Monte Carlo and Empirical Methods for Stochastic Inference (MASM11/FMSN50)

Magnus Wiktorsson
Centre for Mathematical Sciences
Lund University, Sweden

Lecture 11
MCMC for Bayesian computation
February 26, 2019

Plan of today's lecture

We are here $\longrightarrow \bullet$

The Gibbs sampler

In the following,

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

The Gibbs sampler (cont.)

The Gibbs sampler then goes as follows.

Simulate a sequence of values (X_k), forming a Markov chain on X, with the following mechanism: Given X_k,

- 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

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

Variance of MCMC estimators

As mentioned, the MH and Gibbs samplers are geometrically ergodic, implying a LLN for the resulting estimators.

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 the MCMC chain at stationarity.

Theorem

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

where

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

Estimating asymptotic variance using blocking

Use $N = nK$ samples and write

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

where

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

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 CLT to hold at least approximately, implying that

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

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

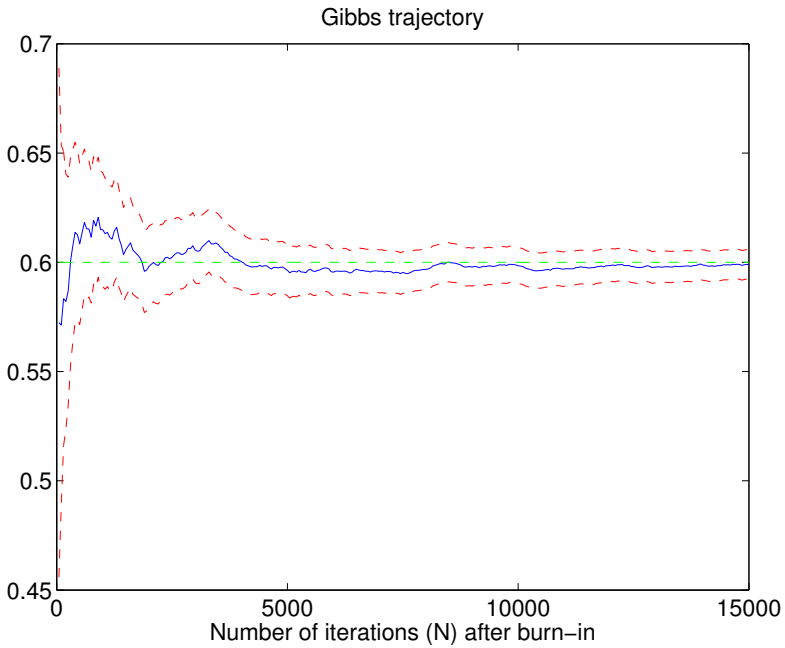
$$\tau = \mathbb{E}(Y)$$

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)

```
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);
```



We are here $\longrightarrow \bullet$

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 the conditional distribution of a block is known, update according to Gibbs.
- If there are blocks for which we cannot find the conditional distribution, just insert a MH step instead!

The resulting chain still satisfies global balance and is thus a valid MCMC sampler.

This will be of great use in HA3!

Hybrid chains (theoretical motivation)

Assume without loss of generality that we have two blocks and want to sample from $f(x_1, x_2)$. The hybrid MCMC goes like this, given $X = x = (x_1, x_2)$.

1. Draw $\tilde{X}_2 \sim q(\tilde{x}_2|x_2)$ where q is an MH kernel for $f(x_2|x_1)$.
2. Draw $\tilde{X}_1 \sim f(\tilde{x}_1|X_2 = \tilde{X}_2)$ i.e. standard Gibbs.

Global balance equation

$$\int\int q(\tilde{x}_2, \tilde{x}_1|x_1, x_2) f(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2$$

$$= \int\int f(\tilde{x}_1|\tilde{x}_2) q(\tilde{x}_2|x_2) f(x_2|x_1) f(x_1) \mathrm{d}x_1 \mathrm{d}x_2$$

$$= f(\tilde{x}_1|\tilde{x}_2) \int\int q(\tilde{x}_2|x_2) f(x_2|x_1) \mathrm{d}x_2 f(x_1) \mathrm{d}x_1$$

$$\overset{\text{Global Balance MH}}{=} f(\tilde{x}_1|\tilde{x}_2) \int f(\tilde{x}_2|x_1) f(x_1) \mathrm{d}x_1$$

$$\overset{\text{Law of total prob}}{=} f(\tilde{x}_1|\tilde{x}_2) f(\tilde{x}_2) = f(\tilde{x}_1, \tilde{x}_2),$$

which concludes the proof $\square$.

Part II: MC methods for statistical inference

We are here $\longrightarrow \bullet$

Overview

We will consider

- some literature,
- stochastic modeling,
- frequentist vs. Bayesian statistics, and
- an example—MCMC.

Alternative literature

For MCMC:

- *Markov Chain Monte Carlo in Practice*, Gilks, Richardson & Spiegelhalter, 1996.
- *Monte Carlo Statistical Methods*, Robert & Casella, 2005.

For Bootstrap (to be discussed next week):

- *Bootstrap Methods and Their Application*, Davison & Hinkley, 1997.
- *An Introduction to the Bootstrap*, Efron & Tibshirani, 1994.

Stochastic modeling: frequentist approach

The basic setup is the following.

- 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 fit 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

- 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

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

  where y is the given observed data.
- Often, a 95% confidence interval is calculated to cover the true value in 95% of the cases.

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,

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

- $f(y|\theta)$ 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 $f(\theta|y)$, i.e. the distribution of the parameters given the observed data.
- By Bayes's Formula:

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

Bayesian vs. frequentist statistics

- Bayesian inference is done using the posterior $f(\theta|y)$.
- Frequentist inference uses the likelihood $f(y|\theta)$.
- 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.

Bayesian vs. frequentist statistics: Example

An example:

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 $Y \sim \text{Po}(4\theta)$.

Bayesian vs. frequentist statistics: Example (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 which gives

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

Bayesian vs. frequentist: Example (cont.)

Bayesian approach:

- However, other information about the underlying risk may exist, such as the previous year's rates or rates in similar hospitals. Suppose this other 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

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

Bayesian vs. frequentist: Example (cont.)

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

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

- 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 $\mathbb{P}(\theta > 4|Y = y) = 0.978$, which indicates strong evidence against $\mathcal{H}_0$.

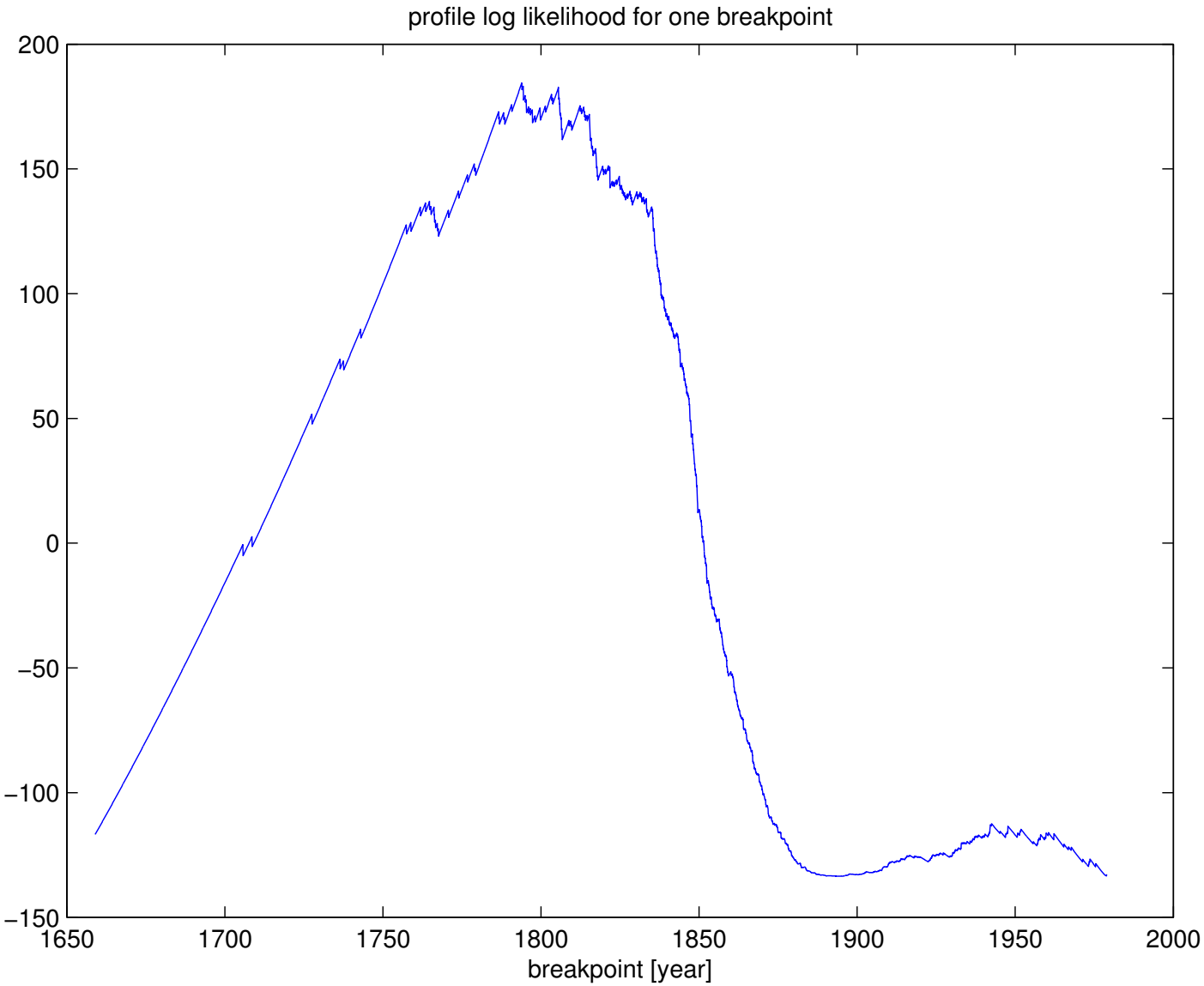
We are here $\longrightarrow \bullet$

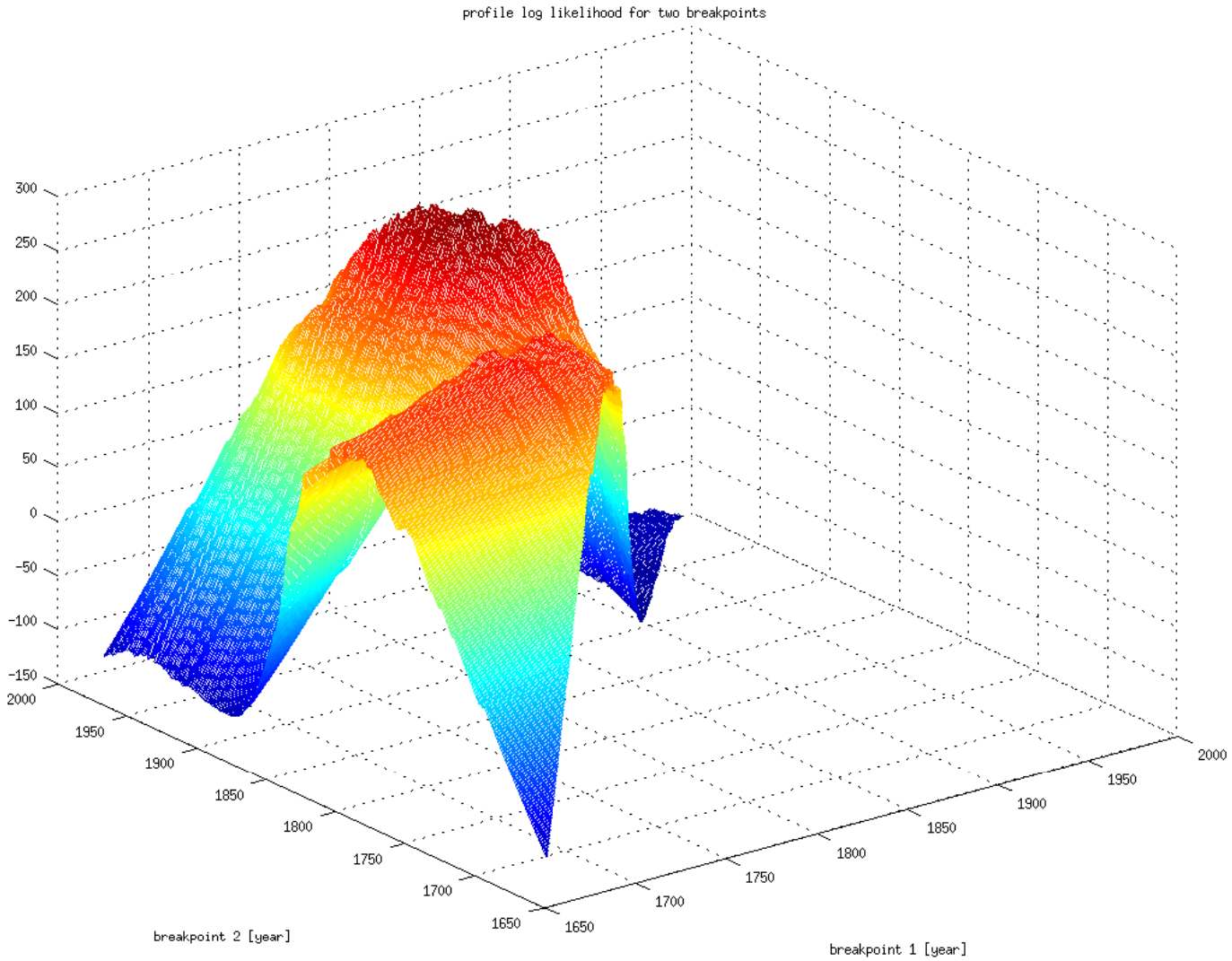
HA3: MCMC and bootstrap

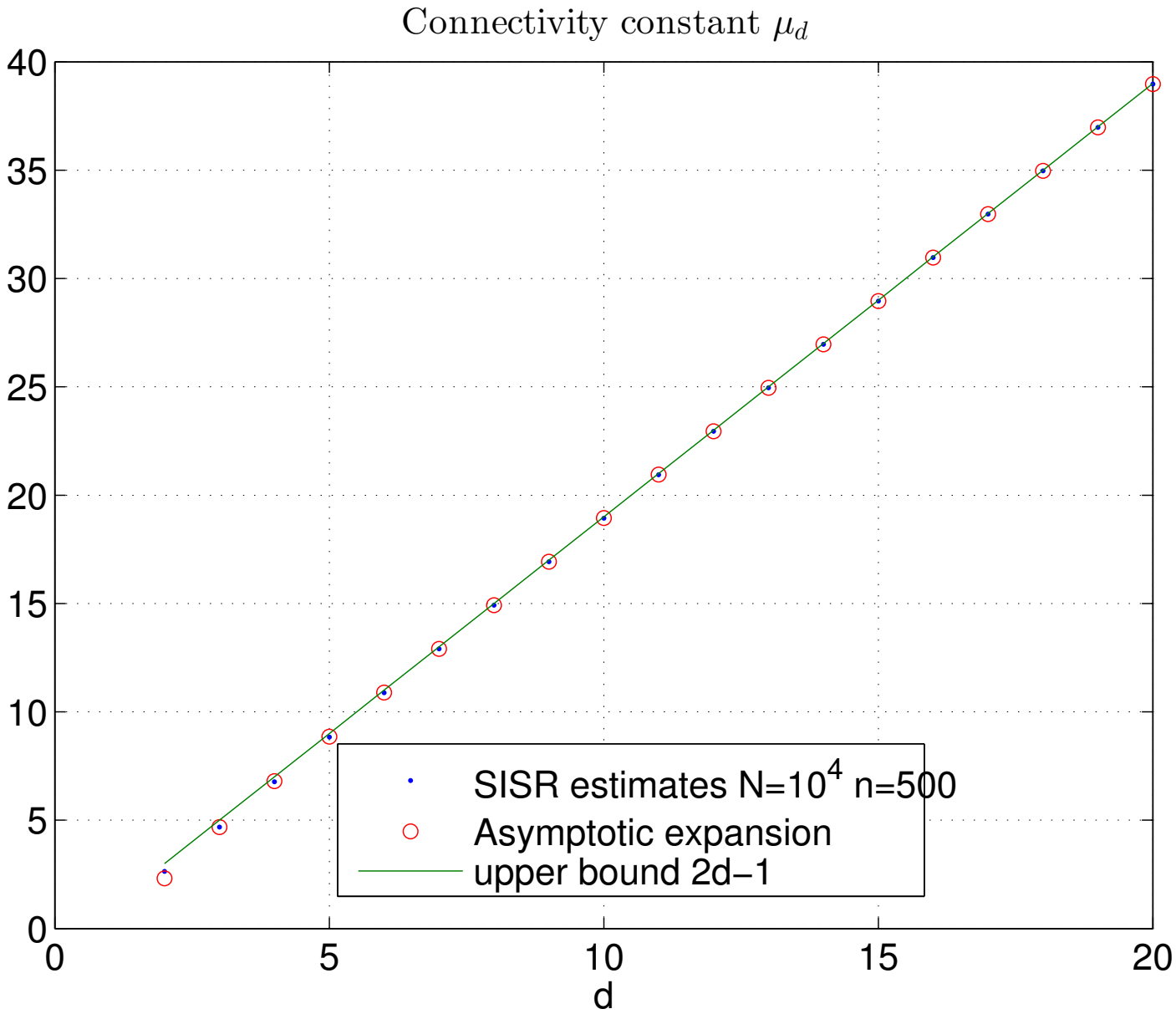
HA3 comprises

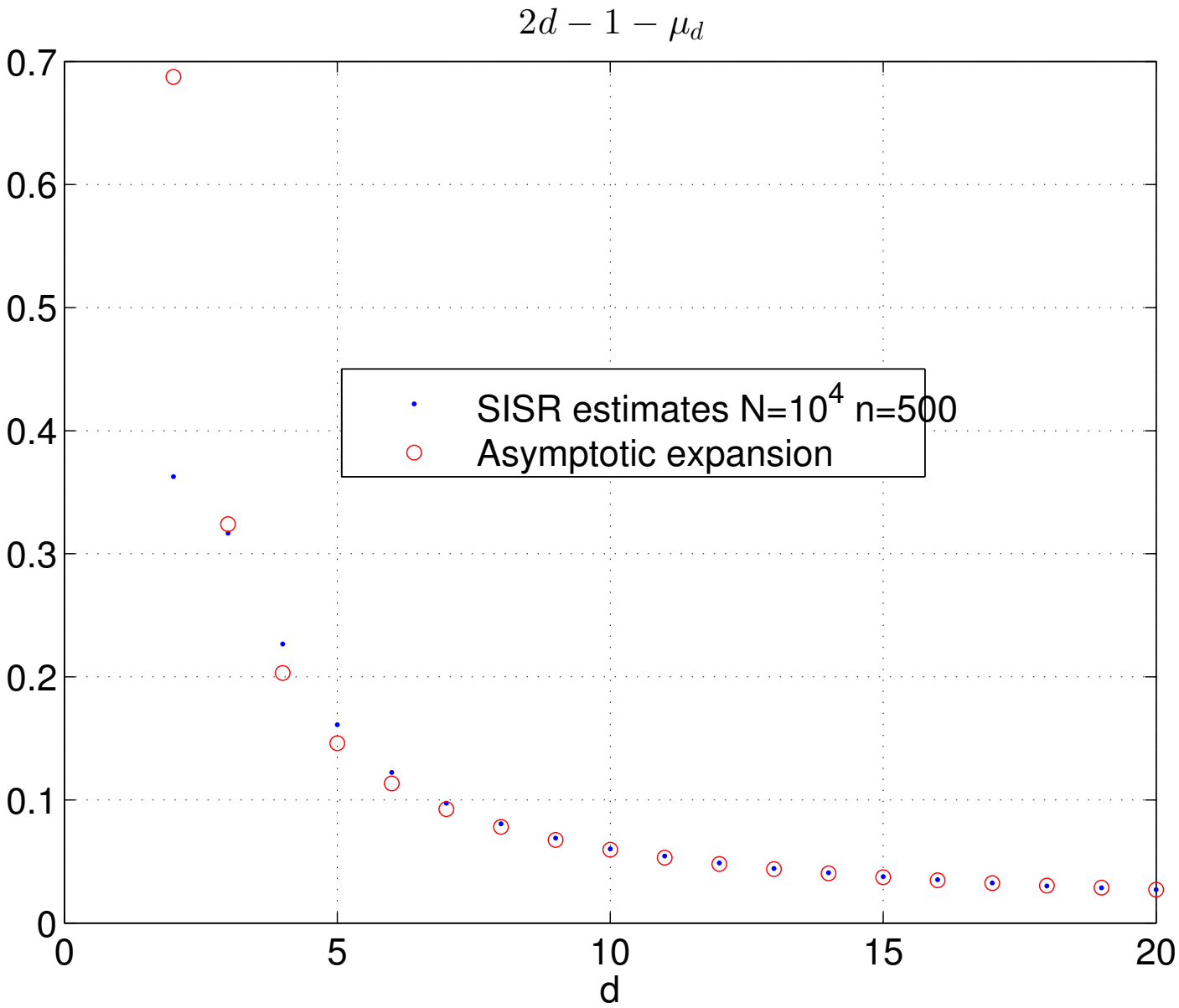
- one problem aiming at detecting change points in cole mine data using hybrid MCMC samplers and
- one problem aiming at estimating the 100-year north Atlantic wave using parametric bootstrap (discussed on Tuesday).

Deadline: **Tuesday 12 Mar, 13:00:00**.


profile log likelihood for one breakpoint
200
150
100
50
0
−50
−100
−150
1650
1700
1750
1800
1850
1900
1950
2000
breakpoint [year]


profile log likelihood for two breakpoints
300
250
200
150
100
50
0
-50
-100
-150
2000
1950
1900
1850
1800
1750
1700
1650
breakpoint 2 [year]
1650
1700
1750
1800
1850
1900
1950
2000
breakpoint 1 [year]


Connectivity constant μ_d
40
35
30
25
20
15
10
5
0
0
5
10
15
20
d
SISR estimates N=10^4 n=500
Asymptotic expansion
upper bound 2d−1


$2d - 1 - \mu_d$
SISR estimates N=10^4 n=500
Asymptotic expansion
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
d