\theta) \end{aligned}$$

For the convenience of sampling, we further modify $f(\theta, u)$ as follows:

$$f(\theta, u) = M \cdot 1(u < \frac{h(\theta)}{M}) \times 1(0 < u < 1) \cdot g(\theta) \tag{4.3}$$

Finally, we run the following algorithm to sample $\theta_1, ..., \theta_M$:

- Sample $\tilde{\theta}$ from $g(\theta)$
- Sample $\tilde{u}$ from $Unif(0, 1)$
- Bank the sample $\tilde{\theta}$ (i.e. accept that $\tilde{\theta}$ is from $f(\theta)$) if $\tilde{u} < \frac{h(\theta)}{M}$
  Else, resample.

[8] This distribution describe the effect of a certain drug. The setup of the function is based on expert knowledge, therefore cannot be somehow "engineered" for the convenience of sampling.

[9] I.e. u is from $Unif(u)$.

[10] $1(0 < u < \frac{h(\theta)}{M})$ is an indicator function, where the function returns 1 if $0 < u < \frac{h(\theta)}{M}$; it returns 0 otherwise.

- Iterate until having M samples accepted.

The acceptance rate of the rejection sampler can be computed as follows:

$$\begin{aligned} P(\tilde{u} < \frac{h(\tilde{\theta})}{M}) &= \int P(\tilde{u} < \frac{h(\tilde{\theta})}{M} | \tilde{\theta}) g(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{M} \int h(\tilde{\theta}) g(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{M} \end{aligned} \tag{4.4}$$

In practice, however, the rejection sampling often fail due to high rejection rate – the sampling thus becomes extremely slow. To demonstrate, consider the following example:

$$\begin{aligned} f(\theta) &= \frac{1}{\Gamma(a)} \theta^{(a-1)} e^{-\theta} \\ h(\theta) &= \frac{1}{\Gamma(a)} \theta^{(a-1)} e^{-\frac{1}{2}\theta}; \quad g(\theta) = \frac{1}{2} e^{-\frac{1}{2}\theta} \end{aligned}$$

Here we have our "named" density $g(\theta)$, and $f(\theta) = h(\theta)g(\theta)$. However, the sampler usually will not work. We know that the rejection rate of the sampler is $\frac{1}{M}$. To compute the upperbound of M, we compute $argmax_\theta h(\theta)$:

$$\begin{aligned} argmax_\theta h(\theta) &= argmax_\theta \frac{1}{\Gamma(a)} \theta^{(a-1)} e^{-\frac{1}{2}\theta} \\ &= argmax_\theta \theta^{(a-1)} e^{-\frac{1}{2}\theta} \end{aligned}$$

To simplify computation, we take $log(\theta^{(a-1)} e^{-\frac{1}{2}\theta})$[11], differentitiate the resulting equation, and set it to 0 to solve for θ.

$$log(\theta^{(a-1)} e^{-\frac{1}{2}\theta}) = (a-1) log(\theta) - \frac{1}{2}\theta \tag{4.5}$$

$$((a-1) log(\theta) - \frac{1}{2}\theta)' = 0 \implies \frac{a-1}{\theta} = \frac{1}{2} \implies \theta = 2(a-1) \tag{4.6}$$

Plug θ into $h(\theta)$, we get the upperbound value of $M = \frac{2}{\Gamma(a)} (2(a-1))^{a-1} e^{-\frac{1}{2}2(a-1)}$. The rejection rate is therefore $\frac{1}{M} = \frac{\frac{1}{2}\Gamma(a) e^{a-1}}{2(a-1)^{a-1}}$. It is immediately obvious that, even with a moderate value of a, $\frac{1}{M}$ becomes extremely small (e.g. $\frac{1}{M} = 10^{-34}$ when $a = 30$). With the hefty rejection rate, the sampling will become insurmountably slow. Fortunately, it is possible to use a technique called *stationary process* to overcome the problem.

[11] $argmax_\theta h(\theta) = argmax_\theta log(h(\theta))$

4.2 Transition Density & Stationary Process

Let $f(\theta)$ be the density of a named distribution from which we wish to sample, and an attempt of rejection sampling failed due to high rejection rate. Suppose we manage to get one sample θ_1 after a long wait. Although one sample is not useful, we know that θ_1 is a quality sample from $f(\theta)$. If there exists some density $\xi(\theta_{i+1}|\theta_i)$, with which we may get a θ_2 from $\xi(\theta_2|\theta_1)$, which depends on θ_1, then we will be able to generate the set of $\{\theta_1, ..., \theta_M\}$, which we may use as the desired sample set.

There is however a problem. For instance, to use $f(\theta)$ to find the predicted distribution $g(\theta)$ of the next θ, we compute the Monte Carlo Integration approximation of the integral $\int g(\theta)f(\theta)d\theta$. I.e. $\frac{1}{M}\sum_{i=1}^{M} g(\theta_i)$, where $\theta_1, ..., \theta_M$ are from $f(\theta)$. Monte Carlo Integration assumes $\{\theta_1, ..., \theta_M\}$ to be i.i.d. It is clear that our set, which is generated from $\xi(\theta_{i+1}|\theta_i)$ using θ_1 as the seed, has dependent samples. Fortunately, there is a theorem which states that if we select a $\xi(\theta_{i+1}|\theta_i)$ which satisfies some criterion, the dependent set $\{\theta_1, ..., \theta_M\}$ can still be used in Monte Carlo Integration:

- **Ergodic[12] Theorem for Monte Carlo Integration**

  Let $\theta_1, ..., \theta_M$ be a *Markov Sequence*[13]. Then, the integral

  $$I_g = \int g(\theta)f(\theta)d\theta$$

  can still be estimated by its Monte Carlo Integration approximation:

  $$\hat{I_g} = \frac{1}{M}\sum_{i=1}^{M} g(\theta_i),$$

  provided there is a *transition density* $\xi(\theta_{i+1}|\theta_i)$ such that

  $$f(\theta_{i+1}) = \int \xi(\theta_{i+1}|\theta_i)f(\theta_i)d\theta_i$$

  The generation $\{\theta_1, ..., \theta_M\}$ using a transition density is called a *stationary process*.

Essentially, the theorem says that sampling, for instance, θ_2 from a transition density $\xi(\theta_2|\theta_1)$ is the same as sampling it from the original distribution $f(\theta)$ directly. This sampling process produces a *pseudo-i.i.d.* sample set which enables us to perform Monte Carlo Integration.

[12] *Ergodicity*: i) every state in the process must be aperiodic — the system does not return to the same state at fixed intervals; and ii) every state must be positive recurrent — the expected number of steps for returning to the same state is finite.

[13] There are some complicated additional constraints which are ignored here because they do not directly relate to the current discussion.

While finding a easy-to-sample transition density for a difficult-to-sample density involves some complex mathematical techniques, it is possible to demonstrate the stationary process using a simple example. Concretely, say we would like to sample from the density $f(\theta) = N(\theta|0,1)$, it can be proved that if $\xi(\theta_{i+1}|\theta_i) = N(\theta_{i+1}|p\theta_i, 1-p^2)$ is selected as the transition density, $f(\theta_{i+1}) = \int \xi(\theta_{i+1}|\theta_i) f(\theta_i) d\theta_i$.

4.3 Gibbs Sampling

When sampling from a difficult-to-sample density where there are 2 or more variables involved (e.g. sampling μ and λ from some distribution $f(\mu, \lambda)$), there is a neat trick to easily build a transition density that allows us to make use of the stationary process. Suppose we have the following n-observation[14] likelihood (i.e. $l(\mu,\lambda) = \prod_{i=1}^{n} g(x_i|\mu,\lambda)$) density, and two prior densities (i.e. $\pi(\mu)$ and $\pi(\lambda)$), which produce a difficult-to-sample posterior density (i.e. $\pi_n(\mu,\lambda)$)[15]:

$$
\begin{aligned}
g(x|\mu,\lambda) &\sim N(x|\mu, 1/\lambda), \text{ where } \lambda = 1/\sigma^2 \\
l(\mu,\lambda) &= \prod_{i=1}^{n} \lambda^{1/2} e^{-\frac{1}{2}\lambda(x_i-\mu)^2} \\
\pi(\mu) &\sim N(\mu|\nu,\tau^2) \\
\pi(\lambda) &\sim Ga(\lambda|a,b)
\end{aligned}
$$

$$
\begin{aligned}
\pi_n(\mu,\lambda) &\propto \pi(\mu)\cdot\pi(\lambda)\cdot l(\mu,\lambda) \\
&\propto e^{-\frac{1}{2}(\mu-\nu)^2/\tau^2} \cdot \lambda^{(a-1)} e^{-\lambda b} \cdot \lambda^{n/2} e^{-\frac{1}{2}\lambda\sum_{i=1}^{n}(x_i-\mu)^2}
\end{aligned} \tag{4.7}
$$

Now, the criterion for stationary process[16] $\pi_n(\mu_2,\lambda_2) = \int\int \xi(\mu_2,\lambda_2|\mu_1,\lambda_1)\pi_n(\mu_1,\lambda_1)d\mu_1 d\lambda_1$ will be satisfied if we choose the transition density as follows:

$$
\xi(\mu_2,\lambda_2|\mu_1,\lambda_1) = \pi_n(\mu_2|\lambda_2)\cdot\pi_n(\lambda_2|\mu_1) \tag{4.8}
$$

[14] That is, n observations from the same density.

[15] The parameters of the priors are known.

[16] Using the transition from the 1st the 2nd sample to demonstrate, for simplicity. This generalizes to all following cases.

PROOF:

$$\begin{aligned}
\pi_n(\mu_2, \lambda_2) &= \int\int \xi(\mu_2, \lambda_2|\mu_1, \lambda_1)\pi_n(\mu_1, \lambda_1)d\mu_1 d\lambda_1 \\
&= \int\int \pi_n(\mu_2|\lambda_2) \cdot \pi_n(\lambda_2|\mu_1) \cdot \pi_n(\mu_1, \lambda_1)d\mu_1 d\lambda_1 \\
&= \int \pi_n(\mu_2|\lambda_2) \cdot \pi_n(\lambda_2|\mu_1) \cdot \pi_n(\mu_1)d\mu_1 \\
&= \pi_n(\mu_2|\lambda_2) \cdot \int \pi_n(\lambda_2|\mu_1) \cdot \pi_n(\mu_1)d\mu_1 \\
&= \pi_n(\mu_2|\lambda_2) \cdot \int \pi_n(\mu_1, \lambda_2)d\mu_1 \\
&= \pi_n(\mu_2|\lambda_2) \cdot \pi_n(\lambda_2) = \pi_n(\mu_2, \lambda_2)
\end{aligned}$$

Now all we have to do is to sample from the transition density in Eq. 4.8. As directly sampling from $\xi(\mu_2, \lambda_2|\mu_1, \lambda_1)$ is not viable, we may conduct the sampling as follows: First assuming we have obtained a μ_1 somehow[17]. Then get λ_2 from $\pi_n(\lambda_2|\mu_1)$ with the initializer. Having obtained λ_2, we may further sample μ_2 from $\pi_n(\mu_2|\lambda_2)$. Now we have the first pair of samples – μ_2 and λ_2, continuing the process, we will be able to get all M samples desired.

Note that we have been assuming that it is possible to directly sample from $\pi_n(\lambda_2|\mu_1)$ and $\pi_n(\mu_2|\lambda_2)$. That is, these are "named" densities. Now we show, through some algebra, that this is indeed the case[18].

$$\begin{aligned}
\pi_n(\lambda|\mu) &\propto \lambda^{a-1}e^{-\lambda b}\lambda^{n/2}e^{-\frac{1}{2}\lambda\sum_{i=1}^{n}(x_i-\mu)^2} \\
&\propto Ga(\lambda|a + \frac{n}{2}, b + \frac{1}{2}\lambda\sum_{i=1}^{n}(x_i - \mu)^2)
\end{aligned} \tag{4.9}$$

$$\begin{aligned}
\pi_n(\mu|\lambda) &\propto e^{-\frac{1}{2}(\mu-\nu)^2/\tau^2}e^{-\frac{1}{2}\sum_{i=1}^{n}\lambda(x_i-\mu)^2} \\
&\propto N(\mu|\frac{n\bar{x}\lambda + \nu/\tau^2}{\lambda n + 1/\tau^2}, \frac{1}{\lambda n + 1/\tau^2})
\end{aligned} \tag{4.10}$$

With all the components ready, we present the *Gibbs Sampling Algorithm*:

- Obtain an initializer μ_1.
- Sample λ_2 from $Ga(\lambda_2|\mu_1)$.

[17]This could be done using rejection sampling, if one wish to initialize the sample process from a good starting point. However it can be proved that it does not matter how the initializer is chosen – the process will gradually move closer to the true density of it. The proof involves complex operations in ergodic theory, and is thus omitted.

[18]In the following, in computing $\pi_n(\lambda|\mu)$, for instance, we take μ to be a constant and thus all the terms that involve instances of it in the original formula $\pi_n(\mu, \lambda)$ can be "proportion to"-ed out.

- Sample μ_2 from $N(\mu_2|\lambda_2)$.
- Iterate until obtaining all M samples: $\{(\mu_2, \lambda_2), ..., (\mu_{M+1}, \lambda_{M+1})\}$

To run the algorithm as described above, we are assuming that the component densities of the transition density (cf. Eq. 4.8) are "named" densities. In practice, however, this assumption is often false and unrealistic. For instance, consider the following example, where the prior $\pi(\theta)$ is an exponential distribution, and the prior $\pi(\phi)$ is a gamma distribution:

$$
\begin{aligned}
f(x|\theta,\phi) &= \phi^{\theta}\frac{x^{\theta-1}}{\Gamma(\theta)}e^{-\phi x} \\
l(\theta,\phi) &= \prod_{i=1}^{n} f(x_i|\theta,\phi) \\
\pi(\theta) &\propto e^{-\theta} \\
\pi(\phi) &\propto \phi^{a-1}e^{-b\phi}
\end{aligned}
$$

$$
\begin{aligned}
\pi_n(\theta,\phi) &\propto \pi(\theta)\cdot\pi(\phi)\cdot l(\theta,\phi) \\
\pi_n(\phi|\theta) &\propto \phi^{\theta n+a-1}e^{-\phi(b+\sum_{i=1}^{n} x_i)} \\
\pi_n(\theta|\phi) &\propto e^{-\theta}\phi^{n\theta}\frac{\prod_{i=1}^{n} {x_i}^{\theta}}{\Gamma(\theta)^n} \\
&\propto \frac{e^{-1}\phi^{n}\prod_{i=1}^{n} {x_i}^{\theta}}{\Gamma(\theta)^n}
\end{aligned}
$$

While $\pi_n(\phi|\theta)$ is a gamma density (i.e. $Ga(\theta n + a, b + \sum_{i=1}^{n} x_i)$), $\pi_n(\theta|\phi)$ is *not* a "named" density, and thus cannot be sampled from directly. As $\pi_n(\theta|\phi)$ only has one parameter θ, we will not be able to implement a Gibbs Sampler for it. In the two following sections, we look at two alternative methods that handle the single-variable difficult-to-sample densities: *Metropolis-Hastings Sampler* and *Latent Variable Sampler*.

4.4 Metropolis Hastings Sampling

Recall from section 4.2 that, to sample from a density $f(\theta)$ which cannot be sampled from directly, we need to find a transition density $\xi(\theta_{i+1}|\theta_i)$ such that $f(\theta_{i+1}) = \int \xi(\theta_{i+1}|\theta_i) f(\theta_i) d\theta_i$. The pseudo-i.i.d sample set $\{\theta_1, ..., \theta_M\}$ obtained from the transition density can then be used in the Monte Carlo Integration to make predictions. The problem we encounter towards the end of section 4.3 is that in the case of some transition densities, it is impossible to sample directly, because the component densities of the transition densities are not "named" densities. The solution we propose for the problem is as follows: Instead of sampling the transition density by sampling from all of its component densities, as we did in Gibbs Sampling, we engineer the transition density

itself algebraically such that one of its components is a "named" density. By sampling from this "named" density, and constrain its output somehow such that the criterion for stationary process is satisfied (cf. section 4.2).

Now we elaborate on the details in the implementation of the idea.

- **Metropolis-Hastings Transition**

  Let $f(\theta)$ be the difficult-to-sample density we would like to sample from, and let there be a density $\xi(\theta_{i+1}|\theta_i)$ such that the following equation is satisfied[19] — the *stationarity* is satisfied:

  $$f(\theta_2)\xi(\theta_1|\theta_2) = f(\theta_1)\xi(\theta_2|\theta_1) \tag{4.11}$$

  Then, it can be shown that $\xi(\theta_{i+1}|\theta_i)$ is a transition density of $f(\theta)$. Specifically, the criterion for stationary process is satisfied:

  $$f(\theta_2) = \int \xi(\theta_2|\theta_1)f(\theta_1)d\theta_1 \tag{4.12}$$

  $\xi(\theta_{i+1}|\theta_i)$ is called the *Metropolis-Hastings Transition*.

PROOF:

$$\begin{aligned} f(\theta_2) &= \int \xi(\theta_2|\theta_1)f(\theta_1)d\theta_1 \\ &= \int \xi(\theta_1|\theta_2)f(\theta_2)d\theta_1, \quad \text{by Eq. 4.11} \\ &= f(\theta_2)\int \xi(\theta_1|\theta_2)d\theta_1 \\ &= f(\theta_2) \end{aligned}$$

There are many ways in which $\xi(\theta_{i+1}|\theta_i)$ can be set up to satisfy Eq. 4.11. We consider the most common setup, the motivation of which we will explore shortly after:

$$\xi(\theta_2|\theta_1) = \alpha(\theta_1,\theta_2)q(\theta_2|\theta_1) + (1 - r(\theta_1)) \cdot 1(\theta_1 = \theta_2), \quad \text{where} \tag{4.13}$$

$q(\theta_2|\theta_1)$ is a "named" density

$$\alpha(\theta_1,\theta_2) = min\{1, \frac{f(\theta_2)q(\theta_1|\theta_2)}{f(\theta_1)q(\theta_2|\theta_1)}\} \tag{4.14}$$

$$r(\theta) = \int \alpha(\theta_1,\theta_2) \cdot q(\theta_2|\theta_1)d\theta_2 \tag{4.15}$$

$1(\theta_1 = \theta_2)$ is an indicator function which is equal to 1 only when $\theta_1 = \theta_2$ (4.16)

[19] Again, for simplicity, we only take the transition from the 1st to 2nd sample for instance.

It can be readily checked that Eq. 4.13 satisfies the condition Eq. 4.11. To explain why we can sample $\xi(\theta_2|\theta_1)$ be sampling from our "named" density $q(\theta_2|\theta_1)$, we rewrite Eq. 4.13 as follows:

$$\xi(\theta_2|\theta_1) = r(\theta_1) \cdot (\frac{\alpha(\theta_1, \theta_2) q(\theta_2|\theta_1)}{r(\theta_1)}) + (1 - r(\theta_1)) \cdot 1(\theta_1 = \theta_2) \tag{4.17}$$

Now, it is quite clear that Eq. 4.17 is a mixed density where, if we sample a θ from $\xi(\theta_2|\theta_1)$, then the probability that the θ is from the first term (i.e. $\frac{\alpha(\theta_1, \theta_2) q(\theta_2|\theta_1)}{r(\theta_1)}$) is $r(\theta_1)$, and the probability that it is from the second term (i.e. $1(\theta_1 = \theta_2)$) is $1 - r(\theta_1)$. To sample from $\xi(\theta_2|\theta_1)$, therefore, we first start with an initializer θ_1[20], and sample a $\tilde{\theta_2}$ from $q(\theta_2|\theta_1)$, check whether it is from the first term[21]. If this is the case, then we bank $\tilde{\theta_2}$ as it is: taking our first sample $\theta_2 = \tilde{\theta_2}$. Otherwise, we set $\theta_2 = \theta_1$, and then bank it. We iterate as such until we have all M samples desired.

To check whether $\tilde{\theta_2}$ is from $\frac{\alpha(\theta_1, \theta_2) q(\theta_2|\theta_1)}{r(\theta_1)}$, we take an auxiliary variable $u \sim Unif(0, 1)$ and check if $u < \alpha(\theta_1, \tilde{\theta_2})$, an event with a probability $r(\theta)$:

$$\begin{aligned} r(\theta) &= \int \alpha(\theta_1, \theta_2) \cdot q(\theta_2|\theta_1) d\theta_2 \\ &= \int \int_0^{\alpha(\theta_1, \theta_2)} 1du \cdot q(\theta_2|\theta_1) d\theta_2 \end{aligned} \tag{4.18}$$

Essentially, the probability of sampling a $\tilde{\theta_2}$ from $q(\theta_2|\theta_1)$ and at the same time having $u \sim Unif(0, 1)$ in the range $\int_0^{\alpha(\theta_1, \theta_2)} 1du$ is $r(\theta)$. In other words, the sampling method guarantees that at a chance $r(\theta)$ the sample $\tilde{\theta_2}$ is from $\frac{\alpha(\theta_1, \theta_2) q(\theta_2|\theta_1)}{r(\theta_1)}$. On the other hand, we therefore know that the event $u \geq \alpha(\theta_1, \tilde{\theta_2})$ happens at a chance $1 - r(\theta)$. So when the alternative event takes place, we bank the new sample as of the same value as θ_1 (i.e. taking it that it is from $1(\theta_1 = \theta_2)$).

The following is a more formal description of the sampling process, which is called the *Metropolis-Hastings Algorithm*:

- Obtain an initializer θ_1.
- Sample $\tilde{\theta_2}$ from $q(\theta_2|\theta_1)$.
- Sample u from $Unif(0, 1)$.
- Bank the new sample as $\theta_2 = \tilde{\theta_2}$, if $u < \alpha(\theta_1, \tilde{\theta_2})$; Otherwise, bank the new sample as $\theta_2 = \theta_1$.
- Iterate until having M samples.

[20] Cf. footnote 16.
[21] Checking technique will soon be discussed.

Now we come back to explore how the acceptance function α is chosen. Recall that, to build a Markov process in sampling, stationarity has to be satisfied (eq. 4.11), and the equation can be rewritten as

$$\frac{\xi(\theta_2|\theta_1)}{\xi(\theta_1|\theta_2)} = \frac{f(\theta_2)}{f(\theta_1)} \tag{4.19}$$

Now, lets break the transition density $\xi(\theta_2|\theta_1)$ into two parts as follows:

$$\xi(\theta_2|\theta_1) = \alpha(\theta_1, \theta_2)q(\theta_2|\theta_1) \tag{4.20}$$

where $\alpha(\theta_1, \theta_2)$ is the acceptance function (i.e., the acceptance for θ_1 transition to θ_2), and $q(\theta_2|\theta_1)$ the *proposal density* — a "named" conditional probability of a transition to θ_2 given θ_1. Now we plug eq. 4.20 into eq. 4.19, we get

$$\frac{\alpha(\theta_1, \theta_2)}{\alpha(\theta_2, \theta_1)} = \frac{f(\theta_2)q(\theta_1|\theta_2)}{f(\theta_1)q(\theta_2|\theta_1)} \tag{4.21}$$

The LHS of eq. 4.21 can be interpreted as the odds of transitioning from θ_1 to θ_2 against the other way around, and the RHS can then be treated as a *transition acceptance indicator*. Thus, we set the acceptance function $\alpha(\theta_1, \theta_2) = min\{1, \frac{f(\theta_2)q(\theta_1|\theta_2)}{f(\theta_1)q(\theta_2|\theta_1)}\}$ (i.e., eq. 4.14) to fulfill the condition 4.21 by saying: we always accept when the acceptance indicator is bigger than 1, and we reject accordingly when the acceptance is smaller than 1 (i.e., *Metropolis Choice*.)

The choice of the "named" density $q(\theta_2|\theta_1)$ depends on the domain of the transition density, which in turn depends on the domain of the original density from which we aim to sample. For instance, if $\theta \in \mathbf{R}$, then we may set $q(\theta_2|\theta_1) \sim N(\theta_2|\theta_1, \sigma^2)$. If $\theta \in (0, \infty)$, then maybe we can employ either a $Ga(\theta|a, b)$ or a lognormal density $logN(\theta_2|\theta_1, \sigma^2)$. Other than the constraint on domain, the choice tends to be by and large empirical.

Finally note that the MH-algorithm can be used in a Gibbs Sampler too. Taking the example towards the end of section 4.3, we may sample from $\pi_n(\phi|\theta)$ as usual, and implement an MH-step to sample from $\pi_n(\theta|\phi)$, which cannot be sampled from directly.

4.5 Latent Variable Sampling

In handling the single-variable difficult-to-sample densities, the MH-algorithm is not the only game in town. We may also turn the single-variable densities into bivariate (or even multivariate, if necessary) densities to enable the implementation of a Gibbs Sampler. As the introduced variable does not actually exist, and is no more than a technical convenience, it is called the *Latent Variable*. The sampling method is thus referred to as the *Latent Variable Sampling*.

We first demonstrate the latent variable sampling works, and then explain why the introduction of an "alien" variable will not affect the validity of the sampling. Consider the single-variable difficult-to-sample following density:

$$f(\theta) \propto \theta^{a-1} e^{-b\theta} (1 - e^{-c\theta})^d \tag{4.22}$$

As we have shown in section 4.1, implementing a rejection sampler for the density, we will suffer from the problem of high rejection rate, which slows down the sample process, if not simply making it utterly impractical. Now we introduce a latent variable v to build the joint density:

$$f(\theta, v) \propto \theta^{a-1} e^{-b\theta} \cdot v^{d-1} 1(v < 1 - e^{-c\theta}) \tag{4.23}$$

It can be readily checked that $\int f(\theta, v) dv = f(\theta)$. This means that if we were to sample $\{(\theta_1, v_1), ..., (\theta_M, v_M)\}$ from $f(\theta, v)$, then the M θ's will be the sample as sampling directly from $f(\theta)$, since the v's can be easily marginalized[22].

Now, we show that it is simple to sample $\{(\theta_1, v_1), ..., (\theta_M, v_M)\}$ from the joint density $f(\theta, v)$ using a Gibbs Sampler:

$$\xi(\theta_2, v_2 | \theta_1, v_1) = f(v_2|\theta_2) \cdot f(\theta_2|v_1) \tag{4.24}$$

$$f(v|\theta) \propto v^{d-1} \cdot 1(v < 1 - e^{-c\theta}) \tag{4.25}$$

$$f(\theta|v) \propto \theta^{a-1} e^{-b\theta} \cdot 1(\theta > -\frac{1}{c} log(1 - v)) \tag{4.26}$$

It is immediately clear that both $f(v|\theta)$ and $f(\theta|v)$ are "named" densities, with the former being a truncated[23] exponential, and the latter a truncated gamma. Therefore, we can set up a Gibbs Sampler as described in section 4.3 without much trouble.

Unlike MH-sampler, however, the latent variable and the setup of the joint density depends on the particular densities from which we wish to sample. Nevertheless, it is usually quite straight forward by using a uniform distribution as demostrated above.

[22] This relates to the basics in probability theory, and the explanation will be rather lengthy, therefore we will not elaborate on it to avoid digression.

[23] This basically means it is a regular "named" density with its domain restricted. In the case of $f(v|\theta)$, for instance, the domain of v is $v < 1 - e^{-c\theta}$, where θ is simply constant in this case.

5 Time Series Models

Suppose we are interested in the transitional behavior of a series of 0/1 values, and have the following transitional probabilities.

$$\mathcal{T} = \begin{bmatrix} 1-p & p \\ 1-q & q \end{bmatrix} \tag{5.1}$$

The $\mathcal{T}$ in Eq. 5.1 is known as a *stochastic/transition matrix*. In this particular example, we know from the matrix that the probabilities of the transition from $0 \to 0$, $0 \to 1$, $1 \to 0$, and $1 \to 1$ are $1-p$, p, $1-q$ and q, respectively. Such a transitional series, where the value/state of a variable depends on the previous states, is referred as a Time Series. Therefore, if we have, say 18 observations of the values: 011000111101011000, we will be able to compute its likelihood by counting the occurence of each type of transition and use the following function[24]:

$$l(p,q) = p^4(1-p)^4q^5(1-q)^4 \tag{5.2}$$

In general, the likelihood density of a $n+1$-observation series of 0/1 values will be:

$$l(p,q) = p^{n_{01}}(1-p)^{n_{00}}q^{n_{11}}(1-q)^{n_{10}} \tag{5.3}$$
$$\text{where } n_{00}+n_{01}+n_{10}+n_{11} = n$$

As Bayesians, we consider the prior knowledge on the transitional behavior. For the binary example, we may set two beta densities $\pi(p) \sim Be(p|a,b)$ and $\pi(q) \sim Be(q|c,d)$ to serve as our priors. Conjugating the priors to the likelihood density, we get two posterior densities:

$$\pi_n(p|q,X) \sim Be(p|a+n_{01}, b+n_{00}) \tag{5.4}$$
$$\pi_n(q|p,X) \sim Be(q|c+n_{11}, d+n_{10}) \tag{5.5}$$

To sample from the joint density of the two betas, therefore, we simply employ a Gibbs Sampler by starting with an initializer (e.g. 0 or 1), and taking turns to sample from Eq. 5.4 and 5.5 (cf. section 4.3).

To make use of the posterior density to make predictions (cf. section 1.2), e.g. compute $P(x_{n+1}=1|x_0,...,x_n)$ (i.e. the probability of the $n+1$ th value being 1) as follows:

$$P(x_{n+1}=1|x_0,...,x_n) = \int \begin{Bmatrix} q \text{ if } x_n = 1 \\ p \text{ if } x_n = 0 \end{Bmatrix} \cdot \pi_n(p,q)dpdq \tag{5.6}$$

In general, we would like to study time series with M states, for which we need to modify our stochastic matrix and prior density. The general-case stochastic

[24] Note that there are 17 transitions in total.

matrix with M states is as follows, where p_{ij} is the probability of the transition from state i to state j:

$$\mathcal{T} = \begin{bmatrix} p_{11} & \cdots & p_{1M} \\ \vdots & \ddots & \vdots \\ p_{M1} & \cdots & p_{MM} \end{bmatrix} \tag{5.7}$$

Notice the sum of each row in Eq. 5.7 is equal to 1, because the sum of the probabilities of a state i transitioning to j where $j = 1, ..., M$ must be equal to 1. Since now we have M possible states where $M > 2$, the beta distribution will not suffice. To generalize the beta distribution, we describe the prior density for the M-state time series with a *Dirichlet Distribution*:

$$\begin{aligned} \pi(p) &= \frac{\Gamma(\sum_{i=1}^{M} \alpha_i)}{\prod_{i=1}^{M} \Gamma(\alpha_i)} {p_{11}}^{\alpha_{11}-1} ... {p_{1M-1}}^{\alpha_{1M-1}-1} (1 - p_{11} - ... - p_{1M-1})^{\alpha_{1M}-1} \\ &\propto \prod_{i=1}^{M} \prod_{j=1}^{M} {p_{ij}}^{\alpha_{ij}-1} \end{aligned} \tag{5.8}$$

The likelihood in the general case will be as follows, where n_{ij} is the number of i to j transitions:

$$l(p) \propto \prod_{i=1}^{M} \prod_{j=1}^{M} {p_{ij}}^{n_{ij}} \tag{5.9}$$

The posterior density is thus:

$$\pi_n(p) \propto l(p) \cdot \pi(p) \propto \prod_{i=1}^{M} \prod_{j=1}^{M} {p_{ij}}^{\alpha_{ij}+n_{ij}-1} \tag{5.10}$$

The predictive is generalized accordingly following Eq. 5.6.