Simulation Based Inference

Chris Men

BI Data Science Capital Markets, Quicken Loans

September 10, 2017

Outline

Motivation

Monte Carlo

Important Sampling

Acceptance Rejection

Conclusion

Why Simulation Based Inference

- In statistics, direct calculation of the moments of a random variable is usually not feasible.

Why Simulation Based Inference

- In statistics, direct calculation of the moments of a random variable is usually not feasible.

- This is even harder in Baysiean statistics because posterior distributions are usually not simple distributions.

Why Simulation Based Inference

- In statistics, direct calculation of the moments of a random variable is usually not feasible.
- This is even harder in Baysiean statistics because posterior distributions are usually not simple distributions.
- Maximum likelihood estimation (MLE) is not always available to fit a model.

Why Simulation Based Inference

- In statistics, direct calculation of the moments of a random variable is usually not feasible.
- This is even harder in Baysiean statistics because posterior distributions are usually not simple distributions.
- Maximum likelihood estimation (MLE) is not always available to fit a model.
- Latent time series models are overly parameterized and simulation based methods are very common.

Why Simulation Based Inference

- In statistics, direct calculation of the moments of a random variable is usually not feasible.
- This is even harder in Baysiean statistics because posterior distributions are usually not simple distributions.
- Maximum likelihood estimation (MLE) is not always available to fit a model.
- Latent time series models are overly parameterized and simulation based methods are very common.
- Markov Chain Monte Carlo (MCMC) methods are more popular but take time to run. C/C++, MATLAB call C, and R call C have to be used.

Why Monte Carlo (MC)?

- Direct calculation is not possible. How to calculate π.

Why Monte Carlo (MC)?

- Direct calculation is not possible. How to calculate π.
- The original MC approach was a method developed by physicists to use random number generation to compute integrals.

Why Monte Carlo (MC)?

- Direct calculation is not possible. How to calculate π.
- The original MC approach was a method developed by physicists to use random number generation to compute integrals.
- Suppose we what to compute an integral

$$\int_a^b f(x)dx = ?$$

Sometimes we can but most of times we can't get this done analytically.

Why Monte Carlo (MC)?

- Direct calculation is not possible. How to calculate π.
- The original MC approach was a method developed by physicists to use random number generation to compute integrals.
- Suppose we what to compute an integral

$$\int_a^b f(x)dx = ?$$

Sometimes we can but most of times we can't get this done analytically.

- If we can decompose $f(x) = h(x)p(x)$, where $p(x)$ is a probability density function (pdf), then we have

$$\int_a^b f(x)dx = \int_a^b h(x)p(x)dx = \mathrm{E}_{p(x)}[h(x)]$$

Law of Large Numbers (LLN)

- Given an independent and identically distributed sequence of randome variables $Y_1, Y_2, ..., Y_n$ with $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$ and $\mathrm{E}(Y_i) = \mu$, then for any $\epsilon > 0$ we have

$$P(|\hat{Y} - \mu| > \epsilon) \to 0$$

as $n \to \infty$.

Law of Large Numbers (LLN)

- Given an independent and identically distributed sequence of randome variables $Y_1, Y_2, ..., Y_n$ with $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$ and $\mathrm{E}(Y_i) = \mu$, then for any $\epsilon > 0$ we have
$$P(|\hat{Y} - \mu| > \epsilon) \to 0$$
as $n \to \infty$.
- According to Law of Large Numbers, as $n \to \infty$, $\bar{Y}_n$ will close to μ, and the approximation will be more accurate.

MC for integral

- If we can generate a random sample $\{x_1, ..., x_n\}$ from $p(x)$, then

$$\int_a^b f(x)dx = \int_a^b h(x)p(x)dx = \mathrm{E}_{p(x)}[h(x)] \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

MC for integral

- If we can generate a random sample $\{x_1, ..., x_n\}$ from $p(x)$, then

$$\int_a^b f(x)dx = \int_a^b h(x)p(x)dx = \mathrm{E}_{p(x)}[h(x)] \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

- A simple case. Integrating a function of $f(x)$ over $[a, b]$ is nothing else than computing the mean of $f(x)$ assuming that $x \sim U[0, 1]$, then

$$\int_a^b f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

MC for integral

- If we can generate a random sample $\{x_1, ..., x_n\}$ from $p(x)$, then

$$\int_a^b f(x)dx = \int_a^b h(x)p(x)dx = \mathrm{E}_{p(x)}[h(x)] \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

- A simple case. Integrating a function of $f(x)$ over $[a,b]$ is nothing else than computing the mean of $f(x)$ assuming that $x \sim U[0,1]$, then

$$\int_a^b f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

- If we can simulate a random variable X over many times we will know almost everything about X.

Monte Carlo Integration in Bayesian Statistics

- In Bayesian inference, MC integration can be used to approximate posterior distributions

$$I(y) = \int_a^b f(y|x)p(x)dx \approx \hat{I}(y) = \frac{1}{n}\sum_{i=1}^{n} f(y|x_i)$$

Monte Carlo Integration in Bayesian Statistics

- In Bayesian inference, MC integration can be used to approximate posterior distributions

$$I(y) = \int_a^b f(y|x)p(x)dx \approx \hat{I}(y) = \frac{1}{n}\sum_{i=1}^{n} f(y|x_i)$$

- The estimated MC standard error is given by

$$\mathrm{SE}^2[\hat{I}(y)] = \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}\Big(f(y|x_i) - \hat{I}(y)\Big)^2\right)$$

Monte Carlo Integration in Bayesian Statistics

- In Bayesian inference, MC integration can be used to approximate posterior distributions

$$I(y) = \int_a^b f(y|x)p(x)dx \approx \hat{I}(y) = \frac{1}{n}\sum_{i=1}^{n} f(y|x_i)$$

- The estimated MC standard error is given by

$$\mathrm{SE}^2[\hat{I}(y)] = \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(f(y|x_i) - \hat{I}(y)\right)^2\right)$$

- This is often used to fit latent time series models where augumented parameters are highly correlated.

Monte Carlo Integration in Bayesian Statistics

- In Bayesian inference, MC integration can be used to approximate posterior distributions

$$I(y) = \int_a^b f(y|x)p(x)dx \approx \hat{I}(y) = \frac{1}{n}\sum_{i=1}^{n} f(y|x_i)$$

- The estimated MC standard error is given by

$$\mathrm{SE}^2[\hat{I}(y)] = \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}\Big(f(y|x_i) - \hat{I}(y)\Big)^2\right)$$

- This is often used to fit latent time series models where augmented parameters are highly correlated.
- This is also widely applied in particle filter methods in terms of one-step and multi-step ahead forecasting.

Monte Carlo Integration
Example 1

- Suppose we want to calculate the expectation of $g(x) = x^d$, where $x \sim N(0, \sigma^2)$ with pdf $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ truncated in $[a, b]$.

$$
\begin{aligned}
\mathrm{E}[g(x)] &= \int_a^b g(x)p(x)dx \\
&= \int_a^b x^d \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx
\end{aligned}
$$

Monte Carlo Integration
Example 1

- Suppose we want to calculate the expectation of $g(x) = x^d$, where $x \sim N(0, \sigma^2)$ with pdf $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ truncated in $[a, b]$.

$$
\begin{aligned}
\mathrm{E}[g(x)] &= \int_a^b g(x)p(x)dx \\
&= \int_a^b x^d \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx
\end{aligned}
$$

- We can simulate either the normal distribution or a uniform distribution on $[a, b]$.
-

$$
\mathrm{E}[g(x)] = \int_a^b \frac{b-a}{\sqrt{2\pi}\sigma} x^d \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \frac{1}{b-a} dx
$$

Monte Carlo Integration

Example 2

- Suppose we want to calculate the expectation of $g(x) = x^d$, where $x \sim Gamma(\alpha, \beta)$ with pdf $p(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right)$ truncated in $[a, b]$.

$$\begin{aligned}
\mathrm{E}[g(x)] &= \int_a^b g(x) f(x) dx \\
&= \int_a^b x^d \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) dx
\end{aligned}$$

Monte Carlo Integration
Example 2

- Suppose we want to calculate the expectation of $g(x) = x^d$, where $x \sim Gamma(\alpha, \beta)$ with pdf $p(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right)$ truncated in $[a, b]$.

$$
\begin{aligned}
\mathrm{E}[g(x)] &= \int_a^b g(x) f(x) dx \\
&= \int_a^b x^d \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) dx
\end{aligned}
$$

- We can simulate either the normal distribution or a uniform distribution on $[a, b]$.

$$
\mathrm{E}[g(x)] = \int_a^b \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{d+\alpha-1} \exp\left(-\frac{x}{\beta}\right) \frac{1}{b-a} dx
$$

Monte Carlo Integration

Example 3

- Suppose we want to estimate $F_Y(y) = P(Y \geq y) = \mathrm{E}[I_{[y,+\infty]}(Y)]$, where $Y \sim N(0,1)$.

$$\begin{aligned}\mathrm{E}[I_{[y,+\infty]}(Y)] &= \int_{y}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) dx \\ &= \int_{-\infty}^{+\infty} h(x) \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) dx\end{aligned}$$

where $h(x) = 0$ if $x < y$ and $h(x) = 1$ if $x \geq y$.

Monte Carlo Integration
Example 3

- Suppose we want to estimate $F_Y(y) = P(Y \geq y) = \mathrm{E}[I_{[y,+\infty]}(Y)]$, where $Y \sim N(0,1)$.

$$\mathrm{E}[I_{[y,+\infty]}(Y)] = \int_y^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) dx$$
$$= \int_{-\infty}^{+\infty} h(x) \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) dx$$

where $h(x) = 0$ if $x < y$ and $h(x) = 1$ if $x \geq y$.

- Draw an iid sample $\{Y_1, ..., Y_n\}$ from $N(0,1)$, then the estimator is

$$\mathrm{E}[I_{[y,+\infty]}(Y)] = \frac{1}{n} \sum_{i=1}^{n} h(Y_i) = \frac{\#\text{ of draws } \geq y}{n}$$

Why Important Sampling (IS)?

- MC sampling is either impossible or not accurate.

Why Important Sampling (IS)?

- MC sampling is either impossible or not accurate.
- Suppose the density $p(x)$ roughly approximates the density (of interest) $q(x)$, then

$$\int_a^b f(x)q(x)dx = \int_a^b f(x)\left(\frac{q(x)}{p(x)}\right)p(x)dx = \mathrm{E}_{p(x)}\left[f(x)\left(\frac{q(x)}{p(x)}\right)\right]$$

Why Important Sampling (IS)?

- MC sampling is either impossible or not accurate.

- Suppose the density $p(x)$ roughly approximates the density (of interest) $q(x)$, then

$$\int_a^b f(x)q(x)dx = \int_a^b f(x)\left(\frac{q(x)}{p(x)}\right)p(x)dx = \mathrm{E}_{p(x)}\left[f(x)\left(\frac{q(x)}{p(x)}\right)\right]$$

- Give a sample $\{x_i, i = 1, ..., n\}$ drawn from $p(x)$, the IS estimator is

$$\int_a^b f(x)q(x)dx = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\left(\frac{q(x_i)}{p(x_i)}\right)$$

Why Important Sampling (IS)?

- In Bayesian statistics.

$$J(y) = \int_a^b f(x)q(x)dx$$

can be approximated by

$$J(y) \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)\left(\frac{q(x_i)}{p(x_i)}\right)$$

An alternative

- An alternative formulation of IS is

$$\int_a^b f(x)q(x)dx \approx \frac{1}{n}\sum_{i=1}^{n}\omega_i f(x_i) \Big/ \sum_{i=1}^{n}\omega_i, \text{ where } \omega_i = \frac{q(x_i)}{p(x_i)}$$

An alternative

- An alternative formulation of IS is

$$\int_a^b f(x)q(x)dx \approx \frac{1}{n}\sum_{i=1}^{n}\omega_i f(x_i) \Big/ \sum_{i=1}^{n}\omega_i, \text{ where } \omega_i = \frac{q(x_i)}{p(x_i)}$$

- The associated MC variance is

$$\mathrm{Var}(\hat{I}) = \frac{1}{n}\sum_{i=1}^{n}\omega_i\big(f(x_i) - \hat{I}\big)^2 \Big/ \sum_{i=1}^{n}\omega_i$$

IS – Example 1

- Suppose we want to estimate $P(Y \geq 3)$, where $Y \sim N(0,1)$.

$$\mathrm{P}(Y > 3) = \int_3^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{x^2}{2}\big) dx = \int_{-\infty}^{+\infty} h(x) \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{x^2}{2}\big) dx,$$

where $h(x) = 0$ if $x < 3$ and $h(x) = 1$ if $x \geq 3$.

IS – Example 1

- Suppose we want to estimate $P(Y \geq 3)$, where $Y \sim N(0,1)$.

$$\mathrm{P}(Y > 3) = \int_3^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) dx = \int_{-\infty}^{+\infty} h(x) \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) dx,$$

where $h(x) = 0$ if $x < 3$ and $h(x) = 1$ if $x \geq 3$.

- Draw an iid sample $\{Y_1, ..., Y_n\}$ from $N(0,1)$, then the estimator is

$$\mathrm{E}[I_{[3,+\infty]}(Y)] = \frac{1}{n} \sum_{i=1}^{n} h(Y_i) = \frac{\#\text{ of draws } \geq 3}{n}$$

IS – Example 1 (Cont'd)

- Draw an iid sample $\{Y_1, ..., Y_{100}\}$ from $N(0, 1)$, then the estimator is

$$\mathrm{E}[I_{[3,+\infty]}(Y)] = \frac{1}{100}\sum_{i=1}^{100} h(Y_i) = \frac{\#\text{ of draws } \geq 3}{100}$$

IS – Example 1 (Cont'd)

- Draw an iid sample $\{Y_1, ..., Y_{100}\}$ from $N(0,1)$, then the estimator is

$$\mathrm{E}[I_{[3,+\infty]}(Y)] = \frac{1}{100}\sum_{i=1}^{100} h(Y_i) = \frac{\#\text{ of draws } \geq 3}{100}$$

- Draw an iid sample $\{Y_1, ..., Y_{100}\}$ from $N(3,1)$, then the estimator is

$$\mathrm{E}[I_{[3,+\infty]}(Y)] = \frac{1}{100}\sum_{i=1}^{100} \frac{h(Y_i)f(y_i)}{g(Y_i)},$$

where $f(y)$ is the density of a $N(0,1)$ and $g(x)$ is the density of $N(4,1)$.

Accept Rejection

- Suppose that we wish to generate samples from a distribution $\pi(x) \propto f(x)/k$, where k is the unknown normalizing constant.

Accept Rejection

- Suppose that we wish to generate samples from a distribution $\pi(x) \propto f(x)/k$, where k is the unknown normalizing constant.
- Let $g(x)$ be the density function that can be sampled by a known method.

Accept Rejection

- Suppose that we wish to generate samples from a distribution $\pi(x) \propto f(x)/k$, where k is the unknown normalizing constant.
- Let $g(x)$ be the density function that can be sampled by a known method.
- Suppose that their is a known constant c satisfying $f(x) \leq cg(x)$ for any x.

The AR Procedure

- The AR procedure
  1. Generate a candidate y from $g(.)$ and a value u from a uniform distribution $\mathcal{U}(0,1)$.
  2. If $u \leq f(y)/(cg(y))$, then return $x = y$; else go to step 1.

The AR Procedure

- The AR procedure
    1. Generate a candidate y from $g(.)$ and a value u from a uniform distribution $\mathcal{U}(0,1)$.
    2. If $u \leq f(y)/(cg(y))$, then return $x = y$; else go to step 1.
- The expected number of iterations of the AR algorithm to generate a sample point from $f(x)$ is $1/c$.

The AR Procedure

- The AR procedure
  1. Generate a candidate y from $g(.)$ and a value u from a uniform distribution $\mathcal{U}(0,1)$.
  2. If $u \leq f(y)/(cg(y))$, then return $x = y$; else go to step 1.

- The expected number of iterations of the AR algorithm to generate a sample point from $f(x)$ is $1/c$.

- The AR sampling method can be optimized by setting

$$c = \sup_x \left\{ \frac{f(x)}{g(x)} \right\}.$$

The AR Procedure

- The AR procedure
  1. Generate a candidate y from $g(.)$ and a value u from a uniform distribution $\mathcal{U}(0,1)$.
  2. If $u \leq f(y)/(cg(y))$, then return $x = y$; else go to step 1.

- The expected number of iterations of the AR algorithm to generate a sample point from $f(x)$ is $1/c$.

- The AR sampling method can be optimized by setting

$$c = \sup_x \left\{ \frac{f(x)}{g(x)} \right\}.$$

- Note: Usually the value of c is small and then the AR method may not be efficient.

AR – Example 1

- Want to simulate a Beta distribution, $f(x) = \gamma x^a (1-x)^b$, where γ is the normalizing constant.

AR – Example 1

- Want to simulate a Beta distribution, $f(x) = \gamma x^a(1-x)^b$, where γ is the normalizing constant.
- We choose a uniform distribution $g(y) = 1$ on the $[0, 1]$ as the proposal distribution.

AR – Example 1

- Want to simulate a Beta distribution, $f(x) = \gamma x^a(1-x)^b$, where γ is the normalizing constant.
- We choose a uniform distribution $g(y) = 1$ on the $[0, 1]$ as the proposal distribution.
- The AR method to simulation $f(x) = \gamma x^a(1-x)^b$ as follows:
  1. Generate a candidate y from $g(y)$ and a value u from a uniform distribution $\mathcal{U}(0, 1)$.
  2. If $u \leq f(y)$, then return $x = y$; else go to step 1.

AR – Example 1

- Want to simulate a Beta distribution, $f(x) = \gamma x^a(1-x)^b$, where γ is the normalizing constant.
- We choose a uniform distribution $g(y) = 1$ on the $[0, 1]$ as the proposal distribution.
- The AR method to simulation $f(x) = \gamma x^a(1-x)^b$ as follows:
    1. Generate a candidate y from $g(y)$ and a value u from a uniform distribution $\mathcal{U}(0, 1)$.
    2. If $u \leq f(y)$, then return $x = y$; else go to step 1.
- Note: For large values of a, this AR method may not be efficient.

AR – Example 2

- Want to simulate a Gamma distribution, $f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp(-x/\beta)$.

AR – Example 2

- Want to simulate a Gamma distribution, $f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta)$.
- Consider a special case: $\alpha = 3/2, \beta = 1$, then

$$f(x) = 2\sqrt{\frac{x}{\pi}} e^{-x}$$

AR – Example 2

- Want to simulate a Gamma distribution, $f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta)$.
- Consider a special case: $\alpha = 3/2, \beta = 1$, then

$$f(x) = 2\sqrt{\frac{x}{\pi}} e^{-x}$$

- Let $g(x) = \lambda e^{-\lambda x}$, then

$$c = \sup_{x>0} \left\{ \frac{f(x)}{g(x)} \right\} = \sup_{x>0} \left\{ \frac{2\sqrt{x} e^{-x}}{\sqrt{\pi}\lambda e^{-\lambda x}} \right\} = \frac{1}{\sqrt{2e\pi\lambda^2(1-\lambda)}}$$

AR – Example 2

- Want to simulate a Gamma distribution, $f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta)$.
- Consider a special case: $\alpha = 3/2, \beta = 1$, then

$$f(x) = 2\sqrt{\frac{x}{\pi}} e^{-x}$$

- Let $g(x) = \lambda e^{-\lambda x}$, then

$$c = \sup_{x>0}\left\{\frac{f(x)}{g(x)}\right\} = \sup_{x>0}\left\{\frac{2\sqrt{x}e^{-x}}{\sqrt{\pi}\lambda e^{-\lambda x}}\right\} = \frac{1}{\sqrt{2e\pi\lambda^2(1-\lambda)}}$$

- Note: The smallest c is most efficient, which is $c^* = 1.257$ with $\lambda = 2/3$.
- That is, $g(x) = \frac{2}{3} e^{-2x/3}$ and $c^* = 1.257$.

Summary

- Monte Carlo methods are useful for obtaining integrals, especially for moment calculation.

Summary

- Monte Carlo methods are useful for obtaining integrals, especially for moment calculation.
- Important sampling is usful for integrals with complicated functions.

Summary

- Monte Carlo methods are useful for obtaining integrals, especially for moment calculation.
- Important sampling is usful for integrals with complicated functions.
- Acceptance-rejection methods are more popular in generating random numbers from partially known density functions.

Summary

- Monte Carlo methods are useful for obtaining integrals, especially for moment calculation.
- Important sampling is usful for integrals with complicated functions.
- Acceptance-rejection methods are more popular in generating random numbers from partially known density functions.
- These three simulation methods are basic tools for Markov chain Monte Carlo (MCMC) methods.

Thank you!

References

[1] Chib, S. and Greenberg E. 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician* Vol. 49, No. 4: 327-335.(full text).

[2] McLeish, D. L. and Z. Men. 2015. Extreme Value Importance Sampling for Rare Event Risk Measurement. K. Glau et al. (eds.), Innovations in Quantitative Risk Management. Springer Proceedings in Mathematics & Statistics. Volume 99, 2015, pp 317-335. (full text).