y. Taking an example, as for the tri-variate random vector (X, Y, Z), if we generate the ith random draws from $f_{x|yz}(x|y_{i-1}, z_{i-1})$, $f_{y|xz}(y|x_i, z_{i-1})$ and $f_{z|xy}(z|x_i, y_i)$, sequentially, we can obtain the random draws from $f_{xyz}(x, y, z)$.

Remark 2: Let X, Y and Z be the random variables. Take an example of the case where X is highly correlated with Y. If we generate random draws from $f_{x|yz}(x|y, z)$, $f_{y|xz}(y|x, z)$ and $f_{z|xy}(z|x, y)$, it is known that convergence of the Gibbs sampler is slow. In this case, without separating X and Y, random number generation from $f(x, y|z)$ and $f(z|x, y)$ yields better random draws from the joint density $f(x, y, z)$.

Example: $X_1, X_2, \cdots, X_n$ are mutually independent with $X_i \sim N(\mu, \sigma^2)$.
Derive Bayesian estimation of μ and σ^2.
Assume that the prior distributions: $\mu \sim N(\mu_0, \sigma_0^2)$ and $\sigma^2 \sim IG(\alpha_0, \beta_0)$, i.e.,

$$\frac{1}{\sigma^2} \sim G(\alpha_0, \beta_0).$$

(*) Note that the gamma distribution $G(\alpha, \beta)$ is given by:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

for $x \geq 0$, $\alpha > 0$ and $\beta > 0$. When $X \sim G(\alpha, \beta)$ and $Y = \frac{1}{X}$, then $Y \sim IG(\alpha, \beta)$. the inverse gamma distribution is:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha) x^{\alpha+1}} \exp(-\frac{1}{\beta x})$$

The prior distribution of σ^2 is:

$$f(\sigma^2) = \frac{1}{\beta_0^{\alpha_0} \Gamma(\alpha_0) (\sigma^2)^{\alpha_0+1}} \exp(-\frac{1}{\beta \sigma^2})$$

Note that the posterior distributions are:

$$
\begin{aligned}
f(\theta|x) &= \frac{f(x,\theta)}{\int f(x,\theta)\mathrm{d}\theta} \propto f(x,\theta) \\
&\propto \begin{cases} f(x,\theta_1|\theta_2)f(\theta_2) \propto f(\theta_1|\theta_2,x) \\ f(x,\theta_2|\theta_1)f(\theta_1) \propto f(\theta_2|\theta_1,x) \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
f(x,\theta) = f(x,\mu,\sigma^2) &= f(x|\mu,\sigma^2)f(\mu)f(\sigma^2) \\
&= (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big) \\
&\times (2\pi\sigma_0^2)^{-1/2}\exp\Big(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\Big) \\
&\times \frac{1}{\beta_0^{\alpha_0}\Gamma(\alpha_0)(\sigma^2)^{\alpha_0+1}}\exp(-\frac{1}{\beta_0\sigma^2})
\end{aligned}
$$

The conditional distribution of μ given σ^2 and x is:

$$
\begin{aligned}
f(\mu|\sigma^2, x) &\propto (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big) \\
&\times (2\pi\sigma_0^2)^{-1/2} \exp\Big(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\Big) \\
&= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2)\Big) \\
&\times (2\pi\sigma_0^2)^{-1/2} \exp\Big(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\Big) \\
&\propto \exp\Big(-\frac{1}{2\sigma^2/n}(\mu - \overline{x})^2\Big) \times \exp\Big(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big(\frac{1}{\sigma^2/n}(\mu^2 - 2\overline{x}\mu) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu_0\mu)\Big)\Big)
\end{aligned}
$$

We focus on the parenthesis in the exponential part.

$$
\frac{1}{\sigma^2/n}(\mu^2 - 2\overline{x}\mu) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu_0\mu)
$$

$$= (\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2})(\mu - \frac{\bar{x}\sigma_0^2 + \mu_0\sigma^2/n}{\sigma^2/n + \sigma_0^2})^2 + \ ...$$

That is,

$$\mu|\sigma^2, x \ \sim N\Big(\frac{\bar{x}\sigma_0^2 + \mu_0\sigma^2/n}{\sigma^2/n + \sigma_0^2},\ (\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2})^{-1}\Big)$$

The conditional distribution of σ^2 given μ and x is:

$$\begin{aligned} f(\sigma^2|\mu, x) &\propto (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big) \\ &\times \frac{1}{\beta_0^{\alpha_0}\Gamma(\alpha_0)(\sigma^2)^{\alpha_0+1}} \exp(-\frac{1}{\beta_0\sigma^2}) \\ &\propto \frac{1}{(\sigma^2)^{n/2+\alpha_0+1}} \exp\Big(-(\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{\beta_0})\frac{1}{\sigma^2}\Big) \end{aligned}$$

which is $IG\Big(\frac{n}{2} + \alpha_0,\ (\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{\beta_0})^{-1}\Big)$, i.e.,

$$\frac{1}{\sigma^2}|\mu, x \ \sim G\Big(\frac{n}{2} + \alpha_0,\ (\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{\beta_0})^{-1}\Big)$$

6.2.2 Metropolis-Hastings Algorithm

This section is based on Geweke and Tanizaki (2003), where three sampling distributions are compared with respect to precision of the random draws from the target density $f(x)$.

The **Metropolis-Hastings algorithm** is also one of the sampling methods to generate random draws from any target density $f(x)$, utilizing sampling density $f_*(x)$, even in the case where it is not easy to generate random draws from the target density.

Let us define the acceptance probability by:

$$\omega(x_{i-1}, x^*) = \min\Big(\frac{q(x^*)}{q(x_{i-1})}, 1\Big) = \min\Big(\frac{f(x^*)/f_*(x^*)}{f(x_{i-1})/f_*(x_{i-1})}, 1\Big),$$

where $q(\cdot)$ is defined as equation (**??**). By the Metropolis-Hastings algorithm, a random draw from $f(x)$ is generated in the following way:

(i) Take the initial value of x as x_{-M}.

(ii) Generate x^* from $f_*(x)$ and compute $\omega(x_{i-1}, x^*)$ given x_{i-1}.

(iii) Set $x_i = x^*$ with probability $\omega(x_{i-1}, x^*)$ and $x_i = x_{i-1}$ otherwise.

(iv) Repeat Steps (ii) and (iii) for $i = -M+1, -M+2, \cdots, 1$.

In the above algorithm, x_1 is taken as a random draw from $f(x)$. When we want more random draws (say, N), we replace Step (iv) by Step (iv)', which is represented as follows:

(iv)' Repeat Steps (ii) and (iii) for $i = -M+1, -M+2, \cdots, N$.

When we implement Step (iv)', we can obtain a series of random draws x_{-M}, x_{-M+1}, $\cdots$, x_0, x_1, x_2, $\cdots$, x_N, where x_{-M}, x_{-M+1}, $\cdots$, x_0 are discarded from further consideration. The last N random draws are taken as the random draws generated from the target density $f(x)$. Thus, N denotes the number of random draws. M is sometimes called the **burn-in period**.

We can justify the above algorithm given by Steps (i) – (iv) as follows. We show that x_i is the random draw generated from the target density $f(x)$ under the assumption x_{i-1} is generated from $f(x)$. Let U be the uniform random variable between zero and one, X be the random variable which has the density function $f(x)$ and x^* be the realization (i.e., the random draw) generated from the sampling density $f_*(x)$. Consider the probability $P(X \leq x | U \leq \omega(x_{i-1}, x^*))$, which should be the cumulative distribution of X, i.e., $F(x)$. The probability $P(X \leq x | U \leq \omega(x_{i-1}, x^*))$ is rewritten as follows:

$$P(X \leq x | U \leq \omega(x_{i-1}, x^*)) = \frac{P(X \leq x, U \leq \omega(x_{i-1}, x^*))}{P(U \leq \omega(x_{i-1}, x^*))},$$

where the numerator is represented as:

$$\begin{aligned} P(X \leq x, U \leq \omega(x_{i-1}, x^*)) &= \int_{-\infty}^{x} \int_{0}^{\omega(x_{i-1}, t)} f_{u,*}(u, t) \, \mathrm{d}u \, \mathrm{d}t \\ &= \int_{-\infty}^{x} \int_{0}^{\omega(x_{i-1}, t)} f_u(u) f_*(t) \, \mathrm{d}u \, \mathrm{d}t = \int_{-\infty}^{x} \Big(\int_{0}^{\omega(x_{i-1}, t)} f_u(u) \, \mathrm{d}u \Big) f_*(t) \, \mathrm{d}t \end{aligned}$$

$$= \int_{-\infty}^{x} \Big(\int_{0}^{\omega(x_{i-1},t)} \mathrm{d}u \Big) f_*(t)\, \mathrm{d}t = \int_{-\infty}^{x} \Big[u \Big]_{0}^{\omega(x_{i-1},t)} f_*(t)\, \mathrm{d}t$$
$$= \int_{-\infty}^{x} \omega(x_{i-1}, t) f_*(t)\, \mathrm{d}t = \int_{-\infty}^{x} \frac{f_*(x_{i-1}) f(t)}{f(x_{i-1})}\, \mathrm{d}t = \frac{f_*(x_{i-1})}{f(x_{i-1})} F(x)$$

and the denominator is given by:

$$P(U \leq \omega(x_{i-1}, x^*)) = P(X \leq \infty, U \leq \omega(x_{i-1}, x^*)) = \frac{f_*(x_{i-1})}{f(x_{i-1})} F(\infty) = \frac{f_*(x_{i-1})}{f(x_{i-1})}.$$

The density function of U is given by $f_u(u) = 1$ for $0 < u < 1$. Let X^* be the random variable which has the density function $f_*(x)$. In the numerator, $f_{u,*}(u, x)$ denotes the joint density of random variables U and X^*. Because the random draws of U and X^* are independently generated, we have $f_{u,*}(u, x) = f_u(u) f_*(x) = f_*(x)$. Thus, the first four equalities are derived. Substituting the numerator and denominator shown above, we have the following equality:

$$P(X \leq x | U \leq \omega(x_{i-1}, x^*)) = F(x).$$

Thus, the x^* which satisfies $u \leq \omega(x_{i-1}, x^*)$ indicates a random draw from $f(x)$. We set $x_i = x_{i-1}$ if $u \leq \omega(x_{i-1}, x^*)$ is not satisfied. x_{i-1} is already assumed to be a random draw from $f(x)$. Therefore, it is shown that x_i is a random draw from $f(x)$. See Gentle (1998) for the discussion above.

As a general formulation of the sampling density, instead of $f_*(x)$, we may take the sampling density as the following form: $f_*(x|x_{i-1})$, where a candidate random draw x^* depends on the $(i-1)$th random draw, i.e., x_{i-1}.

For choice of the sampling density $f_*(x|x_{i-1})$, Chib and Greenberg (1995) pointed out as follows. $f_*(x|x_{i-1})$ should be chosen so that the chain travels over the support of $f(x)$, which implies that $f_*(x|_{i-1})$ should not have too large variance and too small variance, compared with $f(x)$. See, for example, Smith and Roberts (1993), Bernardo and Smith (1994), O'Hagan (1994), Tierney (1994), Geweke (1996), Gamerman (1997), Robert and Casella (1999) and so on for the Metropolis-Hastings algorithm.

As an alternative justification, note that the Metropolis-Hastings algorithm is formulated as follows:

$$f_i(u) = \int f^*(u|v) f_{i-1}(v)\, \mathrm{d}v,$$

where $f^*(u|v)$ denotes the transition distribution, which is characterized by Step (iii). x_{i-1} is generated from $f_{i-1}(\cdot)$ and x_i is from $f^*(\cdot|x_{i-1})$. x_i depends only on x_{i-1}, which is called the **Markov property**. The sequence $\{\cdots, x_{i-1}, x_i, x_{i+1}, \cdots\}$ is called the **Markov chain**. The Monte Carlo statistical methods with the sequence $\{\cdots, x_{i-1}, x_i, x_{i+1}, \cdots\}$ is called the **Markov chain Monte Carlo (MCMC)**. From Step (iii), $f^*(u|v)$ is given by:

$$f^*(u|v) = \omega(v,u) f_*(u|v) + \Big(1 - \int \omega(v,u) f_*(u|v)\, \mathrm{d}u\Big) p(u), \tag{1}$$

where $p(x)$ denotes the following probability function:

$$p(u) = \begin{cases} 1, & \text{if } u = v, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, x is generated from $f_*(u|v)$ with probability $\omega(v, u)$ and from $p(u)$ with probability $1 - \int \omega(v, u) f_*(u|v)\, \mathrm{d}u$. Now, we want to show $f_i(u) = f_{i-1}(u) = f(u)$ as i goes to infinity, which implies that both x_i and x_{i-1} are generated from the invariant distribution function $f(u)$ for sufficiently large i. To do so, we need to consider the condition satisfying the following equation:

$$f(u) = \int f^*(u|v) f(v)\, \mathrm{d}v. \tag{2}$$

Equation (2) holds if we have the following equation:

$$f^*(v|u) f(u) = f^*(u|v) f(v), \tag{3}$$

which is called the **reversibility condition**. By taking the integration with respect to v on both sides of equation (3), equation (2) is obtained. Therefore, we have to check whether the $f^*(u|v)$ shown in equation (1) satisfies equation (3). It is straightforward to verify that

$$\omega(v,u)f_*(u|v)f(v) = \omega(u,v)f_*(v|u)f(u),$$
$$\Big(1 - \int \omega(v,u)f_*(u|v)\,\mathrm{d}u\Big)p(u)f(v) = \Big(1 - \int \omega(u,v)f_*(v|u)\,\mathrm{d}v\Big)p(v)f(u).$$

Thus, as i goes to infinity, x_i is a random draw from the target density $f(\cdot)$. If x_i is generated from $f(\cdot)$, then x_{i+1} is also generated from $f(\cdot)$. Therefore, all the x_i, x_{i+1}, x_{i+2}, $\cdots$ are taken as random draws from the target density $f(\cdot)$.

The requirement for uniform convergence of the Markov chain is that the chain should be **irreducible** and **aperiodic**. See, for example, Roberts and Smith (1993). Let $C_i(x_0)$ be the set of possible values of x_i from starting point x_0. If there exist two

possible starting values, say x^* and x^{**}, such that $C_i(x^*) \cap C_i(x^{**}) = \emptyset$ (i.e., empty set) for all i, then the same limiting distribution cannot be reached from both starting points. Thus, in the case of $C_i(x^*) \cap C_i(x^{**}) = \emptyset$, the convergence may fail. A Markov chain is said to be **irreducible** if there exists an i such that $P(x_i \in C|x_0) > 0$ for any starting point x_0 and any set C such that $\int_C f(x)\,\mathrm{d}x > 0$. The irreducible condition ensures that the chain can reach all possible x values from any starting point. Moreover, as another case in which convergence may fail, if there are two disjoint set C^1 and C^2 such that $x_{i-1} \in C^1$ implies $x_i \in C^2$ and $x_{i-1} \in C^2$ implies $x_i \in C^1$, then the chain oscillates between C^1 and C^2 and we again have $C_i(x^*) \cap C_i(x^{**}) = \emptyset$ for all i when $x^* \in C^1$ and $x^{**} \in C^2$. Accordingly, we cannot have the same limiting distribution in this case, either. It is called **aperiodic** if the chain does not oscillate between two sets C^1 and C^2 or cycle around a partition $C^1, C^2, \cdots, C^r$ of r disjoint sets for $r > 2$. See O'Hagan (1994) for the discussion above.

For the Metropolis-Hastings algorithm, x_1 is taken as a random draw of x from $f(x)$ for sufficiently large M. To obtain N random draws, we need to generate $M + N$ random draws. Moreover, clearly we have $\mathrm{Cov}(x_{i-1}, x_i) > 0$, because x_i is generated based on x_{i-1} in Step (iii).

Based on Steps (i) – (iii) and (iv)', under some conditions the basic result of the Metropolis-Hastings algorithm is as follows:

$$\frac{1}{N}\sum_{i=1}^{N} g(x_i) \longrightarrow \mathrm{E}(g(x)) = \int g(x)f(x)\,\mathrm{d}x, \qquad \text{as } N \longrightarrow \infty,$$

where $g(\cdot)$ is a function, which is representatively taken as $g(x) = x$ for mean and $g(x) = (x - \overline{x})^2$ for variance. $\overline{x}$ denotes $\overline{x} = (1/N)\sum_{i=1}^{N} x_i$. Thus, it is shown that $(1/N)\sum_{i=1}^{N} g(x_i)$ is a consistent estimate of $\mathrm{E}(g(x))$, even though $x_1, x_2, \cdots, x_N$ are mutually correlated.

As an alternative random number generation method to avoid the positive correlation,

we can perform the case of $N = 1$ as in the above procedures (i) – (iv) N times in parallel, taking different initial values for x_{-M}. In this case, we need to generate $M + 1$ random numbers to obtain one random draw from $f(x)$. That is, N random draws from $f(x)$ are based on $N(1 + M)$ random draws from $f_*(x|x_{i-1})$. Thus, we can obtain mutually independently distributed random draws. For precision of the random draws, the alternative Metropolis-Hastings algorithm should be similar to rejection sampling. However, this alternative method is too computer-intensive, compared with the above procedures (i) – (iii) and (iv)', which takes more time than rejection sampling in the case of $M > N_R$.

Furthermore, the sampling density has to satisfy the following conditions: (i) we can quickly and easily generate random draws from the sampling density and (ii) the sampling density should be distributed with the same range as the target density. See, for example, Geweke (1992) and Mengersen, Robert and Guihenneuc-Jouyaux

(1999) for the MCMC convergence diagnostics. Since the random draws based on the Metropolis-Hastings algorithm heavily depend on choice of the sampling density, we can see that the Metropolis-Hastings algorithm has the problem of specifying the sampling density, which is the crucial criticism. Several generic choices of the sampling density are discussed by Tierney (1994) and Chib and Greenberg (1995). We can consider several candidates for the sampling density $f_*(x|x_{i-1})$, i.e., Sampling Densities Iand II.

Sampling Density I (Independence Chain) For the sampling density, we have started with $f_*(x)$ in this section. Thus, one possibility of the sampling density is given by: $f_*(x|x_{i-1}) = f_*(x)$, where $f_*(\cdot)$ does not depend on x_{i-1}. This sampling density is called the **independence chain**. For example, it is possible to take $f_*(x) = N(\mu_*, \sigma_*^2)$, where μ_* and σ_*^2 are the hyper-parameters. Or, when x lies on a certain

interval, say (a, b), we can choose the uniform distribution $f_*(x) = 1/(b-a)$ for the sampling density.

Sampling Density II (Random Walk Chain) We may take the sampling density called the **random walk chain**, i.e., $f_*(x|x_{i-1}) = f_*(x - x_{i-1})$. Representatively, we can take the sampling density as $f_*(x|x_{i-1}) = N(x_{i-1}, \sigma_*^2)$, where σ_*^2 denotes the hyper-parameter. Based on the random walk chain, we have a series of the random draws which follow the random walk process.