Monte-Carlo Methods and Particle Filters: The Gibbs Sampler

Peter Jan van Leeuwen *

Department of Meteorology, University of Reading, United Kingdom
February 6, 2011

* *Corresponding author address:* Peter jan van Leeuwen, Department of Meteorology, University of Reading, United Kingdom
E-mail: p.j.vanleeuwen@reading.ac.uk

Contents

4. The Gibbs sampler

a. Basic methodology

To sample from a complicated joint density, which typically is the posterior density in a data-assimilation problem, one can draw samples from the full marginals. Let us for simplicity write this joint density as $p(x_1, ..., x_d)$. The Gibbs sampler works as follows:

1) Choose a first sample $x^0 = (x_1^0, x_2^0, ..., x_d^0)^T$ from some initial density p^0.

2) Obtain a new sample x^n from x^{n-1} by sampling the values:

$$\begin{aligned} x_1^n &\sim p(x_1^n | x_2^{n-1}, ..., x_d^{n-1}) \\ x_2^n &\sim p(x_2^n | x_1^n, x_3^{n-1}, ..., x_d^{n-1}) \\ &... \\ x_d^n &\sim p(x_n^n | x_1^n, ..., x_{d-1}^n) \end{aligned} \tag{1}$$

3) Change n to $n+1$ and proceed to 2) until convergence.

Note that each new component x_i^n is used immediately to draw the next component x_{i+1}^n 'Convergence' here means that we have reached the stationary joint distribution $p(x_1, ...x_d)$, and one sample is produced. More samples can be produced by rerunning the scheme above, which would be prohibitively expensive in high-dimensional systems. However, other more efficient schemes are possible, as discussed later.

To prove convergence of the Gibbs sampler we first notice that it forms a Markov chain, since each new sample only depends directly on the previous one. Also, the chain is homogeneous since the transitions do not depend on the iteration index explicitly. Next, the distribution $p(x)$ is invariant under each of the Gibbs sampling steps, and hence under the whole Markov chain. This follows from the fact that sampling from $p(x_i|x_{-i})$ leaves the marginal $p(x_{-i})$ invariant because the values x_{-i} ar not changed. (Note x_{-i} denotes the full state vector with x_i excluded.) Furthermore, each step samples from the correct marginal by definition. Because the conditionals and the marginals define the joint, we see that the joint itself is invariant.

We also have to require that the Gibbs sampler is ergodic. A sufficient condition for ergodicity is that none of the conditional densities are zero anywhere. If that is the case, any point in state space can be reached from any other point in a finite number of steps involving one update of each of the component variables.

b. Sampling strategies

Several sampling techniques to improve convergence rates have been proposed:

1) generate n chains using different starting points, each with m iterations to reach the stationary distribution (see chapter Markov Chains). Costs mn. Samples are independent.

2) Generate one single long chain and use the iterates after m iterates to the stationary density. After these m iterates use the next n iterates as samples. Costs $m + n$. Samp[les are not independent, but any statistic derived from them is ok.

3) Sample only every kth value after the burn in period m. Costs $m + kn$. Independent samples if k is large enough. Large enough is determined from the chain autocorrelation.

4) Combine 1) and 2), so l chains with $l < 10$, say, each of length $m + n/l$ (burn in plus n/l samples. Costs $lm + n$. Partly independent samples.

5) Combine 1) and 3), so l chains with $l < 10$, say, and keep each kth sample after m iterates. Costs $lm + kn$.

There are efficiency losses in sampling from a single chain, but it is much cheaper. Strategy 1) is not recommended because it is too expensive. If the posterior is relatively smooth without strong local maxima a single chain will do fine. However, if the high-probability areas are almost disconnected a very long chain is needed to jump to another high-probability area, so a single chain is ineffective.

In very high-dimensional spaces one typically encounters in the geosciences, one tries to generate a first sample directly on the stationary distribution. One of the possibilities is to first generate a 4DVar solution (which can be in a local minimum) and start the Gibbs sampler from there. This has not been explored in any depth in the geosciences yet...

c. Convergence diagnostics

A very difficult problem, especially in high-dimensional spaces, is to determine when enough samples from the stationary distribution have been chosen for e.g. the mean of the samples to have converged. Strictly speaking convergence is a property of the MC chain, regarded as a sequence of random variables. In practice however, one has a single realisation of the chain. Still convergence of a single realisation is a useful concept, and we will use that concept in the following discussion. Several simple techniques to identify convergence that can be applied in high-dimensional systems have been developed.

1) Plot the time series of a few variables in the chain and decide by eye if when the series becomes stationary. One of the questions is how long should we wait? If the chain 'hangs around' a local peak for some time a late transition to another peak can be missed entirely.

2) Plot average values for certain variables, such as the mean and the variance of a variables, as function of the iteration index n. For convergence these quantities should become independent of n. Unfortunately, this independence is no guarantee for convergence, for instance when the chain is sampling from a local peak.

3) When several chains are generated simultaneously one can plot histograms at certain values of n and compare them. Convergence needs these histograms to be the same. This methods needs sufficient simultaneous chains, and is probably out of the question for high-dimansional applications.

4) When several (say m) chains are generated check the variance of the means from each series against that within each series. The former is defined as:

$$B = \frac{1}{m-1} \sum_{i=1}^{m} \left(\overline{f(x_i)} - \overline{f(x)} \right)^2 \tag{2}$$

and the latter is

$$W = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(f(x_i^{(j)}) - \overline{f(x_i)} \right)^2 \tag{3}$$

If the chain has not reached equilibrium B will be too large, and W will be too small. It has been suggested to assume convergence when

$$R = \sqrt{(1 - 1/n) + B/(nW)} \leq 1.2 \tag{4}$$

This statistic can be used on any function of the variables, but also on the complete distribution by taking $f(x) = -2 \log p(x)$.

d. Scanning strategies

Up to know we have considered only scanning a new sample in given fixed order. However, several other strategies are possible:

1) Instead of sampling one component of the state vector at a time one can sample a larger subset of the state vector. In high dimensional state spaces that is certainly recommended, especially when components are highly correlated.

2 Instead of sampling components of the state vector in a fixed order, a random order can be chosen, or a fixed order that depends on a random variable. In some cases it has been shown that these modifications lead to better spread of the samples.

3 Instead of sampling all components from the complete conditional densities it is sometimes faster and more efficient to sample from densities conditioned on only part of the rest of the state vector. For example, in a 3-D space $\{x_1, x_2, x_3\}$ one could sample according to:

$$\begin{aligned} x_1^n &\sim p(x_1^n | x_2^{n-1}, x_3^{n-1}) \\ x_2^n &\sim p(x_2^n | x_1^n) \\ x_3^n &\sim p(x_n^n | x_1^n, x_2^n) \end{aligned} \tag{5}$$

until convergence.

4) When variables are highly correlated the standard Gibbs sampler is not efficiently sampling the full distribution. Coordinate transformations can solve this issue by making the distribution more isotropic, leading to more efficient sampling. One way to implement this is to determine the covariance from the samples and calculate its (approximate) square root A. The variables are then transformed as $x' = A^{-1}x$, leading to approximate independence of the variables.

5) Or one can choose the scan direction randomly, referred to as the *hit-and-run* algorithm. With a uniform distribution for the directions this is known as the *Hypersphere directions algorithm*. Often, however, sampling along arbitrary directions is not easy, and the ordinary Gibbs sampler is preferable.

It is sometimes more efficient to employ the reverse Gibbs sampler, in which new samples are generated using a forward and a backward scan for each sample.

One can also selec the scan order randomly.

e. Accuracy of the sample estimates

When N samples have been generated from the stationary distribution (so after the 'burn in' period) the mean of any function f of the samples can be calculated as:

$$E[f(x)] = \int f(x)p(x)\,dx \approx \frac{1}{N}\sum_i f(x^i) = \overline{f(x)} \tag{6}$$

The accuracy of this estimate depends on the dependence of the samples. Define the *autocovariance* as:

$$R(s) = E\left[\left(f(x^n) - E[f(x)]\right)\left(f(x^{n+s}) - E[f(x)]\right)\right] \tag{7}$$

For a stationary chain R does not depend on n. Note that $R(0) = \sigma^2$, $R(-s) = R(s)$, and $|R(s)| \leq \sigma^2$. Finally, when the samples x^i are independent $R(s) = 0$ for $s \neq 0$. The autocovariance can be used to express the variance as follows:

$$\begin{aligned} Var(\overline{f(x)}) = E[[\overline{f(x)} - E[f(x)])^2] &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} f(x^i) - E[f(x)]\right)^2\right] \\ &= \frac{1}{N^2}\sum_{i,j=1}^{N} E\left[\left(f(x^i) - E[f(x)]\right)\left(f(x^j) - E[f(x)]\right)\right] \\ &= \frac{1}{N^2}\sum_{i,j=1}^{N} R(j-i) \\ &= \frac{1}{N}\sum_{-N<s<N}\left(1 - \frac{|s|}{N}\right)R(s) \end{aligned} \tag{8}$$

If the samples are independent this reduces to

$$Var(\overline{f(x)}) = \frac{\sigma^2}{N} \tag{9}$$

For dependent samples $R(s)$ for $s \neq 0$ tends to be positive, and can contribute significantly to the variance. So using the expression for the independent variables leads to a wrong conclusion on the actual accuracy. For large N the variance equation can be written as:

$$Var(\overline{f(x)}) = \frac{1}{N}\left[\sigma^2 + 2\sum_{s=1}^{\infty} R(s)\right] = \frac{\sigma^2}{N/\tau} \tag{10}$$

in which $\tau = 1 + 2\sum_1^\infty \rho(s)$ where $\rho(s)$ is the autocorrelation. *tau* gives a measure of the number of dependent MC samples needed for one independent MC sample. (Note that it is possible, but uncommon, that $\tau < 1$.)

Actual numbers for the estimate of the accuracy of the $\overline{f(x)}$ can be obtained by using the estimate for $R(s)$ (or σ^2) from the time series. Unfortunately, since the accuracy of these estimates is not known, it is unclear what these expressions will mean. However, often it is the best one can do.

Another possibility is to determine the spectrum of the time series an using $Var(\overline{f(x)}) = S(0)/N$ in which $S(0)$ is the spectral density at zero frequency. Since the spectrum is symmetric around zero , so $S(0)$ is an extremum, the spectra should not be smoothed. To avoid spectral noise one can also fit AR or ARMA models to the time series and determine their spectra. Obviously, the accuracy of the estimate for $S(0)$ will depend on that of the fit.

f. When to use the Gibbs sampler?

As a general rule the Gibbs sampler can be used efficiently in two cases:

1) When the dimension of the system is small the conditional densities can be evaluated before hand, and sampling is just drawing from this small-dimensional density.

2) When the conditional densities are given in parametric form and samples can easily be generated from them.

If we now consider a large-scale geophysical application, e.g. numerical weather forecasting, the first case is not the case, and the second is unclear. It will never be easy to generate samples from conditional of the posterior pdf, but the effort has to be compared with other data assimilation methods like 4DVar or the Ensemble Kalman filter. We know these are expensive too. As far as I know, Gibbs has not been tried with all its tricks on large-scale geophysical problems.