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 $l : \Theta \times \Theta \to \mathcal{R}$:

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

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 $\mathcal{D}$:

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

where

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

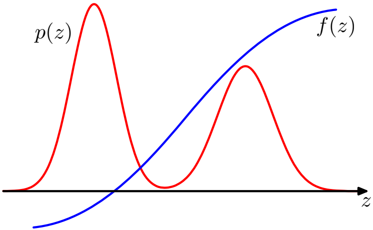
is known as the marginal likelihood.
Computing marginal likelihoods often requires computing very high-dimensional integrals

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

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

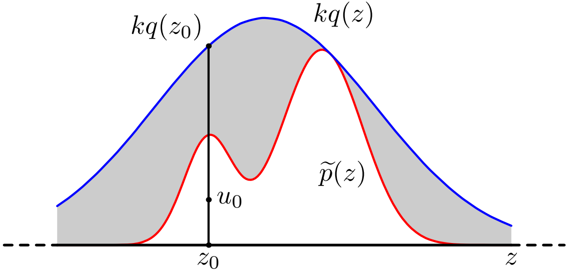
Sampling from **target distribution** $p(z)$ is difficult.

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

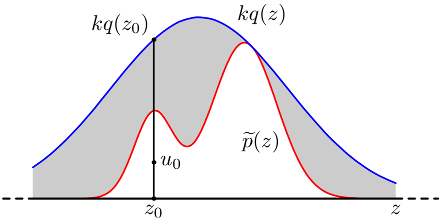
$$p(z) = \tilde{p}(z)/\mathcal{Z}_p$$

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 $kq(z)$.

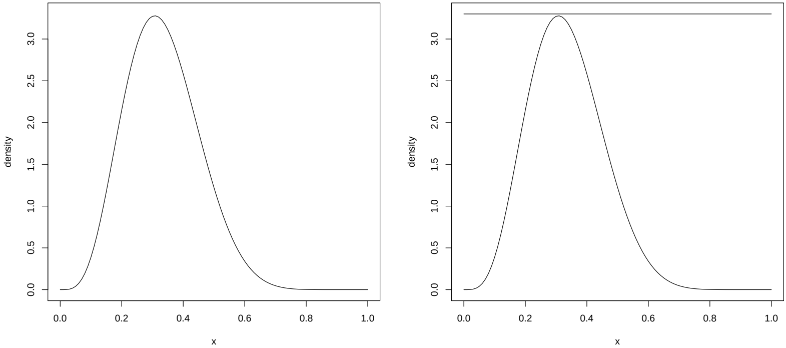
Rejection sampling

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

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

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 $\tilde{p}(z)$**.

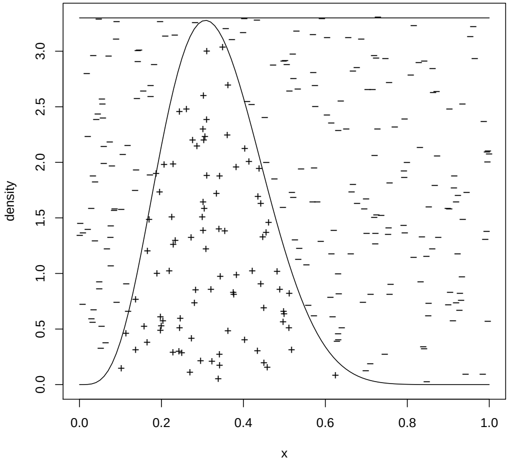
Hard to find appropriate $q(z)$ with optimal k. Useful technique in one or two dimensions. Typically applied as a subroutine in more advanced algorithms.



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

$X \sim \text{Beta}(5, 10)$ 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 $[0, 1]$.

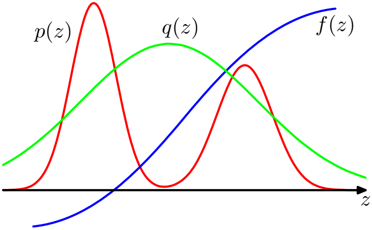
Rejection sampling



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 Chain Monte Carlo (MCMC)