MONASH University

ETC2420

Statistical methods in Insurance

Week 10.
Monte Carlo sampling methods

6 October 2016

Outline

Week	Topic	Lecturer
1	Randomization & Hypothesis Testing I	Souhaib & Di
2	Hypothesis Testing II & Decision Theory	Souhaib
3	Statistical Distributions	Di
4	Model fitting & Linear regression	Di
5	Linear models	Di
6	Bootstrap, Permutation and Linear models	Di
	Multilevel models	Di
7	Generalized Linear models	Di
8	Compiling data for problem solving	Di
9	Bayesian Reasoning I & II	Souhaib
10	Monte Carlo sampling methods I & II	Souhaib
10	Time series models I & II	Souhaib
11	Project presentation	Souhaib

References

- Berger, J. O. 2013. **Statistical Decision Theory and Bayesian Analysis**. Springer Series in Statistics. Springer New York.
- Robert, Christian, and George Casella. 2010. **Introducing Monte Carlo Methods with R**. Springer Science & Business Media.
- Bishop, Christopher M. 2006. **Pattern Recognition and Machine Learning**. Edited by M. Jordan, J. Kleinberg, and B. Scholkopf. Vol. 16. Springer.

Bayesian method

$X_1, \ldots, X_n \sim F_\theta$

$$\pi(\theta|x_1, \ldots, x_n) = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{f(x_1, \ldots, x_n)} \propto \mathcal{L}_n(\theta)\pi(\theta)$$

where

$$\mathcal{L}_n(\theta) = f(x_1, \ldots, x_n|\theta) = \Pi_{i=1}^n f(x_i|\theta)$$

and

$$f(x_1, \ldots, x_n) = \int_\Theta \mathcal{L}_n(\theta)\pi(\theta)d\theta = c_n$$

Bayesian method

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{Bernouilli}(p)$$

$$\hat{p}_{MLE} = \frac{s}{n}$$

$$p|x_1, \ldots, x_n \sim \text{Beta}(s + \alpha, n - s + \beta) = \frac{\mathcal{L}_n(p) \times \text{Beta}(\alpha, \beta)}{c_n}$$

and

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\theta, \sigma_0^2)$$

$$\hat{\theta}_{MLE} = \bar{x}$$

$$\theta \mid x_1 \ldots x_n \sim N(\bar{\mu}, \bar{\sigma}^2) = \frac{\mathcal{L}_n(\theta) \times N(\mu, \tau^2)}{c_n}$$

Bayesian computational challenges

- In the two previous examples, the posterior distribution was available in closed form $\rightarrow$ ☺
- However, often likelihood $\times$ *prior* does not look like any distribution we know (non-conjugacy), and the normalising constant is hard to find
- **Bayesian point estimation** and **prediction** require posterior distribution $\rightarrow$ computing posterior distributions (and hence predictive distributions) is often analytically intractable ☹
- **Model selection** often requires computing very high-dimensional integrals ☹

Bayesian point estimation

Given a loss function $l : \Theta \times \Theta \rightarrow \mathcal{R}$:

$$d^* = \underset{d}{\operatorname{argmin}} \int_\Theta l(d, \theta)\ \pi(\theta|x)\ d\theta$$

If $l(d, \theta) = (d - \theta)^2$:

$$d^* = \int_\Theta \theta\ \pi(\theta|x)\ d\theta = \frac{\int_\Theta \theta\ f(x|\theta)\ \pi(\theta)\ d\theta}{\int_\Theta f(x|\theta)\ \pi(\theta)\ d\theta}$$

Bayesian prediction

The approximation of a distribution related with the parameter of interest, say $g(y|\theta)$, based on the observation $x \sim f(x|\theta)$. The *predictive distribution* is then given by:

$$\pi(y|x) = \int_\Theta g(y|\theta)\, \pi(\theta|x)\, d\theta$$

Bayesian model selection

Compare model classes, e.g. $\mathcal{M}_1$ and $\mathcal{M}_2$. Need to compute posterior probabilities given dataset $\mathcal{D}$:

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}$$

where

$$P(\mathcal{D}|\mathcal{M}) = \int_{\Theta} P(\mathcal{D}|\theta, \mathcal{M})\, P(\theta|\mathcal{M})\, d\theta$$

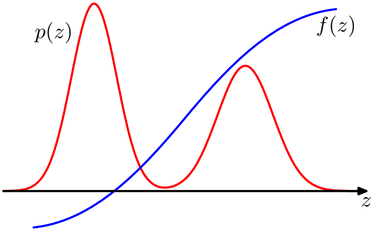
is known as the marginal likelihood.
Computing marginal likelihoods often requires computing very high-dimensional integrals

Bayesian computational challenges

In the different inference problems described above, we often need to compute an expectation:

$$E[f] = \int f(z)\, p(z)\, dz$$

which is to complex to be evaluated exactly using analytical techniques.



Simple Monte Carlo

$$E[f] = \int f(z)\, p(z)\, dz$$

Draw **independent** samples $\{z_1, \ldots, z_n\}$ from distribution $p(z)$ and compute:

$$\hat{f} \approx \frac{1}{N} \sum_{n=1}^{N} f(z_n)$$

Note:

$$E[\hat{f}] = E[f] \textbf{ and } \mathrm{Var}[\hat{f}] = \frac{1}{N} E[(f - E[f])^2]$$

Simple Monte Carlo

$$E[f] = \int f(z)\, p(z)\, dz \approx \frac{1}{N} \sum_{i=1}^{N} f(z_n), \quad z_n \sim p(z)$$

Example (predictive distribution):

$$\pi(y|x) = \int_{\Theta} g(y|\theta)\, \pi(\theta|x)\, d\theta \tag{1}$$

$$\approx \frac{1}{N} \sum_{n=1}^{N} g(y|\theta^n), \quad \theta^n \sim \pi(\theta|x) \tag{2}$$

Problem: It is hard to draw samples from $p(z)$ in general.

Rejection sampling

$$E[f] = \int f(z)\, p(z)\, dz \approx \frac{1}{N} \sum_{i=1}^{N} f(z_n), \quad z_n \sim p(z)$$

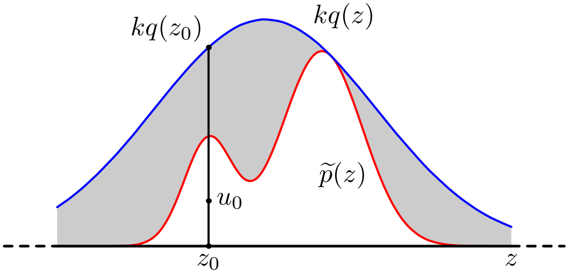
Sampling from **target distribution** $p(z)$ is difficult.

Suppose, as is often the case, that we are easily able to evaluate $p(z)$ for any given value of z, up to some normalising constant $\mathcal{Z}_p$, so that

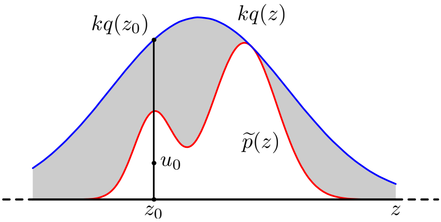
$$p(z) = \tilde{p}(z)/\mathcal{Z}_p$$

Rejection sampling

Suppose we have an easy-to-sample **proposal distribution** $q(z)$, such that $kq(z) \geq \tilde{p}(z), \forall z$.



Rejection sampling



- Sample z_0 from $q(z)$
- Sample u_0 from Uniform$(0, kq(z_0))$
- if $u_0 \leq \tilde{p}(z_0)$, u_0 is retained (white area), otherwise the sample is rejected (grey area).

The pair (z_0, u_0) has uniform distribution under the curve of $kq(z)$.

Rejection sampling

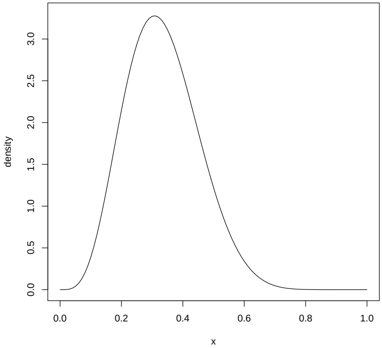
The original values z are **generated** from the distribution q, and these samples are then **accepted** with probability $\tilde{p}(z)/kq(z)$. So, the probability that a sample will be accepted is given by

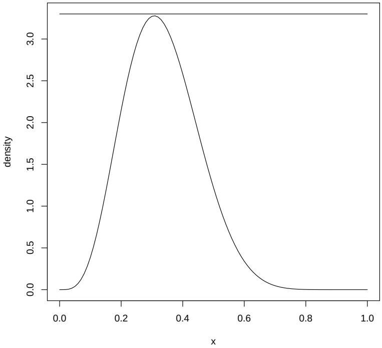
$$P(\mathrm{Accept}) = \int \frac{\tilde{p}(z)}{kq(z)} q(z) dz = \frac{1}{k} \int \tilde{p}(z) dz$$

The fraction of accepted samples depends on the **ratio of the area under** $\tilde{p}(z)$ **and** $kq(z)$. The constant k should be **as small as possible** subject to the limitation that $kq(z)$ **must be nowhere less than** $\tilde{p}(z)$.

Hard to find appropriate $q(z)$ with optimal k. Useful technique in one or two dimensions. Typically applied as a subroutine in more advanced algorithms.

Rejection sampling

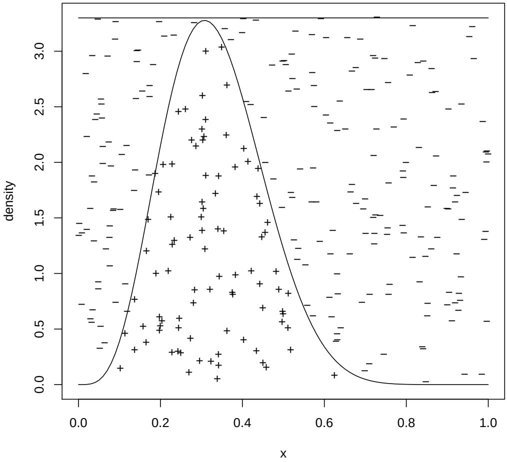




$$f(x;\alpha;\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$X \sim \text{Beta}(5,10)$ and $f(x;5;10) \leq 3.3 \times 1 = 3.3 \times q(x)$ where $q(x)$ is the PDF of a uniform distribution on $[0,1]$.

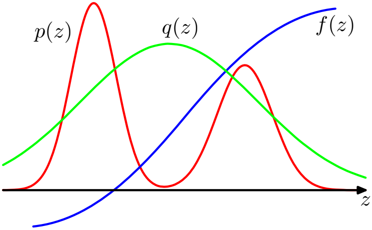
Rejection sampling



Importance sampling

Importance sampling provides a framework for **approximating expectations directly** but does **not** itself provides a mechanism for **drawing samples** from distribution $p(z)$.

Suppose we have an easy-to-sample **proposal distribution** $q(z)$, such that $q(z) > 0$ if $p(z) > 0$



$$
\begin{aligned}
E[f] &= \int f(z)p(z)dz \\
&= \int f(z)\frac{p(z)}{q(z)}q(z)dz \\
&\approx \frac{1}{N}\sum_n \frac{p(z_n)}{q(z_n)}f(z_n), \quad z_n \sim q(z)
\end{aligned}
$$

Importance sampling

- The quantities $w_n = p(z_n)/q(z_n)$ are known as **importance weights**.
- Unlike rejection sampling, all samples are retained.

Suppose $p(z) = \tilde{p}(z)/\mathcal{Z}_p$ and $q(z) = \tilde{q}(z)/\mathcal{Z}_q$:

$$
\begin{aligned}
E[f] &= \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}q(z)dz \\
&= \frac{\mathcal{Z}_q}{\mathcal{Z}_p}\int f(x)\frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)dz \\
&\approx \frac{\mathcal{Z}_q}{\mathcal{Z}_p}\frac{1}{N}\sum_n \frac{\tilde{p}(z_n)}{\tilde{q}(z_n)}f(z_n) = \frac{\mathcal{Z}_q}{\mathcal{Z}_p}\frac{1}{N}\sum_n w_n f(z_n), \quad z_n \sim q(z)
\end{aligned}
$$

Importance sampling

$$\frac{\mathcal{Z}_p}{\mathcal{Z}_q} = \frac{1}{\mathcal{Z}_q}\int \tilde{p}(z)dz = \int \frac{\tilde{p}(z)}{\tilde{q}(z)} q(z)dz$$
$$\approx \frac{1}{N}\sum_n \frac{\tilde{p}(z_n)}{\tilde{q}(z_n)} = \frac{1}{N}\sum_n w_n$$

Hence:

$$E[f] \approx \sum_n \frac{w_n}{\sum_n w_n} f(z_n), \quad z_n \sim q(z)$$

where

$$w_n = p(z_n)/q(z_n)$$

Problems

If our proposal distribution $q(z)$ poorly matches our target distribution $p(z)$ then:

- Rejection Sampling: almost always rejects
- Importance Sampling: has large, possibly infinite, variance (unreliable estimator)

For high-dimensional problems, finding good proposal distributions is very hard. What can we do?

Markov Chains Monte Carlo (MCMC)

Markov Chains Monte Carlo

Recap:

- Analytical calculations on $\pi(z)$ is not possible
- Direct sampling from $\pi(z)$ is not possible

Markov Chains Monte Carlo (MCMC):

1. Construct a Markov chain $\{Z_n\}_0^\infty$ so that

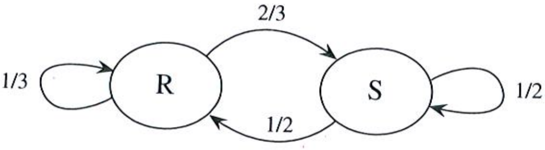
$$lim_{n\to\infty} P(Z_n = z) = \pi(z)$$

2. Simulate the Markov chain for many iterations
3. For m large enough, $z_m, z_{m+1}, z_{m+2}, \dots$ are (essentially) samples from $\pi(z)$

Example: Rainy-sunny Markov chain

If today is rainy, then tomorrow will be rainy with probability $1/3$ and sunny with probability $2/3$. If today is sunny, then tomorrow will be rainy with probability $1/2$ and sunny with probability $1/2$.

If Z_n is the weather on day n, $Z_0, Z_1, Z_2, \ldots$ is a Markov chain on the state space $\{R, S\}$, where R stands for rainy and S for sunny. The transition graph and matrix T are given by



$$\begin{array}{c c} & \begin{array}{cc} S & R \end{array} \\ \begin{array}{c} S \\ R \end{array} & \begin{pmatrix} 1/2 & 1/2 \\ 2/3 & 1/3 \end{pmatrix} \end{array}$$

Markov chains

A **first-order Markov chain**: a series of random variables $\{z_1, \ldots, z_N\}$ such that the following conditional independence property holds for $n \in \{1, \ldots, N-1\}$:

$$p(z_{n+1}|z_1, \ldots, z_n) = p(z_{n+1}|z_n)$$

- Probability distribution for initial state $p(z^1)$
- Conditional probability for subsequent states in the form of transition probabilities $T(z_{n+1} \leftarrow z_n) \equiv p(z_{n+1}|z_n)$

$T(z_{n+1} \leftarrow z_n)$ is also called a **transition kernel**.

Markov chains

The **marginal probability** of a particular state can be computed as:

$$p(z_{n+1}) = \sum_{z_n} T(z_{n+1} \leftarrow z_n) p(z_n)$$

A distribution $\pi(z)$ is said to be **invariant** or **stationary** with respect to a Markov chain if each step in the chain leaves $\pi(z)$ invariant:

$$\pi(z) = \sum_{z'} T(z \leftarrow z') \pi(z')$$

Note: a given Markov chain may have many stationary distributions.

Markov chains

Some Markov chains have a **unique limit distribution**. Our goal is to find conditions under which the Markov chain converges to a unique limit distribution (**independent from its starting state distribution**)

Theorem: If the Markov chain is **irreducible** and **aperiodic**, then the chain will convergence to the unique stationary distribution

- **Irreducibility**: It is possible to get to any state from any state, i.e. $T^K(z^{'} \leftarrow z) > 0, \quad \forall \pi(z^{'}) > 0$
- **Aperiodicity**: The chain cannot get trapped in cycles.

How can we find the limiting distribution of an irreducible and aperiodic Markov chain?

Markov chains

A sufficient (but not necessary) condition for ensuring that $\pi(z)$ is invariant is to choose a transition kernel that satisfies a **detailed balance** property:

$$\pi(z^{'})T(z \leftarrow z^{'}) = \pi(z)T(z^{'} \leftarrow z)$$

A transition kernel that satisfies detailed balance will leave that distribution invariant:

$$\begin{aligned}
\sum_{z^{'}} \pi(z^{'})T(z \leftarrow z^{'}) &= \sum_{z^{'}} \pi(z)T(z^{'} \leftarrow z) \\
&= \pi(z) \sum_{z^{'}} T(z^{'} \leftarrow z) \\
&= \pi(z)
\end{aligned}$$

A Markov chain that satisfied detailed balance is said to be **reversible**.

Recap

We want to sample from target distribution $\pi(z) = \tilde{\pi}(z)/\mathcal{Z}$ (e.g. posterior distribution).

Obtaining independent samples (e.g. using rejection sampling) is difficult.

- Set up a Markov chain with transition kernel $T(z' \leftarrow z)$ that leaves our target distribution $\pi(z)$ invariant.
- If the chain is **irreducible** and **aperiodic**, then the chain will converge to this unique invariant distribution $\pi(z)$.
- We obtain dependent samples drawn approximately from $\pi(z)$ by simulating a Markov chain for some time.

Metropolis-Hasting Algorithm

- A new candidate state z^* is proposed according to some **proposal distribution** $q(z^*|z)$, e.g. $\mathcal{N}(z, \sigma^2)$.
- A candidate state z^* is accepted with probability:

$$min\left(1, \frac{\tilde{\pi}(z^*)q(z|z^*)}{\tilde{\pi}(z)q(z^*|z)}\right)$$

- If accepted, set $z^{'} = z^*$. Otherwise the next state is the copy of the current state ($z^{'} = z$).

Note: no need to know normalising constant $\mathcal{Z}$.

Choice of proposal

Proposal distribution: $q(z^*|z) = \mathcal{N}(z, \sigma^2)$

- σ large: many rejections
- σ small: chain moves too slowly

The specific choice of proposal can greatly affect the performance of the algorithm