Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

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

Lecture 12
MCMC for Bayesian computation
4 May 2017

Plan of today's lecture

Outline

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 $X_k \rightarrow X_{k+1}$ 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

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

Last time: the frequentist approach

- Data y is viewed as an observation of a random variable Y with distribution $\mathbb{P}_0$, which most often is assumed to be a member of an exponential family

$$\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}.$$

- Estimates $\widehat{\theta}(y)$ are realizations of random variables.
- The point estimate is often equipped with a confidence bound on level, say, 95%.
- Hypothesis testing is done by rejecting a hypothesis $\mathcal{H}_0$ if $\mathbb{P}(\text{data } y \| \mathcal{H}_0)$ is small.

Last time: the Bayesian approach

- The uncertainty concerning θ is modeled by viewing θ as a random variable, and inference is based completely on the posterior distribution $f(\theta \mid y)$.
- It is possible to incorporate prior information via the prior distribution $f(\theta)$.
- A 95% credible or posterior probability interval contains θ with a probability of 95% given the observations.
- Hypothesis tests are done by studying $\mathbb{P}(\mathcal{H}_0 \| \text{data } y)$.

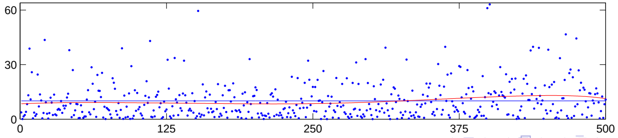
Plan of today's lecture

1. Last Time
2. MCMC for Bayesian computation
3. Prior distributions
4. Interlude: Mixing of MCMC samplers
5. HA2

Outline

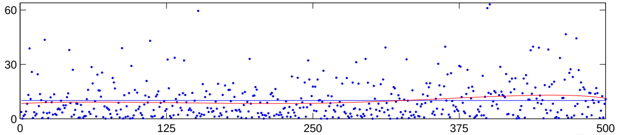
Example: change point detection

- We have measured the waiting times in a system and suspect that the expected waiting time changed during the monitoring period.
- The observations $(y_i)_{i=1}^{n}$ are assumed to follow exponential distributions with parameter θ_1 for $i \in \{1, \ldots, n_b\}$ and parameter θ_2 for $i \in \{n_b + 1, \ldots, n\}$.
- Further, we put a Gamma prior on θ_k, $\theta_k \sim \Gamma(a, b)$, with $a = 40$ and $b = 4$, and a uniform prior on n_b



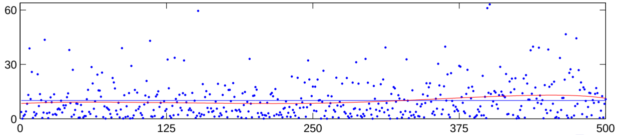
Example: change point detection

- We have measured the waiting times in a system and suspect that the expected waiting time changed during the monitoring period.
- The observations $(y_i)_{i=1}^n$ are assumed to follow exponential distributions with parameter θ_1 for $i \in \{1, \ldots, n_\text{b}\}$ and parameter θ_2 for $i \in \{n_\text{b} + 1, \ldots, n\}$.
- Further, we put a Gamma prior on θ_k, $\theta_k \sim \Gamma(a, b)$, with $a = 40$ and $b = 4$, and a uniform prior on n_b



Example: change point detection

- We have measured the waiting times in a system and suspect that the expected waiting time changed during the monitoring period.
- The observations $(y_i)_{i=1}^n$ are assumed to follow exponential distributions with parameter θ_1 for $i \in \{1, \ldots, n_\text{b}\}$ and parameter θ_2 for $i \in \{n_\text{b} + 1, \ldots, n\}$.
- Further, we put a Gamma prior on θ_k, $\theta_k \sim \Gamma(a, b)$, with $a = 40$ and $b = 4$, and a uniform prior on n_b



Example: change point detection (cont.)

- Thus, we have unknown parameters $(\theta_1, \theta_2, n_b)$ and data $Y = (y_1, \ldots, y_n)$. The posterior becomes

$$
\begin{aligned}
& f(n_b, \theta_1, \theta_2 \mid y_1, \ldots, y_n) \\
& \propto f(\theta_1) f(\theta_2) f(n_b) \prod_{i=1}^{n} f(y_i \mid n_b, \theta_1, \theta_2) \\
& = \theta_1^{n_b + a - 1} \exp\left(-\theta_1 \left(b + \sum_{i=1}^{n_b} y_i\right)\right) \\
& \qquad \times \theta_2^{n - n_b + a - 1} \exp\left(-\theta_2 \left(b + \sum_{i=n_b+1}^{n} y_i\right)\right).
\end{aligned}
$$

Example: change point detection (cont.)

- This posterior is complicated ...
- However, the conditional distributions of θ_1 and θ_2 are easily calculated according to

$$\theta_1 \mid n_b, y_1, \ldots, y_n \sim \Gamma\left(n_b + a, b + \sum_{i=1}^{n_b} y_i\right),$$

$$\theta_2 \mid n_b, y_1, \ldots, y_n \sim \Gamma\left(n - n_b + a, b + \sum_{i=n_b+1}^{n} y_i\right).$$

- The conditional distribution of n_b is however more complicated. Thus, we draw n_b by inserting an MH step in the Gibbs sampler, yielding a hybrid sampler.

Example: change point detection (cont.)

- This posterior is complicated ...
- However, the conditional distributions of θ_1 and θ_2 are easily calculated according to

$$\theta_1 \mid n_b, y_1, \ldots, y_n \sim \Gamma\left(n_b + a, b + \sum_{i=1}^{n_b} y_i\right),$$

$$\theta_2 \mid n_b, y_1, \ldots, y_n \sim \Gamma\left(n - n_b + a, b + \sum_{i=n_b+1}^{n} y_i\right).$$

- The conditional distribution of n_b is however more complicated. Thus, we draw n_b by inserting an MH step in the Gibbs sampler, yielding a hybrid sampler.

Example: change point detection (cont.)

- The MH step is as follows.
- Given n_b, we propose a candidate n_b^* uniformly on the integers $\{n_b - R, \ldots, n_b, \ldots, n_b + R\}$, for some R. This forms a symmetric proposal on $\{1, \ldots, n\}$.
- Thus, the acceptance probability for the MH step becomes

$$
\alpha(n_b, n_b^*) = 1 \wedge \frac{\theta_1^{n_b^*} \theta_2^{-n_b^*} \exp(-\theta_1 \sum_{i=1}^{n_b^*} y_i) \exp(-\theta_2 \sum_{i=n_b^*+1}^{n} y_i)}{\theta_1^{n_b} \theta_2^{-n_b} \exp(-\theta_1 \sum_{i=1}^{n_b} y_i) \exp(-\theta_2 \sum_{i=n_b+1}^{n} y_i)}.
$$

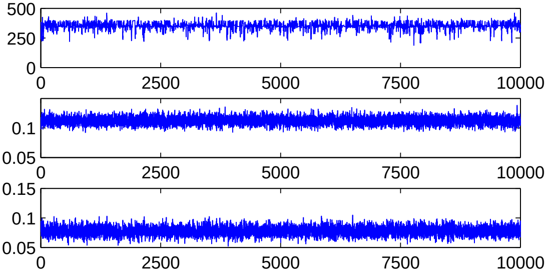
Example: change point detection (cont.)

- The MH step is as follows.
- Given n_b, we propose a candidate n_b^* uniformly on the integers $\{n_b - R, \ldots, n_b, \ldots, n_b + R\}$, for some R. This forms a symmetric proposal on $\{1, \ldots, n\}$.
- Thus, the acceptance probability for the MH step becomes

$$\alpha(n_b, n_b^*) = 1 \wedge \frac{\theta_1^{n_b^*} \theta_2^{-n_b^*} \exp(-\theta_1 \sum_{i=1}^{n_b^*} y_i) \exp(-\theta_2 \sum_{i=n_b^*+1}^{n} y_i)}{\theta_1^{n_b} \theta_2^{-n_b} \exp(-\theta_1 \sum_{i=1}^{n_b} y_i) \exp(-\theta_2 \sum_{i=n_b+1}^{n} y_i)}.$$

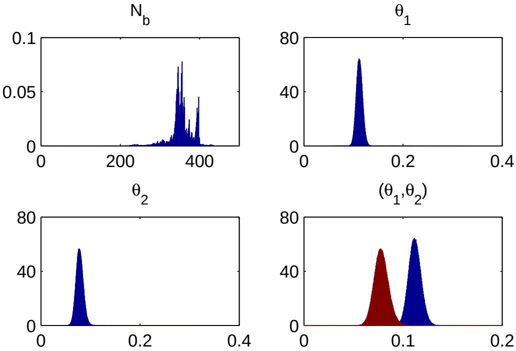
Example: change point detection (cont.)

- Running this Gibbs sampler with $R = 75$ gives an acceptance rate of 33%.



Example: change point detection (cont.)

- The resulting histograms of the parameters are as follows:



Outline

Selecting priors

- Recall that the posterior is computed via Bayes's formula

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

- In Bayesian modeling there is always an interplay between the prior and the data:
  - The posterior is drawn away from the data towards the prior. How far depends on the strength of the prior.
  - However, enough data will most likely overwhelm the prior.
- Two common prior-types are
  - conjugate priors.
  - improper (flat) priors.

Selecting priors

- Recall that the posterior is computed via Bayes's formula

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

- In Bayesian modeling there is always an interplay between the prior and the data:
  - The posterior is drawn away from the data towards the prior. How far depends on the strength of the prior.
  - However, enough data will most likely overwhelm the prior.
- Two common prior-types are
  - conjugate priors.
  - improper (flat) priors.

Conjugate priors

- Conjugate priors
  - are such that the prior and the posterior belong to the same distribution class for a given likelihood.
  - allow for straightforward theoretical calculations and Gibbs sampling.
  - are sometimes criticized since we select priors for ease of calculation.
  - may be hard to derive for complex models.

Conjugate priors

- Conjugate priors for θ for some common likelihoods. All parameters except θ are assumed fixed and known and data $(y_i)_{i=1}^n$ are assumed to be conditionally independent given θ.

Likelihood	*Prior*	*Posterior*
$\text{Bin}(n, \theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + y, \beta + n - y)$
$\text{Ge}(\theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + n, \beta + \sum_{i=1}^n y_i - n)$
$\text{NegBin}(n, \theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + n, \beta + y - n)$
$\Gamma(k, \theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + nk, \beta + \sum_{i=1}^n y_i)$
$\text{Po}(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + \sum_{i=1}^n y_i, \beta + n)$
$\text{N}(\mu, \theta^{-1})$	$\Gamma(\alpha, \beta)$	$\Gamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right)$
$\text{N}(\theta, \sigma^2)$	$\text{N}(m, s^2)$	$\text{N}\left(\frac{m/s^2 + n\bar{y}/\sigma^2}{1/s^2 + n/\sigma^2}, \frac{1}{1/s^2 + n/\sigma^2}\right)$

Improper priors

- Improper, or flat, priors are used when prior information is deficient.
- For instance, if $\theta \in \mathbb{R}$, $f(\theta) \propto 1$ is an improper prior since it is not integrable; however, we allow this as long as the posterior is a well-defined density.
- For instance, let y be an observation from $Y \sim \mathrm{N}(\theta, 1)$, where $\theta \in \mathbb{R}$. Since we do not have any prior information concerning θ we put $f(\theta) \propto 1$ for all $\theta \in \mathbb{R}$. After this we proceed, formally, like

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{\int f(y \mid \theta') f(\theta')\, d\theta'} = \frac{\mathrm{N}(y; \theta, 1) \cdot 1}{\int \mathrm{N}(y; \theta', 1) \cdot 1\, d\theta'}$$

$$\overset{\text{symm.}}{=} \frac{\mathrm{N}(\theta; y, 1) \cdot 1}{\int \mathrm{N}(\theta'; y, 1) \cdot 1\, d\theta'} = \mathrm{N}(\theta; y, 1).$$

Outline

Mixing of MCMC samplers

- We recall that the asymptotic variance of τ_N^{MCMC} is given by

$$\sigma^2 = r(0) + 2\sum_{\ell=1}^{\infty} r(\ell) \quad \text{with} \quad r(\ell) = \lim_{n\to\infty} \mathbb{C}(\phi(X_{n+\ell}), \phi(X_n)).$$

- Consequently, in order to obtain a low variance of τ_N^{MCMC}, the covariance function $r(\ell)$ should decrease rapidly with ℓ.
- For geometrically ergodic chains $r(\ell)$ tends to zero geometrically fast.
- The speed of which $r(\ell)$ tends to zero is typically described using the term “mixing”.
    - Strong mixing = fast forgetting = rapidly decreasing $r(\ell)$.
    - Bad mixing = slow forgetting = slowly decreasing $r(\ell)$.

Why is good mixing important?

- Bad choices of proposal distributions may lead to bad mixing, which causes problems for the MCMC algorithm in the sense that it may
  - need a very long time to converge.
  - exhibit high autocorrelation, implying high variance and the need of a large MC sample size to ensure good estimates.

Optimal mixing for the MH algorithm

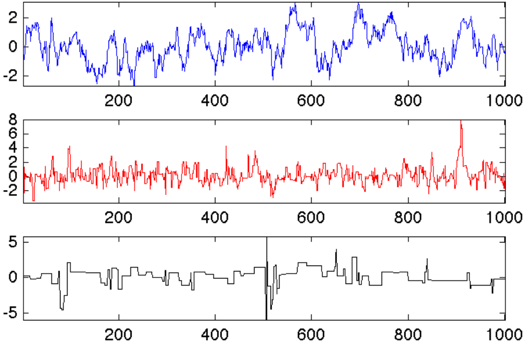
- When designing a random walk proposal, $X_k^* = X_k + \varepsilon$ with $\varepsilon \sim \mathrm{N}(\mathbf{0}, s\mathbf{\Sigma})$, two things effect the acceptance rate:
  1. how well $\mathbf{\Sigma}$ captures the dependence structure of the target distribution,
  2. how appropriate the scaling $s > 0$ is.
- One way to obtain a covariance matrix $\mathbf{\Sigma}$ that captures well the dependence structure of the target distribution $f(x)$ is to let

$$\mathbf{\Sigma}_{ij} = \frac{2.38}{d} \left(- \frac{\partial^2 \log f(x)}{\partial x_i \partial x_j} \bigg|_{x = x_{\text{mode}}} \right)^{-1}.$$

- Rule of thumb: a good acceptance rate is around 30% (23%–44%)!

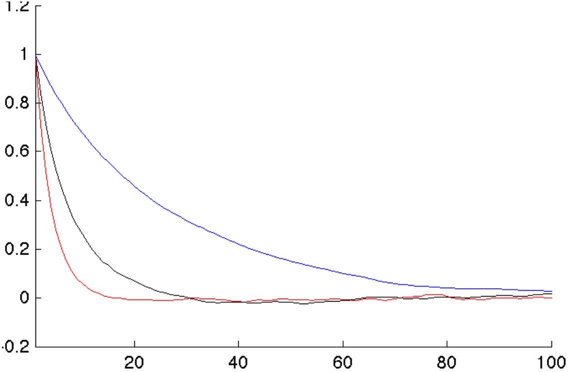
Mixing—Random walk proposal

- Using symmetric normal proposal with three different values for s (small, medium, large, respectively) yields typically trajectories of the following form:



Mixing—Random walk proposal

- Correlation function for the three chains:



Outline

HA2: MCMC and bootstrap

- HA2 comprises
  - one problem aiming at detecting change points in cole mine data using hybrid MCMC samplers and
  - one problem aiming at estimating the 100-year north Atlantic wave using parametric bootstrap (to be discussed).
- Submission:
  - A written report in PDF format.
  - An email containing all your m-files. With a file that runs your analysis.
  - Follow the same instructions as for HA1.
  - Deadline: Thursday 18 May, 13:00:00.