Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

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

Lecture 11
MCMC for Bayesian computation
28 April 2017

Peer review form for project reports

	weak	satisfactory	strong	comments
Contents *(covering, relevance)*				
Presentation *(guiding the reader, flow of ideas, typesetting and spelling)*				
Evidence *(credibility, correctness)*				
Overall effectiveness				

Peer review (cont.)

- The review form (in LaTeX) is available through the course home page.
- For identifiability, start each review by quoting a part of the first sentence of the corresponding report.
- Provide at least three substantial comments for each item (contents, presentation, etc.)
- The peer review is performed group-wise, i.e., the members of a group do not write seperate reports.
- Email your reviews to the lecturer (`johawes@kth.se`) by Tuesday 9 May, 23:59.

Outline

1 Last time: Gibbs sampling and estimation of variance

2 Hybrid MCMC samplers

3 Stochastic modeling and Bayesian inference

The Gibbs sampler

- 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 itself vector-valued.
- 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 $X^{-k} = x^{-k}$.
- Assume 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 a sequence (X_k) of values forming a Markov chain on X using with the following mechanism: given X_k,
  - draw $X_{k+1}^1 \sim f_1(x^1|X_k^2, \dots, X_k^m)$,
  - draw $X_{k+1}^2 \sim f_2(x^2|X_{k+1}^1, X_k^3, \dots, X_k^m)$,
  - draw $X_{k+1}^3 \sim f_3(x^3|X_{k+1}^1, X_{k+1}^2, X_k^4, \dots, X_k^m)$,
  - $\dots$
  - draw $X_{k+1}^m \sim f_m(x^m|X_{k+1}^1, X_{k+1}^2, \dots, X_{k+1}^{m-1})$.
- 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

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

Variance of MCMC estimators

- As mentioned, the MH and Gibbs samplers are geometrically ergodic, implying an LLN in each case.
- In addition, one may establish the following CLT. Let

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

be the covariance function of (X_k) at stationarity.

Theorem

Assume that

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

Then

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

Estimating asymptotic variance using blocking

- Use $N = nK$ samples and write

$$\tau_N^{\text{MCMC}} = \frac{1}{N} \sum_{k=1}^{N} \phi(X_k) = \frac{1}{n} \sum_{\ell=1}^{n} T_\ell,$$

where

$$T_\ell \overset{\text{def}}{=} \frac{1}{K} \sum_{m=(\ell-1)K+1}^{\ell K} \phi(X_m), \quad \ell = 1, 2, \ldots, n.$$

- If the blocks are large enough we can view these as close to independent and identically distributed.

Estimating asymptotic variance using blocking (cont.)

- We may thus expect the standard CLT to hold at least approximately, implying that

$$\mathbb{V}(\tau_N^{\text{MCMC}}) = \mathbb{V}\left(\frac{1}{n}\sum_{\ell=1}^{n} T_\ell\right) \approx \frac{\mathbb{V}(T_1)}{n},$$

where $\mathbb{V}(T_1)$ can be estimated using the standard estimator

$$\mathbb{V}(T_1) \approx \frac{1}{n-1}\sum_{m=1}^{n}(T_m - \overline{T}_n)^2,$$

with $\overline{T}_n = \sum_{m=1}^{n} T_m/n$ denoting the sample mean. The latter is easily computed using MATLAB's `var` function.

Example: A tricky bivariate distribution (again)

- We let again (X, Y) have bivariate distribution

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

  on $\{0, 1, 2, \ldots, n\} \times (0, 1)$ and estimate the marginal expectation

  $$\tau = \mathbb{E}(Y)$$

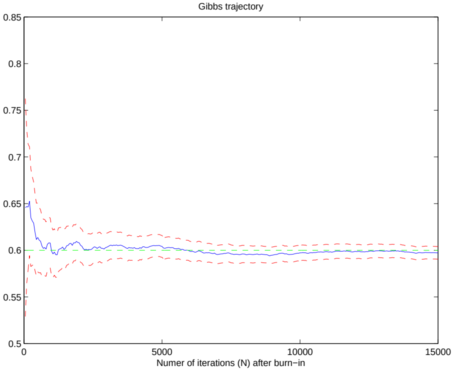
  using the output (X_k, Y_k) of the Gibbs sampler.
- In addition, we construct a 95% confidence bound on τ using the blocking method.

Example: A tricky bivariate distribution (again)

- In MATLAB:

```
K = 50; % block size
n = N/K; % number of blocks
T = zeros(1,n);
for k = 1:n, % take means over n blocks
    T(k) = mean(Y((burn_in + (k - 1)*K + 1):(burn_in + K*k)));
end
LB = tau - norminv(0.975)*std(T)/sqrt(n); % confidence bound
UB = tau + norminv(0.975)*std(T)/sqrt(n);
```

Example: A tricky bivariate distribution (again)



Outline

1 Last time: Gibbs sampling and estimation of variance

2 Hybrid MCMC samplers

3 Stochastic modeling and Bayesian inference

Hybrid MCMC samplers

- It is often very convenient to consider hybrids between Gibbs and MH:
  - Divide the space into blocks and aim for Gibbs sampling.
  - If it is possible to sample directly from the conditional distribution of a block, update according to Gibbs.
  - If it is not, just insert a local MH step instead!
- The resulting chain satisfies still global balance and is thus a valid MCMC sampler (referred to as the hybrid sampler or Metropolis-within-Gibbs).

Hybrid MCMC samplers (cont.)

- More specifically, assume that q_ℓ is some Markov transition density allowing $f_\ell(x^\ell \mid x^{-\ell})$ (i.e., the conditional density of the ℓ^{th} block) as a stationary distribution. The density q_ℓ may depend on $x^{-\ell}$.
- For instance, q_ℓ may be an MH kernel for $f_\ell(x^\ell \mid x^{-\ell})$ based on some proposal density r_ℓ.
- In the particular case where r_ℓ is the independent proposal $f_\ell(x^\ell \mid x^{-\ell})$ the acceptance probability becomes identically one, and we are back at a standard—ideal—Gibbs sub-step!

Hybrid MCMC samplers (cont.)

- We may now consider the generalized Gibbs scheme with one iteration (sweep) given by

$$\begin{pmatrix} X_k^1 \\ X_k^2 \\ X_k^3 \\ \vdots \\ X_k^m \end{pmatrix} \xrightarrow{q_1} \begin{pmatrix} X_{k+1}^1 \\ X_k^2 \\ X_k^3 \\ \vdots \\ X_k^m \end{pmatrix} \xrightarrow{q_2} \begin{pmatrix} X_{k+1}^1 \\ X_{k+1}^2 \\ X_k^3 \\ \vdots \\ X_k^m \end{pmatrix} \xrightarrow{q_3} \cdots \xrightarrow{q_m} \begin{pmatrix} X_{k+1}^1 \\ X_{k+1}^2 \\ X_{k+1}^3 \\ \vdots \\ X_{k+1}^m \end{pmatrix}.$$

- In order to show that one full iteration $X_k \to X_{k+1}$ allows f as a stationary distribution it is enough to show that each sub-step allows f as a stationary distribution (see E4, Problem 3).

Hybrid MCMC samplers (cont.)

- The ℓ^{th} sub-step follows the transition $q_\ell(\tilde{x}^\ell \mid x^\ell)\delta_{x^{-\ell}}(\tilde{x}^{-\ell})$.
- This transition density allows indeed f as a stationary distribution, as

$$\begin{aligned}
&\int f(x)q_\ell(\tilde{x}^\ell \mid x^\ell)\delta_{x^{-\ell}}(\tilde{x}^{-\ell})\,dx \\
&= \int \left[\int f_\ell(x^\ell \mid x^{-\ell})q_\ell(\tilde{x}^\ell \mid x^\ell)\,dx^\ell \right] f(x^{-\ell})\delta_{x^{-\ell}}(\tilde{x}^{-\ell})\,dx^{-\ell} \\
&= \int f_\ell(\tilde{x}^\ell \mid x^{-\ell})f(x^{-\ell})\delta_{x^{-\ell}}(\tilde{x}^{-\ell})\,dx^{-\ell} \\
&= \int f(\tilde{x}^\ell, x^{-\ell})\delta_{x^{-\ell}}(\tilde{x}^{-\ell})\,dx^{-\ell} \\
&= f(\tilde{x}).
\end{aligned}$$

Part II

MC methods for statistical inference

Statistical inference: data ⇒ knowledge

- ”Inference is the problem of turning data into knowledge, where knowledge often is expressed in terms of entities that are not present in the data per se but are present in models that one uses to interpret the data.”

  Committee on the Analysis of Massive Data: *Frontiers in Massive Data Analysis.* The National Academies Press, Washington D.C., 2013, p.3.

Outline

1 Last time: Gibbs sampling and estimation of variance

2 Hybrid MCMC samplers

3 Stochastic modeling and Bayesian inference

Overview

We will consider

- some literature,
- stochastic modeling,
- frequentist vs. Bayesian statistics

Alternative literature

- MCMC:
  - *Markov Chain Monte Carlo in Practice*, Gilks, Richardson & Spiegelhalter, 1996.
  - *Monte Carlo Statistical Methods*, Robert & Casella, 2005.
- Bayesian statistics:
  - *The Bayesian Choice*, Robert, 2001.
- Bootstrap (to be discussed on Friday and next week):
  - *Bootstrap Methods and Their Application*, Davison & Hinkley, 1997.
  - *An Introduction to the Bootstrap*, Efron & Tibshirani, 1994.

Stochastic modeling: frequentist approach

- In the frequentist approach to stochastic modeling, the setup is as follows.
- We observe *data* y.
- The data y is assumed to be an observation of a (typically multivariate) random variable Y with distribution $\mathbb{P}_0$.
- A statistical model is a set $\mathcal{P}$ of probability distributions that is assumed to contain $\mathbb{P}_0$.
- The largest possible model would be

$$\mathcal{P} = \{\text{all possible distributions } \mathbb{P} \text{ that could generate } y\}.$$

- An inference problem refers to the problem of selecting a distribution from $\mathcal{P}$ that fits the observed data y.

Stochastic modeling: frequentist approach (cont.)

- Commonly we restrict the set of distributions to a come from a parametric family

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\},$$

where Θ is called the parameter space.

- For instance,

$$\mathcal{P} = \{\text{all normal distributions with mean } \theta \text{ and variance } 1\}.$$

In this case, $\Theta = \mathbb{R}$ and the true parameter θ_0 is seen as having an unknown but fixed value.

Stochastic modeling: frequentist approach (cont'd)

- An estimate of θ_0 is formed using a function $\widehat{\theta}(y)$ of the data. The function $\widehat{\theta}(y)$ is called estimator. The estimate $\widehat{\theta}(y)$ (i.e. the value taken by the estimator) should be close to θ_0.
- Since y is a random sample from $\mathbb{P}_0$, the estimate $\widehat{\theta}(y)$ is a realization of the random variable $\hat{\theta}(Y)$.
- A common estimator is the maximum likelihood estimator (MLE), which is obtained as the parameter $\widehat{\theta}(y)$ maximizing the maximum likelihood function

$$\theta \mapsto f(y \mid \theta),$$

where y is the given observed data.
- The point estimate is often equipped with a 95% confidence interval.

Stochastic modeling: Bayesian approach

- In Bayesian inference, the setup is the following.
- Our uncertainty concerning the parameters θ is modeled by letting the parameters be random variables.
- Thus, a Bayesian model is the joint distribution $f(y, \theta)$ of Y and θ. By Bayes's formula,

$$f(y, \theta) = f(y \mid \theta) f(\theta).$$

- $f(y \mid \theta)$ is the likelihood that describes how the data Y behaves conditionally on the parameters θ.
- $f(\theta)$ is called the prior distribution and summarizes our prior belief about θ before observing Y.

Stochastic modeling: Bayesian approach (cont.)

- Since θ is viewed as a random variable, inference is based on the posterior (or a posteriori) distribution $f(\theta \mid y)$, i.e., the distribution of the parameters given the observed data.
- By Bayes's Formula:

$$f(\theta \mid y) = \frac{f(y, \theta)}{f(y)} = \frac{f(y \mid \theta) f(\theta)}{\int f(y \mid \theta') f(\theta') \, d\theta'} \propto f(y \mid \theta) f(\theta).$$

Bayesian vs. frequentist statistics

- Bayesian inference is done using the posterior $f(\theta \mid y)$.
- Frequentist inference uses the likelihood $f(y \mid \theta)$.
- A Bayesian makes statements about the relative evidence for parameter values given a dataset.
- A Frequentist compare the relative chance of datasets given a parameter value.

Example: methicillin-resistant Staphylococcus aureus

- Suppose a hospital has around 200 beds occupied each day and that we want to know the underlying risk that a patient will be infected by MRSA (methicillin-resistant Staphylococcus aureus).
- Looking back at the first six months of the year, we count $y = 20$ infections in 40,000 bed-days.
- Let θ be the expected number of infections per 10,000 bed-days. A reasonable model is that y is an observation of $Y \sim \mathrm{Po}(4\theta)$.

Example: MRSA (cont.)

- Frequentist approach:
  - MLE: $\widehat{\theta}(y) = y/4 = 20/4 = 5$.
  - An approximate confidence interval based on a normal approximation is given by

$$\widehat{\theta}(y) \pm \lambda_{\alpha/2}\sqrt{\frac{\widehat{\theta}(y)}{4}} = (2.81, 7.19).$$

  - A hypothesis test of $\mathcal{H}_0 : \theta = 4$ vs. $\mathcal{H}_1 : \theta > 4$ can be carried through using the direct method. This gives

$$\begin{aligned}&\mathbb{P}\left(\text{get what we got or worse under } \mathcal{H}_0 \| \mathcal{H}_0 \text{ true}\right)\\ &\quad = \mathbb{P}(Y \geq 20 \| Y \sim \text{Po}(16)) = 0.188.\end{aligned}$$

Example: MRSA (cont.)

- Bayesian approach:
  - However, additional information about the underlying risk may exist, such as previous years' rates or rates in similar hospitals. Suppose this additional information, on its own, suggests plausible values of θ of around 10 per $10{,}000$, with 95% of the support for θ lying between 5 and 17.
  - This can be expressed through the prior

$$\theta \sim \Gamma(a, b), \quad a = 10,\ b = 1.$$

  - The posterior distribution is now

$$f(\theta \mid y) \propto f(y \mid \theta) f(\theta) \propto \theta^y e^{-4\theta} \theta^{a-1} e^{-b\theta} \propto \theta^{y+a-1} e^{-\theta(4+b)}.$$
$$\Rightarrow \theta \mid Y = y \sim \Gamma(y + a, 4 + b).$$

Example: MRSA (cont.)

- Thus, the posterior is $\theta \mid Y = y \sim \Gamma(y + a, 4 + b)$.
- If we want a point estimate of θ, one may use Bayes's estimator

$$\widehat{\theta} = \mathbb{E}(\theta \mid Y = y) = \int \theta' f(\theta' \mid y)\, d\theta' = \frac{y + a}{4 + b} = \frac{20 + 10}{4 + 1} = 6.$$

- A credible or posterior probability interval can be found using the quantiles of the posterior distribution.
- A hypothesis test of $\mathcal{H}_0 : \theta = 4$ vs. $\mathcal{H}_1 : \theta > 4$ can be carried through by computing $\mathbb{P}(\theta \geq 4 \mid Y = y) = 0.978$, which indicates strong evidence against $\mathcal{H}_0$.

Next Week

- Using MCMC for Bayesian computation.
- Prior distributions.
- Mixing of MCMC samplers.
- This leads to HA2 which will cover Bayesian Inference using MCMC and the Bootstrap method.