Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

Department of mathematics
KTH Royal Institute of Technology
johawes@kth.se

Lecture 10
Markov chain Monte Carlo III
26 April 2017

Plan of today's lecture

Outline

Last time: the Metropolis-Hastings (MH) algorithm

- We assumed that were able to simulate from a transition density $r(z \mid x)$, referred to as the proposal kernel, on X.
- The MH algorithm simulates recursively a sequence of draws (X_k), forming a Markov chain on X, through the following mechanism: given X_k,
  - draw $X^* \sim r(z \mid X_k)$ and
  - set $X_{k+1} = \begin{cases} X^* & \text{w. pr. } \alpha(X_k, X^*) \overset{\text{def}}{=} 1 \wedge \dfrac{f(X^*)r(X_k \mid X^*)}{f(X_k)r(X^* \mid X_k)}, \\ X_k & \text{otherwise.} \end{cases}$

  (Here we used the notation $a \wedge b \overset{\text{def}}{=} \min\{a, b\}$.) The scheme is initialized by drawing X_1 from some arbitrary initial distribution χ.

Last time: the MH algorithm: pseudo-code

draw $X_1 \sim \chi$;
for $i = 1 \to (N-1)$ **do**
  draw $X^* \sim r(z \mid X_k)$;
  set $\alpha \leftarrow 1 \wedge \frac{f(X^*) r(X_k | X^*)}{f(X_k) r(X^* | X_k)}$;
  draw $U \sim \mathrm{U}(0,1)$;
  if $U \leq \alpha$ **then**
    $X_{k+1} \leftarrow X^*$;
  else
    $X_{k+1} \leftarrow X_k$;
  end
end
set $\tau_N^{\mathrm{MCMC}} \leftarrow \sum_{k=1}^{N} \phi(X_k)/N$;
return τ_N^{MCMC}

Last time: different types of proposal kernels

- There are a number of different ways of constructing the proposal kernel r.
- The three main classes are
  - independent proposals,
  - symmetric proposals, and
  - multiplicative proposals.

Last time: convergence of the MH algorithm

- The following results are fundamental:

Theorem (detailed balance of the MH sampler)

The MH sampler satisfies detailed balance for the target density f.

Corollary (global balance of the MH sampler)

The Markov chain generated by the MH sampler allows f as a stationary distribution.

- In addition, one may prove, under weak assumptions, that the MH algorithm is also geometrically ergodic, implying that it satisfies an LLN.

Outline

The Gibbs sampler

- In the following,
  - assume that the space X can be divided into m blocks, i.e., $x = (x^1, \ldots, x^m) \in$ X, where each block may be vector-valued itself.
  - assume that we want to sample a multivariate distribution f on X.
  - denote by x^k the kth component of x and by $x^{-k} = (x^\ell)_{\ell \neq k}$ the set of remaining components.
  - denote by $f_k(x^k \mid x^{-k}) = f(x)/\int f(x)\,dx^k$ the conditional distribution of X^k given the other components $X^{-k} = x^{-k}$ and
  - assume (initially) that it is easy to simulate from $f_k(x^k \mid x^{-k})$ for all $k = 1, \ldots, m$.

The Gibbs sampler (cont.)

- The Gibbs sampler simulates recursively a sequence of values (X_k), forming a Markov chain on X, using the following mechanism.
- Given $X_k = (X_k^1, \ldots, X_k^m)$,
  - draw $X_{k+1}^1 \sim f_1(x^1 | X_k^2, \ldots, X_k^m)$,
  - draw $X_{k+1}^2 \sim f_2(x^2 | X_{k+1}^1, X_k^3, \ldots, X_k^m)$,
  - draw $X_{k+1}^3 \sim f_3(x^3 | X_{k+1}^1, X_{k+1}^2, X_k^4, \ldots, X_k^m)$,
  - ...
  - draw $X_{k+1}^m \sim f_m(x^m | X_{k+1}^1, X_{k+1}^2, \ldots, X_{k+1}^{m-1})$.
- In other words, at the ℓth round of the cycle generating X_{k+1}, the ℓth component of X_{k+1} is updated by simulation from its conditional distribution given all other components.

Convergence of the Gibbs sampler

- As for the MH algorithm, the following holds true.

Theorem

The chain (X_k) generated by the Gibbs sampler has f as stationary distribution.

- In addition, one may prove, under weak assumptions, that the Gibbs sampler is also geometrically ergodic, implying that

$$\tau_N^{\text{MCMC}} = \frac{1}{N} \sum_{k=1}^{N} \phi(X_k) \to \tau \quad \text{as} \quad N \to \infty.$$

Example: a tricky bivariate distribution

- Suppose that we want to sample the distribution on $\{0, 1, 2, \ldots, n\} \times (0, 1)$ given by

$$f(x, y) \propto \frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}.$$

which is very complex and hard to sample from.

- The conditional distributions are however simple; indeed
  - $X \mid Y = y \sim \mathrm{Bin}(n, y)$,
  - $Y \mid X = x \sim \mathrm{Beta}(x + \alpha, n - x + \beta)$.
- Thus, the problem of sampling $f(x, y)$ can be perfectly cast into the framework of the Gibbs sampler.

Example: a tricky bivariate distribution (cont.)

```
burn_in = 1000;
M = N + burn_in;
X = zeros(1,M);
Y = X;
X(1) = 5;
Y(1) = 0.5;
for k = 1:(M - 1),
    x = binornd(n,Y(k));
    X(k + 1) = x;
    Y(k + 1) = betarnd(x + alpha,n - x + beta);
end
```

Example: a tricky bivariate distribution (cont.)

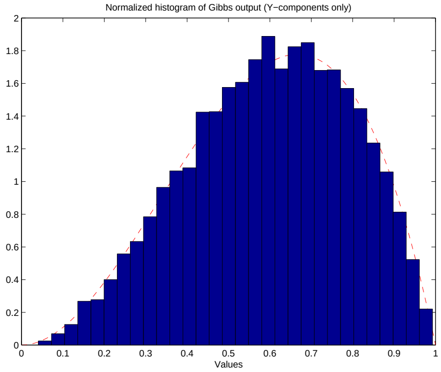

Figure: Comparison between the true density and the histogram of Y_k, $k = 1001, \ldots, 11000$.

Outline

1. Last time: the Metropolis-Hastings (MH) algorithm
2. The Gibbs sampler (Ch. 5.4)
3. Variance of MCMC samplers

Variance of MCMC estimators

- As mentioned, the MH and Gibbs samplers are geometrically ergodic, implying an LLN for the resulting estimators. In addition, one may establish a CLT.
- For this purpose, let

$$r(\ell) = \lim_{n\to\infty} \mathbb{C}(\phi(X_{n+\ell}), \phi(X_n))$$

be the covariance function of the MCMC chain at stationarity.

Variance of MCMC estimators (cont.)

- The following holds true.

Theorem

For the MCMC samplers discussed above it holds that

$$\sqrt{N}(\tau_N^{MCMC} - \tau) \stackrel{d.}{\longrightarrow} \mathrm{N}(0, \sigma^2) \quad \textit{as} \quad N \to \infty,$$

where

$$\sigma^2 = r(0) + 2\sum_{\ell=1}^{\infty} r(\ell).$$

Estimating the variance of MCMC samplers

- For i.i.d.-based Monte Carlo integration we used the sample variance (Matlab: `var`) to estimate $\mathbb{V}(\phi(X))$.
- However, now we need the entire covariance function $r(\ell)$. A number of different approximative solutions are possible, e.g.,
  - assume a parametric form of the covariance function, usually that of an AR process of low order, and estimate it,
  - use only samples that are far apart, ensuring approximate independence,
  - divide the samples into blocks that are large enough to be approximately independent. Then calculate averages of each block and use these to estimate the standard deviation.

Next lecture

- Block-based estimation of the MCMC asymptotic variance,
- Statistics!