How to assess the output of an MCMC sampler

Kubo formula:

Let $x_i, i = 1, 2, \ldots, Ne$ be a random sequence.

$$\underbrace{Var\left(\sum_{i=1}^{Ne} x_i\right)}_{\text{variance of n steps of the "random walk"}} \approx Ne \cdot \hat{D}$$

"effective diffusion time"

(see books for specifics of D)

E.g. if $x_i \sim N(0,1)$ iid, then

$$Var\left(\sum_{i=1}^{Ne} x_i\right) = \sum_{i=1}^{Ne} Var(x_i) = Ne \quad \Rightarrow D = 1 \text{ for iid Gaussians.}$$

$\Rightarrow$ D describes how quick the random sequence/random walk diffuses.

We can use this as follows:

Suppose $x_1, \ldots x_{Ne}$ is ~~gets then~~ the output of your MCMC sampler.

You estimate $E(x) \approx \frac{1}{Ne} \sum_{i=1}^{Ne} x_i$

The variance of this estimator is

$$Var\left(\frac{1}{Ne}\sum_{i=1}^{Ne} x_i\right) = \frac{1}{Ne^2} var\left(\sum_{i=1}^{Ne} x_i\right) = \frac{1}{Ne^2} Ne \cdot D = \frac{D}{Ne}$$

D in Kubo's formula is defined as follows:

Define: $C(t,s) = \mathrm{cov}(X_t, X_s) = E[(X_t - \mu_t)(X_s - \mu_s)]$

integer, discrete time!

If we assume that the process is stationary, then

$$C(t,s) = C(t-s) = \mathrm{Cov}(X_t, X_{t+s})$$ ~~vector correlation~~ "auto covariance".

Renaming variables:

$$C(t) = \mathrm{Cov}(X_s, X_{s+t})$$

Define: $\rho(t) = C(t)/C(0)$ "auto-correlation".

With these definitions:

$$D = C(0) + 2\sum_{t=1}^{\infty} C(t)$$

$$= C(0)\underbrace{\left(1 + 2\sum_{t=1}^{\infty}\rho(t)\right)}_{:=\,\tau}$$

$:= \tau$, the integrated auto correlation time

$$\boxed{D = C(0)\cdot\tau}$$

Then $\mathrm{var}\left(\frac{1}{N_e}\sum_{i=1}^{N_e} x_i\right) \overset{\text{Kubo}}{\approx} \frac{D}{N_e} \overset{\text{see Defs.}}{=} \frac{C(0)\cdot\tau}{N_e} = \frac{C(0)}{(N_e/\tau)}$

a variance

effective sample size

$$\boxed{N_{eff} = N_e/\tau}$$

! When you use MCMC, you must report N_e and τ !

(E.g. if $N_e = 1000$, but $\tau = 100$, then accuracy of your estimates is ~~the same~~ comparable to an accuracy of 10 samples, i.e., not much!)

Results/lessons

- Always compute τ
- Always report τ
- An effective MCMC sampler requires that τ is "small".
- For MH, many samplers require that proposal variance decreases with dimension (to keep acceptance ratio reasonable). This increases τ. MCMC may be problematic in high dimensions.
- As in importance sampling: N_{eff} (or τ) is computed from the samples of the ~~the~~ sampler. This makes it difficult to compute.

Suggested procedure:

Run MCMC sampler

compute τ

disregard first 5-10τ as "burn-in" (transition to steady state)

Compute estimates.

Then keep the chain running and repeat above steps for much larger N_e.

If the results are consistent, then you *might* get the right answer.

How large should N_e be?

$10 \cdot \tau$ samples for burn in

τ samples per effective sample

$\Rightarrow N_e \approx 1000 \cdot \tau$ (to be safe)

How to Compute τ:

$$\tau = 1 + 2 \sum_{t=1}^{\infty} \rho(t)$$

We want to compute τ given $x_1, x_2, \dots, x_{N_e}$, the samples produced by the MCMC sampler.

Strategy:

1) Estimate $C(t)$ using "naive" estimator $\rightarrow \hat{C}(t)$
2) Compute $\hat{\rho}(t) = \dfrac{\hat{C}(t)}{\hat{\rho}(t)}$
3) Compute: $\boxed{\hat{\tau} = 1 + 2 \sum_{t=1}^{\omega \cdot \hat{\tau}} \hat{\rho}(t)}$

Note: $C(t), \rho(t), \tau$ are numbers

$\hat{C}(t), \hat{\rho}(t), \hat{\tau}$ are r.v. and we want that

$$E[\hat{C}] \rightarrow C \qquad \qquad \mathrm{Var}(\hat{C}) \rightarrow 0$$
$$E[\hat{\rho}] \rightarrow \rho \qquad \text{and} \qquad \mathrm{Var}(\hat{\rho}) \rightarrow 0 \qquad \text{as } N_e \rightarrow \infty.$$
$$E[\hat{\tau}] \rightarrow \tau \qquad \qquad \mathrm{Var}(\hat{\tau}) \rightarrow 0$$

The above/below estimators have the property, but be careful $\hat{\tau}_{BS} = 1 + 2 \sum_{t=1}^{N/2} \hat{\rho}(t)$ has variance that does not go to zero, hence it should not be used.

Estimators for $\hat{C}$:

$$\hat{C}(t) = \frac{1}{N_e - t - 1} \sum_{j=1}^{N_e} (x_j - \hat{\mu})(x_{j+t} - \hat{\mu})$$

$$\hat{\mu} = \frac{1}{N_e} \sum_{j=1}^{N_e} x_j$$