STAT 513/413: Lecture 17
Markov chain Monte Carlo: practical use

(including crash course in Bayesian statistics)

Why bother at all?

The possibility of generating from $\ell(x)$

without knowing the exact value of $\int \ell(x)\,\mathrm{d}x$

Who needs that??

Bayesian statistics ("inverse probability")

We observe x, as an outcome of a random variable X that has probability distribution depending on ϑ

Unlike x, the value of ϑ is unknown - but we would like to say something about it on the basis of x

To this end, we need to give ϑ a probability distribution, $g(\vartheta)$

- which does not have an interpretation in any repeated or potentially repeated event, but rather alluding to our uncertainty about ϑ; in particular, $g(\vartheta)$ is interpreted as the "initial", *prior* distribution: quantifies our uncertainty *before* observing any x

The distribution of x can be then viewed as $f(x|\vartheta)$, conditional distribution of x given ϑ

(all of those may be densities or probability mass functions)

We look then at $g(\vartheta|x)$, the conditional distribution of ϑ given x - which expresses our uncertainty after observin x

We are thus "inverting the probability":

start with $f(x|\vartheta)$ (and $g(\vartheta)$), end up with $g(\vartheta|x)$

How do we do that?

The Bayes theorem

The probabilistic definition of conditional probability/density says:

$$f(x|\vartheta) = \frac{f(x,\vartheta)}{p(\vartheta)}$$ where $f(x,\vartheta)$ is the *joint* distribution of X and ϑ

(Note: $f(x,\vartheta)$ is quantity different from $f(x|\vartheta)$)

We have then also: $$g(\vartheta|x) = \frac{f(\vartheta,x)}{f(x)} = \frac{f(x,\vartheta)}{f(x)}$$

because the joint distribution of ϑ and x is the same as the joint distribution of x and ϑ

How do we obtain $f(x)$, the (marginal) distribution of x?

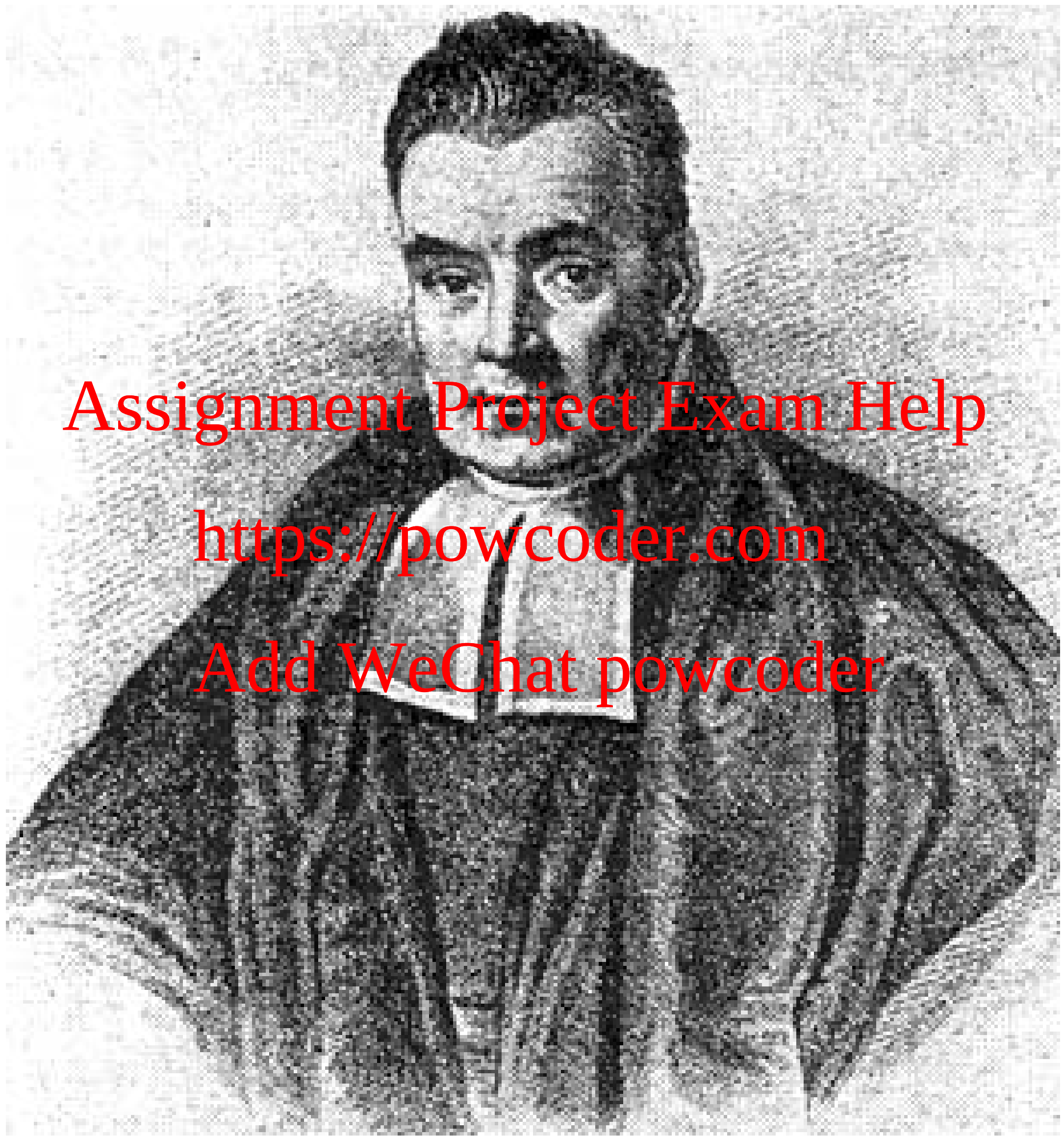
$$f(x) = \int f(x,\vartheta)\, d\vartheta$$ (replace integral by sum if necessary)

Thus:

$$g(\vartheta|x) = \frac{f(x,\vartheta)}{f(x)} = \frac{f(x,\vartheta)}{\int f(x,\vartheta)\, d\vartheta} = \frac{f(x|\vartheta)g(\vartheta)}{\int f(x|\vartheta)g(\vartheta)\, d\vartheta}$$

Numerator is easy - but denominator may be difficult

Reverend Thomas Bayes



Pierre-Simon Laplace

Example

We observe the outcomes of n 0-1 independent random variables X_i, which have the same probability $\vartheta = P[X_i = 1]$; we are interested only in the number of 1's among those

Our X is thus the sum $$X = \sum_i X_i$$

which has binomial distribution $$f(x|\vartheta) = \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x}$$

We assume ϑ to have the beta distribution $B(\alpha, \beta)$

$$g(\vartheta) = \frac{\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}}{B(\alpha, \beta)}$$ for $\vartheta \in [0, 1]$ (otherwise $g(\vartheta) = 0$)

Why? Good question... What is $B(\alpha, \beta)$?

Well, something that makes $g(\vartheta)$ integrating to 1

- otherwise, we do not have to worry about it...

... and be a bit more relaxed

A blasé way of doing Bayesian derivations

A wondrous symbol: $\propto$ reads "is proportional to" and means "up to a constant, dependent on context, equal to"

$$g(\vartheta) \propto \vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}$$

$$f(x|\vartheta) \propto \vartheta^{x}(1-\vartheta)^{n-x}$$

$$f(x,\vartheta) \propto f(x|\vartheta)g(\vartheta) \propto \vartheta^{x+\alpha-1}(1-\vartheta)^{n-x+\beta-1}$$

$$g(\vartheta|x) \propto f(x,\vartheta)$$

really: as $g(\vartheta|x) = \dfrac{f(x,\vartheta)}{f(x)}$ and $f(x)$ is with respect to ϑ only a constant (albeit depending on x)

$$g(\vartheta|x) \propto \vartheta^{x+\alpha-1}(1-\vartheta)^{n-x+\beta-1}$$

Now you perhaps can guess why $g(\vartheta)$ was selected $B(\alpha,\beta)$: merely out of convenience. Which distribution is $\propto \vartheta^{x+\alpha-1}(1-\vartheta)^{n-x+\beta-1}$? It is $B(\alpha+x,\beta+n-x)$

There is no need to calculate $\displaystyle\int \binom{n}{x}\vartheta^{x}(1-\vartheta)^{n-x}\frac{\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}}{B(\alpha,\beta)}\,d\vartheta$

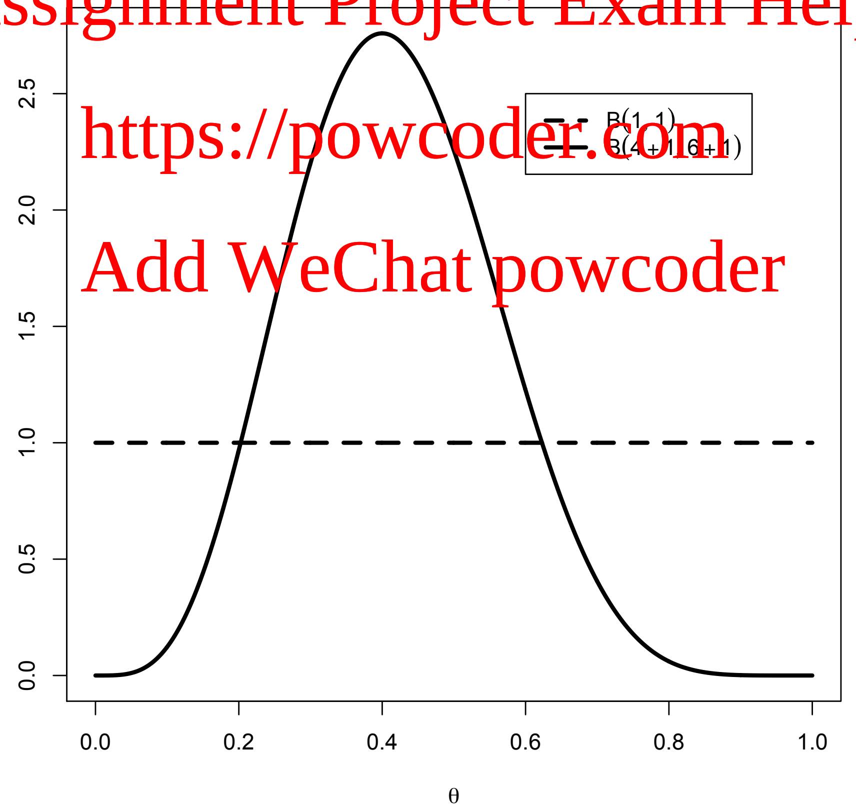
For instance

Out of 10 trials, we observed 4 ones and 6 zeros

For $g(\vartheta)$, we put $B(1,1)$ - assuming "we know nothing":

$$g(\vartheta) \propto \vartheta^{1-1}(1-\vartheta)^{1-1} = 1$$

- which is nothing but uniform distribution on $[0,1]$



So what?

Well, these are somewhat lucky, and not that frequent circumstances; albeit in this case it can be argued that the family of beta distributions has many members, allowing thus for flexible "expression of initial uncertainty" about ϑ, it it is quite clear that there will be many more instance which will not go that smoothly

Once again, first steps are easy: once $g(\vartheta)$ and $f(x|\vartheta)$ are specified, then they are multiplied

But the results has to be integrated, which may be difficult - even numerically (and even Monte Carlo)

Yes, but why do we need to integrate it at all? The product is proportional to $g(\vartheta|x)$ - so it may perhaps tell us something about ϑ itself?

Well, it can indicate the shape ("where ϑ is concentrated"), but if we need to calculate something out of $g(\vartheta|x)$

For instance, the mean:

it is known that the mean of $B(\alpha, \beta)$ is $\dfrac{\alpha}{\alpha + \beta}$

If only we could do Monte Carlo

Imagine that we know $g(\vartheta|x)$, and can generate random numbers Z_i out of this distribution

Then we can get estimates of its expected value: $$\frac{1}{n}\sum_i Z_i$$

Well, we know $g(\vartheta|x)$ *up to a constant*...

... but neither inversion nor acceptance/rejection method would work without the knowledge of the constant!

But Markov chain Monte Carlo algorithms do!!!

Another example

Suppose the data come from the distribution

$$f(x|\vartheta) = \frac{1}{c\sqrt{2\pi}} e^{-\frac{(x-\vartheta)^2}{2c}} \qquad \text{that is, } N(\vartheta, c^2)$$

and the prior distribution of ϑ is

$$g(\mu) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{(\vartheta-m)^2}{2s^2}} \qquad \text{that is, } N(m, s^2)$$

where c, m, s are considered known constants

A possible story: measuring IQ. The population IQ is normally distributed, centered at $m = 100$ (by definition), with $s = 15$ (approximately; US 1970's). We are measuring IQ by tests: their outcomes are $x_1, \ldots, x_n$, each with standard deviation c

(In practice we often observe note one x, but several of those. However, the notation is easier with one.)

So... the result is sort of intuitive

$$
\begin{aligned}
f(x|\vartheta) &\propto e^{-\frac{(x-\vartheta)^2}{2c^2}} \\
g(\vartheta) &\propto e^{-\frac{(\vartheta-m)^2}{2s^2}} \\
f(x,\vartheta) &\propto e^{-\frac{(x-\vartheta)^2}{2c^2}} e^{-\frac{(\vartheta-m)^2}{2s^2}} = e^{-\frac{(x-\vartheta)^2}{2c^2}-\frac{(\vartheta-m)^2}{2s^2}} = e^{A\vartheta^2+B\vartheta+C}
\end{aligned}
$$

with $A < 0$ - another normal distribution

The result is $N\left(\frac{\frac{x}{c^2}+\frac{m}{s^2}}{\frac{1}{c^2}+\frac{1}{s^2}}, \frac{1}{\frac{1}{c^2}+\frac{1}{s^2}}\right) = N\left(\frac{\frac{x}{c^2}+\frac{m}{s^2}}{\frac{1}{c^2}+\frac{1}{s^2}}, \frac{1}{\frac{1}{c^2}+\frac{1}{s^2}}\right)$

More generally, if we have n tests with outcomes $x_1, x_2, \ldots, x_n$, each with standard deviation c, then the result is

$$
N\left(\frac{\frac{n\bar{x}}{c^2}+\frac{m}{s^2}}{\frac{n}{c^2}+\frac{1}{s^2}}, \frac{1}{\frac{n}{c^2}+\frac{1}{s^2}}\right) = N\left(\frac{\frac{\bar{x}}{c^2}+\frac{m}{ns^2}}{\frac{1}{c^2}+\frac{1}{ns^2}}, \frac{1}{\frac{n}{c^2}+\frac{1}{s^2}}\right)
$$

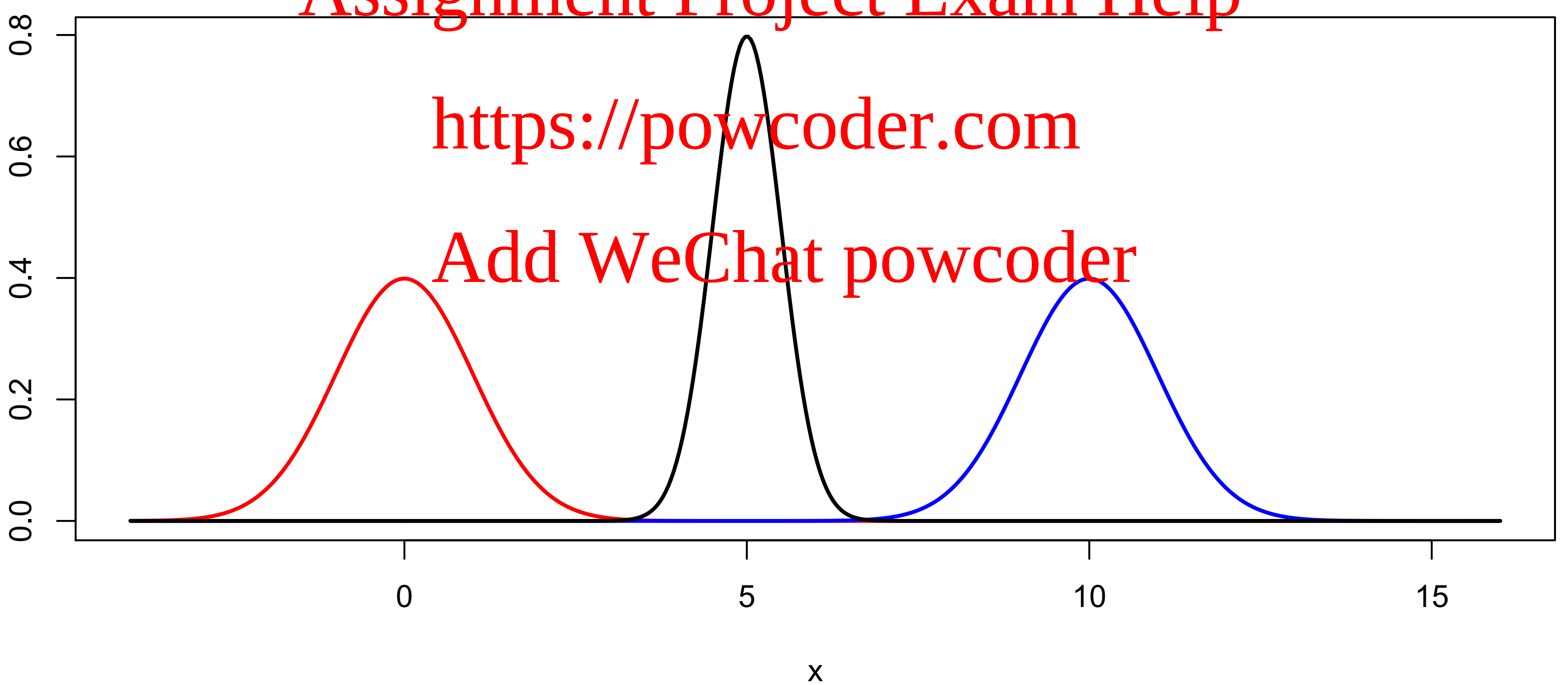
where $\bar{x}$ is the average of the x_i's - and this has a natural interpretation

But sometimes a bit funny

Suppose that $ns = c^2$; then the result is $N\left(\frac{\bar{x}+m}{2}, \frac{c^2}{2}\right)$

If, say, $n = 1$, $s = c = 1$, $m = 0$, and $x = 10$

then the result looks like this



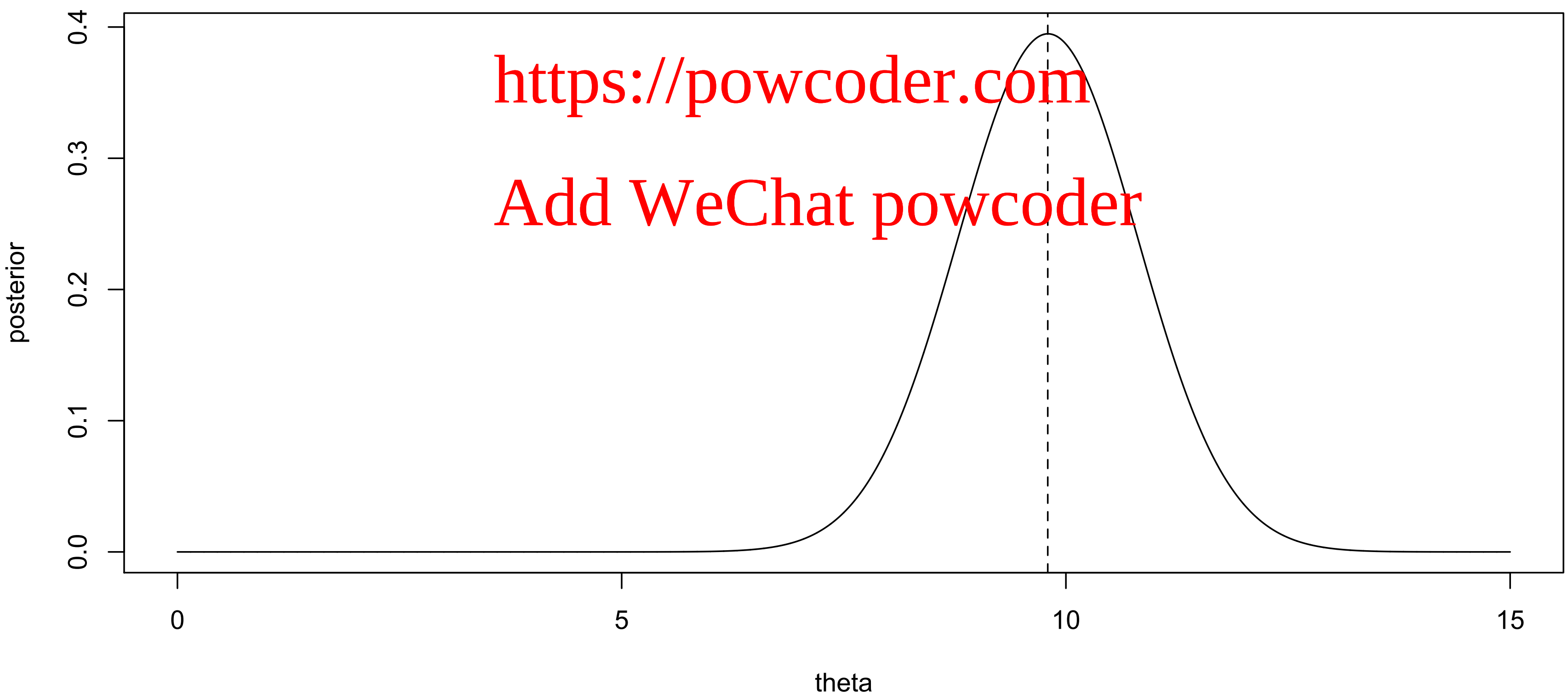
Have you got a good story for that? (There are bad stories...)

A fix?

If, instead of the normal distribution used before, $g(\vartheta)$ is a t distribution, say, with 1 degree of freedom (the Cauchy distribution)

$$g(\vartheta) = \frac{1}{\pi(1+\vartheta^2)} \qquad \text{(this distribution is centered at 0)}$$

and as before, $x = 10$, where x is $N(\vartheta, 1)$, then the result is



Markov chain Monte Carlo: concluding remarks

But such fixes have to be done numerically: and Markov chain Monte Carlo is a convenient vehicle for that, as it allows to generate random numbers following a density $c\ell(x)$ without knowing c.

It is still a methodology in development; compared to classical Monte Carlo methods, there are still a couple of aspects that have to be paid attention to, and adequately tuned.

Burn-in. The generated random numbers follow the desired distribution closely only later in the sequence - thus the initial random numbers are discarded. However, it is hard to say exactly how many.

Convergence. Unlike in the independent case, there are no performance guarantees like Chebyshev or Hoeffing inequalities. To assess whether the generated sequence is long enough, some empirical criteria have been proposed. (The book of Rizzo mentions Gelman-Rubin.)

There are also other, more specific aspects to be tuned. For instance, when Metropolis algorithm uses random walk as $q_x(y)$ (a "proposal transition distribution"), then it is important to scale it (set the length of the average step) so that it is neither too small nor too big (which is expressed by the rejection rate).

Monte Carlo: concluding remarks

We bid a farewell to random numbers at this point - in a hope that what we have seen so far reinforced certain principles, rather than gave mere recipes, how these numbers can be effectively used in numerical computations. In particular, we hope that we understand more the following aspects now:

- how these numbers are generated by computers

- how they can be used in numerical experiments that aim at exploration of the properties of probabilistic phenomena that are hard to be obtained in theoretical fashion, or by classical numerical methods

- how they can be used for numerical calculations that may be hard to accomplish by classical numerical methods, like the computation of integrals; and how these methods can be improved

- how to obtain, in certain cases, validation criteria, precision estimates, and performance guarantees

- and also, what are the principles of some more specific, and up-to-date technologies, applicable in specific situations - like Markov chain Monte Carlo methods