Quantitative Applications for Data Analysis: Monte Carlo methods

Erik Spence

SciNet HPC Consortium

21 March 2023

SciNet compute • calcul CANADA

Today's slides

Today's slides can be found here. Go to the "Quantitative Applications for Data Analysis" page, under Lectures, "Monte Carlo methods".

`https://scinet.courses/1276`

SciNet compute • calcul CANADA

Today's class

Today we will visit the following topics:

- Monte Carlo methods, in general,
- Monte Carlo sampling,
- Monte Carlo integration,
- Markov Chain Monte Carlo.

SciNet compute • calcul CANADA

Monte Carlo analysis

Monte Carlo analyses are a collection of techniques whose unifying feature is the use of random sampling to generate results. These analyses generally fall into one of three categories:

- Adding randomness to otherwise-deterministic dynamics, and studying how the dynamics are changed, or the resulting data distributions.
- Generating samples from a given probability distribution, $\boldsymbol{P}(\boldsymbol{x})$, usually a distribution that is complicated and can't be dealt with nicely in closed form.
- Estimating expectation values under this distribution, e.g.

$$\langle \boldsymbol{A}(\mathbf{x}) \rangle = \int \boldsymbol{P}(\mathbf{x}) \boldsymbol{A}(\mathbf{x}) d\mathbf{x}$$

  where $\mathbf{x}$ is typically high dimensional.

These depend on having a good random number generator!

Random sampling, example

We can use a Monte Carlo analysis to help with decision making.

- Determine all the variables that go into making a particular decision.
- Assign a weight to each variable, indicating each variable's relative importance.
- Determine the probability distribution for each variable. This may be uniform, Gaussian, whatever seems appropriate. The wider the range, or the larger the standard deviation, the less certain you are about that particular variable.
- Sample each variable randomly from its distribution, many times.
- Multiply each sample by its weight, and sum up the values.
- Repeat many times to get distributions for each option.

Obviously the distributions and weights are subjective.

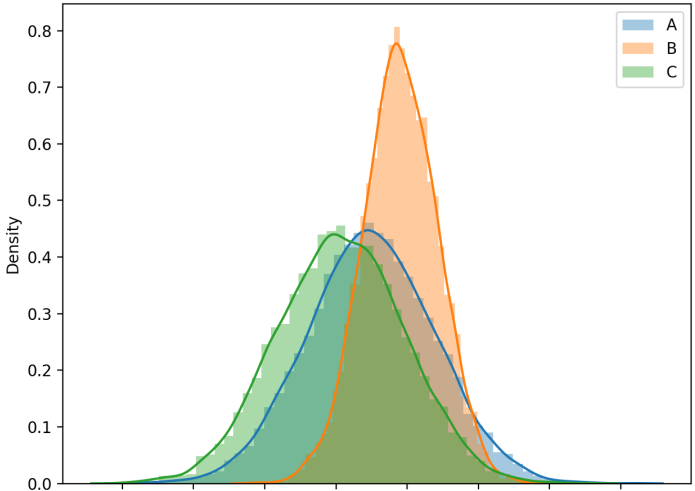
Let's do an example. Suppose we want to choose a university to attend. We have three universities which we're considering. Let us run a Monte Carlo analysis for this decision.

Random sampling, example, continued

```
import numpy.random as npr

vars = ['rep', 'prog', 'cost',
        'loc', 'atm']

def uni_MC(grade, num):
  final_results = []
  weights = [0.15, 0.2, 0.35, 0.1, 0.1]
  for n in range(num):
    results = 0
    for i in range(len(weights)):
      results += weights[i] * \
              npr.normal(grade[i][0],
                         grade[i][1])

    final_results.append(results)
  return final_results
```

```
In [1]:
In [1]: # The sub-lists represent the mean and sd.
In [1]: a = uni_MC([[8,2],[8,2],[6,2],[9,1],[7,2]], 10000)
In [2]: b = uni_MC([[8,2],[7,1],[8,1],[7,1],[8,1]], 10000)
In [3]: c = uni_MC([[6,1],[7,2],[7,2],[7,2],[6,3]], 10000)
In [4]:
In [4]: import seaborn as sns
In [5]:
In [5]: sns.distplot(a, label = 'A')
In [6]: sns.distplot(b, label = 'B')
In [7]: sns.distplot(c, label = 'C')
In [8]: legend()
In [9]:
```

Random sampling, example, continued more



Random sampling example, notes

Though this example is a bit contrived, it demonstrates some useful properties:

- Random sampling from distributions can be used to combine together different distributions, especially if the relationships between the distributions are complicated.
- The final distribution which results can be used to draw conclusions about the final possible outcomes.

This is a common use-case in many branches of science.

SciNet compute • calcul CANADA

Monte Carlo example: integration

The basic idea of Monte Carlo integration is very simple and only requires elementary statistics. Suppose we want to find the value of

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

.

The quantity S is usually estimated using the expression

$$S \simeq \frac{(b-a)}{n} \sum_{i=1}^{n} f(a + U_i(b-a))$$

where U is the uniform distribution sampled n times between b and a. As you can see, this estimation is simply calculating the average value of f in the interval and then multiplying by $(b-a)$ to get the value of the area.

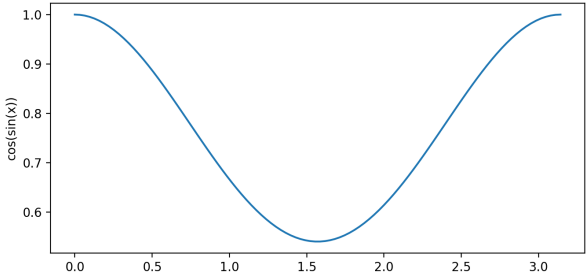
Integration, example

Let's integrate the function

$$f(x) = \cos(\sin(x))$$

from 0 to π.

```
In [9]:
In [9]: import numpy as np
In [10]: import matplotlib.pyplot as plt
In [11]:
In [11]: x = np.linspace(0, pi, 100)
In [12]: plt.plot(x, np.cos(np.sin(x)))
In [13]:
```



Integration, example, continued

Let's integrate the function
$f(x) = \cos(\sin(x))$ from 0 to π.

```
# int_MC.py
import numpy.random as npr
import numpy as np

def f(x):
  return np.cos(np.sin(x))

def int_MC(num, a, b):
  results = 0
  for n in range(num):
    results += f(a + npr.uniform(0, b - a))

  return results * (b - a) / num
```

```
In [13]:
In [13]: import int_MC
In [14]:
In [14]: int_MC.int_MC(10000, 0, pi)
Out[14]: 2.4101580829908666
In [15]:
In [15]: import scipy.integrate as si
In [16]:
In [16]: si.simps(np.cos(np.sin(x)), x)
Out[16]: 2.403936768802837
In [17]:
```

Multidimensional integration

Using a Monte Carlo integration technique on a 1D problem is inefficient. There are much more efficient techniques out there, such as Simpson's rule.

But suppose our integral is of a higher dimension, say, 4D. This is where Monte Carlo integration techniques start to become more useful. They can efficiently reach into as many dimensions as necessary.

$$S = \int_V f(\mathbf{x}) d\mathbf{x}$$

Where now we are integrating over the 3D domain V, and $\mathbf{x}$ is a 3D vector.

Multidimensional integration, continued

Let us integrate over 3 dimensions, rather than 1.

$$S = \int_V f(\mathbf{x}) d\mathbf{x}$$

The quantity S is estimated using a similar expression to the 1D example.

$$S \simeq \frac{V}{n} \sum_{i=1}^{n} f(a_x + U_i(b_x - a_x), a_y + U_i(b_y - a_y), a_z + U_i(b_z - a_z))$$

where a_x, b_x are the limits of integration for x, and the samples from the uniform distribution must be evenly sampled throughout V. Obviously, if a data point is randomly sampled outside of V it cannot be used.

SciNet compute • calcul CANADA

Multidimensional integration, example

Let's integrate over the 4-sphere, to calculate its volume.

Rather than limit the range of npr.uniform to V, we keep sampling points until all points are within V. This makes sure the points are spaced evenly.

Note that the volume of a 4-sphere is $\pi^2 r^4/2$. We get half this value, since we're only integrating half of the 4-sphere.

```
In [17]: import MultiD_int_MC as MD
In [18]:
In [18]: MD.multiD_int_MC(1, 10000)
Out[18]: 2.4791925774411387
```

```
# MultiD_int_MC.py
import numpy.random as npr, numpy as np

def f(r, x, y, z):
  return np.sqrt(r**2 - x**2 - y**2 - z**2)

def my_samp(r):
  x = npr.uniform(-r, r); y = npr.uniform(-r, r)
  z = npr.uniform(-r, r); return x, y, z

def multiD_int_MC(r, num):
  results = 0
  for n in range(num):
    x, y, z = my_samp(r)
    while (x**2 + y**2 + z**2 > r**2):
      x, y, z = my_samp(r)
    results += f(x, y, z)
  return results / num * (4 / 3 * np.pi * r**3)
```

Monte Carlo integration, summary

A few notes about Monte Carlo integration.

- MC integration works under minimal assumptions (the desired mean must exist, then (law of large numbers) $\mathcal{P}(\lim_{n\to\infty} \hat{\mu} = \mu) = 1$.
- MC integration does not deliver extreme accuracy

$$\mathbf{RMSE} = E((\hat{\mu} - \mu)^2) = \sigma/\sqrt{n}$$

- MC integration is very competitive in high dimensional or non-smooth problems.
- MC integration has good error estimation.
- There are ways to improve the approach we've used, such as using using non-uniform sampling (Importance sampling).

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is considered one of the most important algorithms of the 20th century.

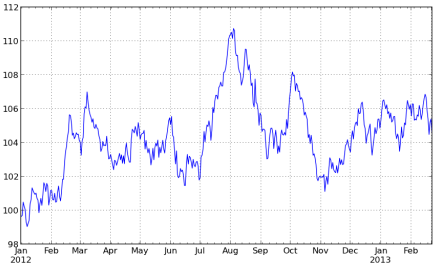
- It combines two techniques:
  - Monte Carlo: estimating a distribution's properties by randomly sampling from it, and
  - Markov Chains: the random samples are generated by a restricted sequential process.
- The resulting chain of samples is essentially a 'random walk' through a high-dimensional space.
- This walk is used for optimization problems, in particular Bayesian inference.
- Once you have your chain, you can use this data to determine the distributions of whatever parameters you're after.

SciNet compute • calcul CANADA

Markov chain

A Markov chain is a chain of 'steps' through parameter space, with particular characteristics.

- Usually when we deal with random samples they are independent and identically distributed. New samples don't depend on previous samples.
- A *Markov chain* is a sequence of random numbers $\boldsymbol{X_0, X_1, ..., X_n}$ where the probability of $\boldsymbol{X_{i+1}}$ depends on $\boldsymbol{X_i}$, but does NOT depend on $\boldsymbol{X_{i-1}}$.
- Distribution is $\boldsymbol{P(X_{i+1}|X_i)}$ instead of $\boldsymbol{P(X_i)}$.
- A classic example of a Markov chain is a random walk: $\boldsymbol{X_{i+1} = X_i + \epsilon}$



SciNet compute • calcul CANADA

Bayesian Inference

Bayesian inference is a process for updating our beliefs when we acquire new information, using Bayes theorem:

$$\boldsymbol{P(X|d) = \frac{P(d|X)P(X)}{P(d)}}$$

Terminology:

- $\boldsymbol{d}$ is the data, $\boldsymbol{X}$ are the model parameters.
- $\boldsymbol{P(X|d)}$ is called the *posterior*, the probability distribution of $\boldsymbol{X}$ after knowing $\boldsymbol{d}$.
- $\boldsymbol{P(X)}$ is called the *prior*, $\boldsymbol{\pi(X)}$, the *a priori* probability distribution of $\boldsymbol{X}$ (our beliefs about $\boldsymbol{X}$ prior to knowing $\boldsymbol{d}$).
- $\boldsymbol{P(d|X)}$ is called the *likelihood*, $\boldsymbol{\mathcal{L}(X, d)}$, the probability distribution of $\boldsymbol{d}$ given $\boldsymbol{X}$.
- $\boldsymbol{P(d)}$ is called the *model evidence*: $\boldsymbol{P(d) = \int P(d|X)P(X)dX}$.

SciNet compute • calcul CANADA

MCMC

MCMC is most often used to determine the posterior distribution, $\boldsymbol{P(X|d)}$, for some problem. So what do we need to do this?

- Data, $\boldsymbol{d}$. Presumably you already have this.
- A model with which you can represent your data. The model parameters, $\boldsymbol{X}$, are contained herein.
- A sampling algorithm, with which you generate new values for $\boldsymbol{X}$, your model parameters.
- An equation for your likelihood, $\boldsymbol{\mathcal{L}(X, d)}$,.
- An equation for the prior, $\boldsymbol{\pi(X)}$. This is needed, as you need to describe what you think the parameters look like before you have any data.

Once you have these pieces you can begin to perform MCMC.

SciNet compute • calcul CANADA

Bayesian Inference, likelihood

So how do we calculate the likelihood, $P(d|X)$ ($\mathcal{L}(X, d)$),? We make the usual assumption, that our data contains noise, $y_{\mathbf{obs}} = y_{\mathbf{true}} + \epsilon$, and then assume that the noise is Gaussian, and the data are uncorrelated. As a result, we have

$$P(d|X) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2}$$

Recall that $y_i - f(x_i)$ are just the residuals of the model, f, and σ_i is the uncertainty for data point i. This is sometimes modelled as the log of the likelihood.

$$\log P(d|X) = \sum_i -\log(2\pi) - \log(\sigma_i^2) - \left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2$$

where we've multiplied the right side by 2. Logs are easier to deal with and are more numerically stable.

Bayesian Inference, prior

So how do we calculate the prior, $\pi(\boldsymbol{X})$? This is subjective, based on the prior knowledge of the researcher. There are several commonly used options:

- uniform prior: prior is a constant over some range, no particular value of $\boldsymbol{X}$ is any better than any other.
- log-uniform prior: useful if your parameter many orders of magnitude (doesn't work near zero).
- posterior prior: use a previously-calculated posterior as your prior, if the posterior is related to the problem you're working on.
- observation-based prior: using observations to create a distribution to use as your prior.

It's not uncommon to use the log of the prior, since it's more numerically stable.

SciNet compute • calcul CANADA

MCMC, sampler

So what does the sampling algorithm do?

- It generates a new value for $\boldsymbol{X_{i+1}}$, given $\boldsymbol{X_i}$.
- This is done by sampling from the "proposal distribution", $\boldsymbol{q(X_{i+1}|X_i)}$, (typically Gaussian, for continuous model parameters). This value is added to $\boldsymbol{X_i}$ to get $\boldsymbol{X_{i+1}}$. The proposal distribution should be symmetric and centred at zero.
- This value of $\boldsymbol{X_{i+1}}$ is then passed to the model, which, using the data, calculates the likelihood, $\boldsymbol{\mathcal{L}}$.
- The value is also passed to the prior, $\boldsymbol{\pi(X)}$, to calculate its value.
- The likelihood and prior are then combined to calculate the posterior, $\boldsymbol{P(X|d)}$.

$$\boldsymbol{P(X|d) \propto \mathcal{L}(X,d)\pi(X)}$$

We ignore the denominator $\boldsymbol{P(d)}$, as this is just a normalization constant.

SciNet compute • calcul CANADA

MCMC, continued

We will use a particular type of MCMC called the Metropolis algorithm.

1. Choose a starting position, $\boldsymbol{X_0}$.
2. Propose the next data point using the sampler.
3. Compare the posterior's new value, $\boldsymbol{P(X_{i+1}|d)}$, to the previous value, $\boldsymbol{P(X_i|d)}$. If $\boldsymbol{P(X_{i+1}|d) > P(X_i|d)}$ then take $\boldsymbol{X_{i+1}}$ as our current point.
4. If $\boldsymbol{P(X_{i+1}|d) < P(X_i|d)}$ then randomly take $\boldsymbol{X_{i+1}}$ to be the current value with probability $\boldsymbol{P(X_{i+1}|d)/P(X_i|d)}$.
5. Repeat, starting at step 2.

The resulting chain, $\boldsymbol{X_0, X_1, ..., X_n}$, is the data used to calculate $\boldsymbol{P(X|d)}$.

Bayesian inference, example

Suppose we've got some data and, having plotted it, have decided that the data follows a linear relationship. Let's use MCMC to generate the distributions of our model parameters.

What do we need to do MCMC?

- A model: $\boldsymbol{y = mx + b + \epsilon}$. Let us assume that the noise is Gaussian, with a mean of zero and a standard deviation of $\boldsymbol{\sigma}$.
- This means there are 3 model parameters: $\boldsymbol{X = (m, b, \sigma)}$.
- A likelihood, $\boldsymbol{\mathcal{L}(X, d)}$: we will use a Gaussian of the residuals.
- A prior, $\boldsymbol{\pi(X)}$: we will use uniform priors for this calculation.

We will code this ourselves, as it's not too difficult.

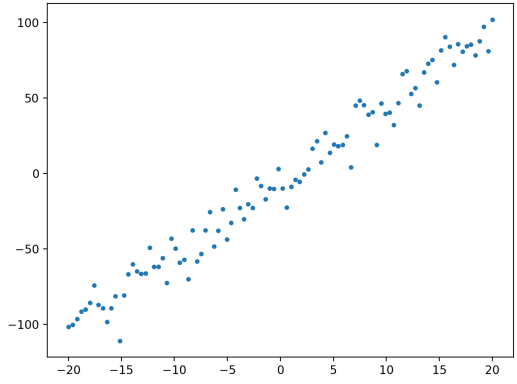
Bayesian inference, example, continued

```
# my_mcmc.py
import scipy.stats as ss, numpy as np

def likelihood(params, x, y):
  m = params[0];   b = params[1]
  sd = params[2]
  pred = m * x + b

  # We want the log of the Gaussian.  We take
  # the sum because we are taking the log.
  return np.sum(ss.norm.logpdf(y,
    loc = pred, scale = sd))

def prior(params):
  # We will take all priors to be 1, meaning
  # all equally probable.  But we return the
  # log of the result, which is 0.
  return 0
```

```
# my_mcmc.py, continued
def posterior(params, x, y):
  return likelihood(params, x, y) +
    prior(params)

def proposal_func(params):
  # We propose 3 new values based on the
  # existing values, using hard-coded
  # standard deviations.
  return ss.norm.rvs(size = 3, loc = params,
    scale = np.array([0.1, 0.1, 0.1])

def calc_chisq(params, x, y):
  m = params[0];   b = params[1]
  sd = params[2]
  return np.sum(((m * x + b - y) / sd)**2)
```

Bayesian inference, example, continued more

```
In [19]:
In [19]: import matplotlib.pyplot as plt
In [20]:
In [20]: trueM, trueB, trueSd = 5, -5, 10
In [21]:
In [21]: n = 100
In [22]: x = np.linspace(-20, 20, n)
In [23]:
In [23]: y = trueM * x + trueB
In [24]: y += ss.norm.rvs(size = n, loc = 0,
                  scale = trueSd)
In [25]:
In [25]: plt.plot(x, y, '.')
In [26]:
```



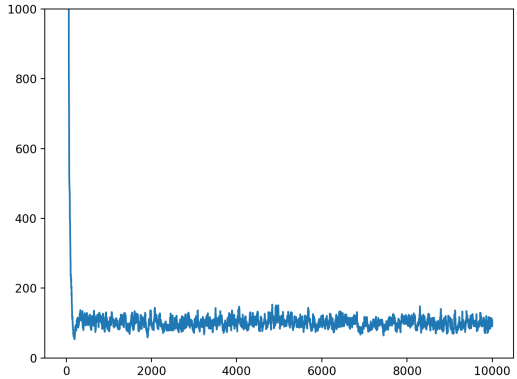
Bayesian inference, example, continued some more

```
In [26]:
In [26]: import my_mcmc
In [27]:
In [27]: startvalue = np.array([1, 0, 1])
In [28]:
In [28]: num = 10000
In [29]:
In [29]: chain = my_mcmc.run_mcmc(startvalue,
                 x, y, num)
In [30]:
In [30]: chain = np.array(chain)
In [31]:
```

```
# my_mcmc.py, continued
def run_mcmc(startvalue, x, y, n):
  chain = [startvalue]

  for i in range(n):
    proposal = proposal_func(chain[i])
    old_p = posterior(chain[i], x, y)
    new_p = posterior(proposal, x, y)

    if new_p > old_p:
      chain.append(proposal)
    else:
      p = np.exp(new_p - old_p)
      if ss.uniform.rvs() < p:
        chain.append(proposal)
      else: chain.append(chain[i])
  return chain
```

Bayesian inference, example, burn in

```
In [31]: chisq = np.zeros(num + 1)
In [32]:
In [32]: for i in range(num + 1):
...:     chisq[i] = my_mcmc.calc_chisq(chain[i,],
...:                         x, y)
In [33]:
In [33]: plt.plot(chisq)
In [34]: plt.ylim(0,1000)
In [35]:
In [35]: trueM, np.mean(chain[2000:,0])
Out[35]: (5, 4.973537491078847)
In [36]: trueB, np.mean(chain[2000:,1])
Out[36]: (-5, -4.719488741813247)
In [37]: trueSd, np.mean(chain[2000:,2])
Out[37]: (10, 10.325602452263714)
```



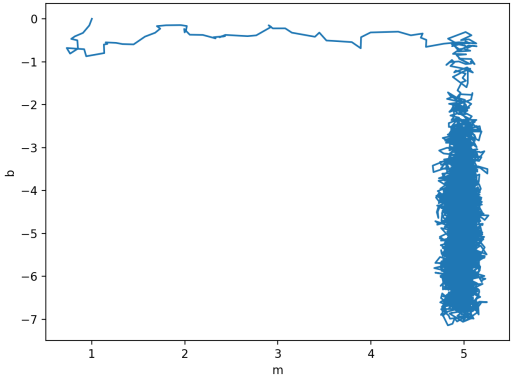
"Burn in" is the period before the MCMC lands near the correct values, as can be seen in the chi-squared values for the chain.

Bayesian inference, example, burn in, continued

We can plot the walk as it tries to find the correct values of the parameters.

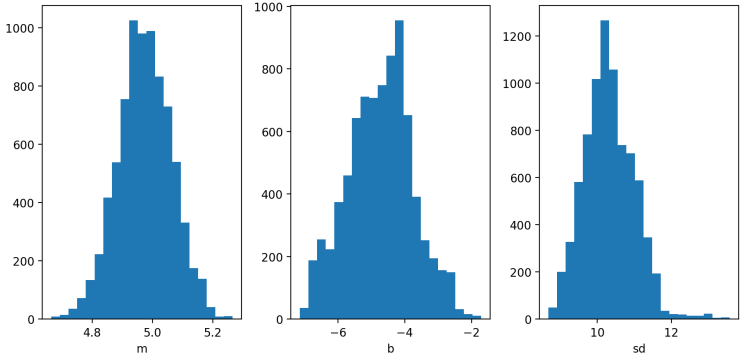
We can see that the chain has a much better sense of the value of m than b.

```
In [38]:
In [38]: plt.plot(chain[:,0], chain[:,1])
In [39]:
```



Bayesian inference, example, burn in, continued

```
In [39]: ans = chain[2000:,]
In [40]:
In [40]: plt.subplot(131)
In [41]: h = plt.hist(ans[:,0])
In [42]: plt.xlabel('m')
In [42]:
In [42]: plt.subplot(131)
In [43]: h = plt.hist(ans[:,1])
In [44]: plt.xlabel('b')
In [44]:
In [44]: plt.subplot(131)
In [45]: h = plt.hist(ans[:,2])
In [46]: plt.xlabel('sd')
In [46]:
```



Summary

Some notes from today's class.

- MCMC is a powerful technique, but it's not foolproof.
- How to know if the chain has adequately sampled the distribution (aka **converged**)?
  - Run multiple chains with different starting points, and compare the inter-chain and intra-chain variances (*Gelman-Rubin test*).
- an MCMC used for approx. a multi-dim integral $\rightsquigarrow$ an ensemble of "walkers" moving around randomly. At each point where a walker steps, the integrand value at that point is counted towards the integral.
- the random samples of the integrand used in a conventional MC integration are statistically independent, those used in MCMC methods are correlated.
- A Markov chain is constructed in such a way as to have the integrand as its equilibrium distribution.