Introduction to Bayesian Econometrics

Gibbs Sampling and Metropolis-Hasting Sampling

Tao Zeng

Wuhan University

Dec 2016

- In statistics, there is a distinction between two concepts of probability, **frequentist probability** and **subjective probability.**
- **Frequentists** restrict the assignment of probabilities to statements that describe the outcome of an experiment that can be repeated.

Example

A coin tossed three times will come up heads either two or three times. We can imagine repeating the experiment of tossing a coin three times and recording the number of times that two or three heads were reported.

$$Pr(A_1) = \lim_{n \longrightarrow \infty} \frac{\text{number of times two or three heads coocurs}}{n}.$$

Fact

Those who take the subjective view of probability believe that probability theory is applicable to any situation in which there is uncertainty.

- Outcomes of repeated experiments fall in that category, but so do statements about tomorrow's weather, which are not the outcomes of repeated experiments.
- Calling probabilities 'subjective' does not imply that they can be set arbitrarily, and probabilities set in accordance with the axioms are consistent.

Example

(**Subjective view of probability**) Let Y a binary variable with Y = 1 if a coin toss results in a head and 0 otherwise, and let

$$\begin{aligned} \Pr(Y=1) &= \theta \\ \Pr(Y=0) &= 1-\theta \end{aligned}$$

which is assummed to be constant for each trial. In this model, θ is a parameter and the value of Y is the data (realisation y).

- From the frequentist point of view, probability theory can tell us something about the distribution of the data for a given θ.
- It is not given a probability distribution of θ, since it is not regarded as being the outcome of a repeated expriment.

- In a frequentist approach, the parameters θ are considered as constant terms and the aim is to study the distribution of the data given θ, through the likelihood of the sample.
- The likelihood of the sample $(y_1, \cdots y_n)$ is

$$L_n(\theta; y_1, \cdots, y_n) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}.$$

- From the subjective point of view, however, θ is an **unknown quantity.**
- Since there is uncetainty over its value, it can be regarded as a random variable and assigned a prior distribution.
- Before seeing the data, it is assigned a **prior distribution**

$$\pi(\theta) \text{ with } 0 \leq \theta \leq 1.$$

Definition

Prior distribution In a Bayesian framework, the parameters θ associated to the distribution of the data, are considered as random variables. Their distribution is called the prior distribution and is denoted by $\pi(\theta)$.

- In most of cases, the prior distribution is parametrised, i.e. the pdf $\pi(\theta; \gamma)$ depends on a set of parameters γ where γ are the parameters of the prior distribution, called **hypeparameters**.

Example

(Hyperparameters) If $\theta \in R$ and if the prior distribution is normal

$$\pi(\theta; \gamma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta - \mu)^2}{2\sigma^2}\right)$$

with $\gamma = (\mu, \sigma^2)$ the vector of hyperparameters.

Example

(Beta prior distribution) If $\theta \in [0,1]$, a common (parametrised) prior distribution is the Beta distribution denoted $B(\alpha, \beta)$.

$$\pi(\theta; \gamma) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad \alpha, \beta > 0 \; \theta \in [0,1]$$

with $\gamma = (\alpha, \beta)^T$ the vector of hyper parameters.

- Depending on the choice of α and β, this prior can capture beliefs that indicate θ is centered at $1/2$, or it can shade θ toward zero or one.

Definition

(Posterior distribution) Bayesian inference centers on the posterior distribution $\pi\left(\theta|y\right)$, which is the conditional distribution of the random variable θ given the data (realisation of the sample) $y = \left(y_1, \ldots y_n\right)$.

$$\theta|\left(Y_1 = y_1, \ldots Y_n = y_n\right) \sim \text{posterior distrbution}$$

Theorem

(Bayes Theorem) For events A and B, the conditional probability of event A and given that B has occurred is

$$\Pr\left(A|B\right) = \frac{\Pr\left(B|A\right) \times \Pr\left(A\right)}{\Pr\left(B\right)}$$

Definition

For one observation y_i, the posterior distribution is the conditional distribution of θ given y_i, defined as to be

$$\pi\left(\theta|y_i\right) = \frac{f_{Y_i|\theta}\left(y_i|\theta\right) \times \pi\left(\theta\right)}{f_{Y_i}\left(y_i\right)}$$

where

$$f_{Y_i}\left(y_i\right) = \int_{\Theta} f_{Y_i|\theta}\left(y_i|\theta\right) \times \pi\left(\theta\right) d\theta$$

and Θ the support of the distribution of θ, where the term $f_{Y_i|\theta}\left(y_i|\theta\right)$ corresponds to the likelihood of the observation y_i,

$$f_{Y_i|\theta}\left(y_i|\theta\right) = L_i\left(\theta; y_i\right).$$

Definition

(**Posterior distribution, sample**) For sample $(y_1, \ldots, y_n)$, the posterior distribution is the conditional distribution of θ given y_i, defined as to be

$$\pi\left(\theta | y_1, \ldots, y_n\right) = \frac{L_n\left(\theta; y_1, \ldots, y_n\right) \times \pi\left(\theta\right)}{f_{Y_1, \ldots Y_n}\left(y_1, \ldots y_n\right)}$$

where $L_n\left(\theta; y_1, \ldots, y_n\right)$ is the likelihood of the sample and

$$f_{Y_1, \ldots Y_n}\left(y_1, \ldots y_n\right) = \int_{\Theta} L_n\left(\theta; y_1, \ldots, y_n\right) \times \pi\left(\theta\right) d\theta$$

and Θ the support of the distribution of θ.

- In this setting, the data $(y_1, \dots y_n)$ are viewed as constants whose marginal distributions do not involve the parameters of interest θ, that is

$$f_{Y_1,\dots Y_n}(y_1, \dots y_n) = \text{constant}.$$

Definition

(**Unormalised posterior distribution**) The unormalised posterior distribution is the product of the likelihood of the sample and the prior distribution:

$$\pi(\theta | y_1, \dots, y_n) \propto L_n(\theta; y_1, \dots, y_n) \times \pi(\theta)$$

or with simplified form

$$\pi(\theta | y) \propto L_n(\theta; y) \times \pi(\theta)$$

where the symbol "$\propto$" means "is proportional to".

Example

(Beta distribution) Consider an i .i .d. sample $(Y_1, \dots Y_n)$ of binary variables, such that $Y_i \sim Be(\theta)$ and:

$$f_{Y_i}(y_i;\theta) = \Pr(Y_i = y_i) = \theta^{y_i}(1-\theta)^{1-y_i},$$

We assume that the uninformative prior distribution for θ is an Beta $B(\alpha, \beta)$ with a pdf

$$\pi(\theta;\gamma) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \alpha, \beta > 0 \quad \theta \in [0,1]$$

with $\gamma = (\alpha, \beta)^T$ the vector of hyperparameters.

Example

(Beta distribution, cont) The likelihood of the sample $(y_1, \ldots, y_n)$ is

$$L_n(\theta; y_1, \ldots, y_n) = \theta^{\sum y_i} (1-\theta)^{\sum(1-y_i)},$$

hence the unormalised posterior distribution is

$$\begin{aligned}
\pi(\theta | y_1, \ldots, y_n) &\propto L_n(\theta; y_1, \ldots, y_n) \times \pi(\theta) \\
&= \theta^{\sum y_i} (1-\theta)^{\sum(1-y_i)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\
&\propto \theta^{\sum y_i} (1-\theta)^{\sum(1-y_i)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\
&= \theta^{(\alpha+\sum y_i)-1} (1-\theta)^{(\beta+\sum(1-y_i))-1}.
\end{aligned}$$

Example

(Beta distribution, cont) Recall that the pdf of a Beta distribution is

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

The posterior distribution is in the form of a Beta distribution with parameters

$$\alpha_1 = \alpha + \sum y_i \quad \beta_1 = \beta + n - \sum y_i \ .$$

This is an example of a **conjugate prior**, where the posterior distribution is in the same family as the prior distribution.

Example

(Beta distribution, cont) Note that

$$E\left(\theta | y_1, \ldots, y_n\right) = \frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{\alpha + \sum y_i}{\alpha + \beta + n}$$

which can be expressed as a function of the MLE estimator $\bar{y}_n = n^{-1} \sum_{i=1}^{N} y_i$ as follows

$$\begin{aligned}
\underbrace{E\left(\theta | y_1, \ldots, y_n\right)}_{\text{posterior mean}} &= \frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{\alpha}{\alpha + \beta + n} + \frac{\sum y_i}{\alpha + \beta + n} \\
&= \frac{\alpha + \beta}{\alpha + \beta + n} \underbrace{\frac{\alpha}{\alpha + \beta}}_{\text{prior mean}} + \frac{n}{\alpha + \beta + n} \underbrace{\bar{y}_n}_{\text{MLE}}.
\end{aligned}$$

Prior and posterior distribution
Posterior distribution

- If $n \longrightarrow \infty$, then the weight on the prior mean approaches zero, and the weight on the MLE approaches one, implying

$$\lim_{n \longrightarrow \infty} E\left(\theta | y_1, \ldots, y_n\right) = \bar{y}_n.$$

- If the sample size is very small, $n \longrightarrow 0$, then we have

$$\lim_{n \longrightarrow 0} E\left(\theta | y_1, \ldots, y_n\right) = \frac{\alpha}{\alpha + \beta}.$$

- Bayesian updating

$$\pi\left(\theta | y_1, y_2\right) \propto f\left(y_1, y_2 | \theta\right) \pi\left(\theta\right) = f\left(y_2 | y_1, \theta\right) \pi\left(\theta | y_1\right)$$

- As new information is required, the posterior distribution becomes the prior for the next experiment.

Prior and posterior distribution

Posterior distribution - some intuition of prior

- α can be interpreted as "the number of heads obtained in the expriment on which the prior is based".
- If, for example, you had seen this coin tossed a large number of times and heads appeared frequently, we can set a large number of α.
- $\alpha = \beta = 1$ yields uniform distribution which indicates that both head and tail can appear but otherwise have no strong opnion about the distribution of θ.

- In many cases, although we can always get the analytical form of the posterior density up to a constant, the characteristic of the density, such as **mean**, **variance**, **median**, are not easy to compute.
- Generate random sample from the posterior distribution to approximate these characteristics, such as the posterior mean

$$\widehat{E\left(\theta\right)} = \frac{1}{M}\sum_{m=1}^{M}\theta^{(m)}, \ \ m = 1, 2, 3, \ldots M$$

  where $\theta^{(m)}$ is the random sample from the posterior density $\pi\left(\theta|y\right)$.
- **Monte Carlo** - draw the random samples **identically** and **independently**.
- **Markov Chain Monte Carlo** - draw the random samples **dependently.**

- Suppose we wish to draw a sample of values from a random variable with d.f. $F(\cdot)$ which is nondecreasing.
- Consider the distribution Z, which is obtained by drawing U from $U(0,1)$ and setting $Z = F^{-1}(U)$, then $U = F(Z)$

$$P(Z \leq z) = P(F(Z) \leq F(z)) = P(U \leq F(z)) = F(z).$$

- **Probability integral transform method:**
  1. Draw u from $U(0,1)$.
  2. Return $y = F^{-1}(u)$ as a draw from $f(y)$.
- Requires that $F(\cdot)$ be known (including constant) and $F^{-1}(\cdot)$ can be readily computed.

- $f(\cdot)$ is diffcult to simulate but it is possible to simulate values from $g(\cdot)$ and a number $c \geq 1$ can be found such that $f(Y) \leq cg(Y)$ for all Y in the support of $f(\cdot)$.
- Accepted-Reject Algorithm
  1. Generate a value y from $g(\cdot)$.
  2. Draw a value u from $U(0,1)$.
  3. Return y as a draw from $f(\cdot)$ if $u \leq f(y)/cg(y)$. If not, reject it and return to step 1.
- The density $f(\cdot)$ is only need to be known up to a constant.

Proof.

Consider the distribution of the accepted values of y, $h\left[y|u \leq f\left(y\right)/cg\left(y\right)\right]$, we have

$$\begin{aligned} h\left[y|u \leq f\left(y\right)/cg\left(y\right)\right] &= \frac{P\left[u \leq f\left(y\right)/cg\left(y\right)\right]g\left(y\right)}{\int P\left[u \leq f\left(y\right)/cg\left(y\right)\right]g\left(y\right)dy} \\ &= \frac{f\left(y\right)/cg\left(y\right)g\left(y\right)}{\int f\left(y\right)/cg\left(y\right)g\left(y\right)dy} = f\left(y\right) \end{aligned}$$

$\square$

- Note that
$$\int P\left[u \leq f\left(y\right)/cg\left(y\right)\right]g\left(y\right)dy = 1/c$$
is the probality that a generated value of y is accepted.

- MC methods are not easy to implemented in multivariate case.
- For AR method, it is difficult to find a suitable $g(\cdot)$.
- A sequence $X_1, X_2, \cdots$ of random variables is called **Markov Chain** if the conditional distribution of X_{n+1} given $X_1, \cdots, X_n$ depends on X_n only

$$P(X_{n+1}|X_n, \cdots, X_1) = P(X_{n+1}|X_n),$$

  for instance, the AR(1) proess.
- **Markov Chain Monte Carlo** (MCMC) is a class of algorithms that produce a chain of simulated draws from a distribution where each draw is **dependent** on the previous draw.

- A stochastic process X_t, takes the values in the finite set $S = \{1, 2, \ldots, s\}$.
- Define p_{ij} as the probability that $X_{t+1} = j$ given that $X_t = i$

$$p_{ij} = P\left(X_{t+1} = j | X_t = i\right), \ \ i, j \in S$$

  which is called transition probability, $\sum_{j=1}^{S} p_{ij} = 1$.
- The probability distribution at time $t + 1$ only depends on the system at t is called the Markov property, and the resulting process is a Markov process.

- Irreducible: starting from state i, the process can reach any other state with positive probability, a counter example

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$$

where P_1, P_2 are $m \times m$, strating from the first m states, it will never arrive the second m states.

- Aperiodic: starting from state i, the process can return *ith* state in one period, a counter example

$$P = \begin{bmatrix} 0 & P_1 \\ P_2 & 0 \end{bmatrix}$$

where strarting from the first m states, it takes 2 periods to return.

- Invariant distribution: The probability distribution $\pi^* = (\pi_1^*, \pi_2^*, \dots \pi_s^*)'$ is an invariant distribution for P if $\pi' = \pi' P$.

Example

If we set $P = \begin{pmatrix} 0.75 & 0.25 \\ 0.125 & 0.875 \end{pmatrix}$, from $\pi^{*\prime} = \pi^{*\prime} P$, we have

$$(\pi_1^*, \pi_2^*) = (\pi_1^*, \pi_2^*) \begin{pmatrix} 0.75 & 0.25 \\ 0.125 & 0.875 \end{pmatrix},$$

the solution is $\pi^{*\prime} = (1/3, 1/2)$.

Theorem

Suppose S is finite and $p_{ij} > 0$ for all i, j. Then there exists a unique probability distribution π_j^*, $j \in S$, such that $\sum_i \pi_i^* p_{ij} = \pi_j^*$ for all $j \in S$. Moreover,

$$\left| p_{ij}^{(n)} - \pi_j^* \right| \leq r^n,$$

where $0 < r < 1$, for all i, j and $n \geq 1$.

- For large enough n, the initial state i plays almost no role.
- P^n converges quickly to a matrix whose rows are all $\pi^{*\prime}$.
- If a Markov Chain satisfy some conditions, the probability distribution of its *nth* iterate is very close to its invariant distribution for large n.

- **If we can find a Markov process for which the invariant distribution is the target distribution, we can simulate draws from the process to generate values from the target distribution.**

Theorem

Let P be irreducible and aperiodic over a finite state space. Then there is a unique probability distribution π^* such that $\sum_i \pi_i^* p_{ij} = \pi_j^*$ for all $j \in S$ and

$$\left| p_{ij}^{(n)} - \pi_j^* \right| \leq r^{n/v},$$

for all $i, j \in S$, where $0 < r < 1$, for some positive integer v.

Definition

Transition Kernel: $K : S \times S \longrightarrow R_0^+$:

$$P\left(X_{t+1} \in A \middle| X_t = x_t\right) = \int_A K\left(x_{t+1} \middle| x_t\right) dx_{t+1}$$

for $A \in S$.

Definition

Invariant distribution: A distribution μ with density function f_μ is said to be the invariant distribution of a Markov chain X with transition kernel K if

$$f_\mu\left(y\right) = \int_S f_\mu\left(x\right) K\left(y \middle| x\right) dx$$

for almost all $y \in S$.

Example

(MCMC, Simple example) Suppose we want to sample from the following distribution

$$\pi\left(\theta\right)=\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1-\phi^{2}}{\sigma^{2}}}\exp\left(-\frac{\left(1-\phi^{2}\right)\left(\theta-\mu/\left(1-\phi\right)\right)^{2}}{2\sigma^{2}}\right)$$

where $|\phi|<1$ and pretend that we do not know how to draw i.i.d. samples from this distribution which means that the target ditribution is $N\left(\mu/\left(1-\phi\right),\sigma^{2}/\left(1-\phi^{2}\right)\right)$.

Example

(MCMC, simple example, cont) Then suppose we use the following transition kernel to generate draws

$$q\left(\theta_t|\theta_{t-1}\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(\theta_t - \mu - \phi\theta_{t-1}\right)^2}{2\sigma^2}\right),$$

that is θ_t belonging to an $AR\left(1\right)$ process

$$\theta_t = \mu + \phi\theta_{t-1} + \varepsilon_t$$

where $\varepsilon_t \sim N\left(0, \sigma^2\right)$. We can show that $\pi\left(\theta\right)$ is the invariant distribution of $q\left(\theta_t|\theta_{t-1}\right)$.

Example

(MCMC, simple example, cont) If θ_{t-1} is sampled from the target ditribution $\theta_{t-1} \sim N\left(\mu/\left(1-\phi\right), \sigma^2/\left(1-\phi^2\right)\right)$, then we can easily get that

$$\theta_t = \phi\theta_{t-1} + \varepsilon_t, \varepsilon_t \sim N\left(0, \sigma^2\right),$$

hence

$$\theta_t \sim N\left(\mu/\left(1-\phi\right), \sigma^2/\left(1-\phi^2\right)\right)$$

since

$$\begin{aligned} E\left(\theta_t\right) &= \phi E\left(\theta_{t-1}\right) + E\left(\varepsilon_t\right) = \mu/\left(1-\phi\right) \\ Var\left(\theta_t\right) &= \phi^2 Var\left(\theta_{t-1}\right) + Var\left(\varepsilon_t\right) = \sigma^2/\left(1-\phi^2\right). \end{aligned}$$

Hence π is the invariant distribution of $q\left(\theta_t|\theta_{t-1}\right)$.

- Gibbs sampling was proposed in the early 1990s (Geman and Geman, 1984; Gelfand and Smith, 1990) and fundamentally changed Bayesian computing.
- Gibbs sampling is attractive because it can sample from high-dimensional posteriors.
- The main idea is to break the problem of sampling from the high-dimensional joint distribution into a series of samples from low-dimensional conditional distributions.
- Because the low-dimensional updates are done in a loop, samples are not independent as in rejection sampling.
- The dependence of the samples turns out to follow a Markov distribution, leading to the name Markov chain Monte Carlo (MCMC).

- Gibbs Sampling is used to find the transition kernal which based on the condition that it is possible to sample from each conditional distribution.
- Gibbs algorithm with two blocks
  1. Choose a starting value $x_2^{(0)}$.
  2. At the first iteration, draw
  $$x_1^{(1)} \text{ from } f\left(x_1 | x_2^{(0)}\right),$$
  $$x_2^{(1)} \text{ from } f\left(x_2 | x_1^{(1)}\right).$$
  3. At the gth iteration, draw
  $$x_1^{(g)} \text{ from } f\left(x_1 | x_2^{(g-1)}\right),$$
  $$x_2^{(g)} \text{ from } f\left(x_2 | x_1^{(g-1)}\right).$$

- The Gibbs kernel is

$$p(x, y) = f(y_1|x_2)\, f(y_2|y_1),$$

from which we can compute

$$\begin{aligned} \int p(x, y)\, f(x)\, dx &= \int f(y_1|x_2)\, f(y_2|y_1)\, f(x_1, x_2)\, dx_1 dx_2 \\ &= f(y_2|y_1) \int f(y_1|x_2)\, f(x_1, x_2)\, dx_1 dx_2 \\ &= f(y_2|y_1)\, f(y_1) = f(y_1, y_2). \end{aligned}$$

hence $f(\cdot)$ is the invariant distribution for the Gibbs kernel.