

Computer Intensive Statistics

Paul Jenkins & Adam M. Johansen
p.jenkins@warwick.ac.uk a.m.johansen@warwick.ac.uk

9rd–13th July, 2018

Compiled: June 29, 2018: 15:03

Part 1

Introduction, Motivation & Basics

Motivation
oooo
oooooooooooo

Randomized Testing
oooooooo
oooo

Bootstrap Methods
oooo
oooo

What *is* Computer Intensive Statistics

Computer, *n.* A device or machine for performing or facilitating calculation.

Compare Middle French computeur person who makes calculations (1578).

Intensive, *adj.* Of very high degree or force, vehement.

French intensif, -ive (14–15th cent. in Hatzfeld & Darmesteter).

Statistics, *n.* The systematic collection and arrangement of numerical facts or data of any kind; (also) the branch of science or mathematics concerned with the analysis and interpretation of numerical data and appropriate ways of gathering such data.
In early use after French statistique and German Statistik.

Motivation
oooo
oooooooooooo

Randomized Testing
oooooooo
oooo

Bootstrap Methods
oooo
oooo

What Makes Statistics Computer Intensive?

Some *good* reasons for using computer-intensive methods:

Complexity Complex models cannot often be dealt with analytically.

Intractability Models which are not available analytically.

Laziness Computer time is cheap; human time isn't.

Scale Large data sets bring fresh challenges.

We won't address the *bad* reasons here. . .

Part 1— Section 1

Motivation

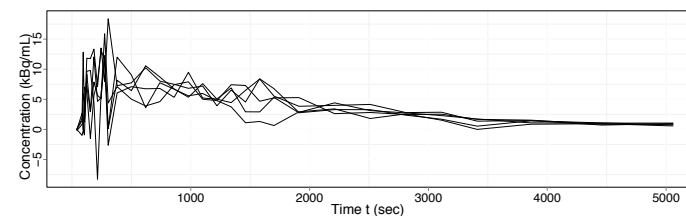
Motivation
●○○○
○○○○○○○○○○○○

Randomized Testing
○○○○○○○○
○○○○

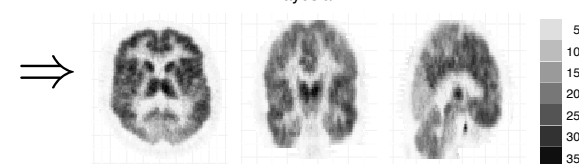
Bootstrap Methods
○○○○
○○○○○

Problems

Motivating Problem: Positron Emission Tomography I



Bayesian



6

Motivation
●○○○
○○○○○○○○○○○○

Randomized Testing
○○○○○○○○
○○○○

Bootstrap Methods
○○○○
○○○○○

Problems

Motivating Problem: Positron Emission Tomography II

Dynamic model:

$$\frac{dC_T}{dt}(s) = AC_T(s) + bC_P(s)$$

$$C_T(0) = \mathbf{0}$$

$$\bar{C}_T(s) = \mathbf{1}^T C_T(t),$$

with solution:

$$\bar{C}_T(t) = \int_0^t C_P(t-s) H_{TP}(s) ds \quad (1)$$

$$H_{TP}(t) = \sum_{i=1}^m \phi_i e^{-\theta_i t},$$

where the ϕ_i and θ_i are functions of A .

7

Motivation
●○○○
○○○○○○○○○○○○

Randomized Testing
○○○○○○○○
○○○○

Bootstrap Methods
○○○○
○○○○○

Problems

Motivating Problem: Positron Emission Tomography III

Interested in the *Volume of Distribution*:

$$V_D := \int_0^\infty H_{TP}(t) dt = \sum_{i=1}^m \frac{\phi_i}{\theta_i}.$$

But have noisy measurements of $\bar{C}_T(t_j)$ for $j = 1, \dots, n$:

$$y_j = \bar{C}_T(t_j; \phi_{1:m}, \theta_{1:m}) + \sqrt{\frac{\bar{C}_T(t_j; \phi_{1:m}, \theta_{1:m})}{t_j - t_{j-1}}} \varepsilon_j$$

$$\bar{C}_T(t_j; \phi_{1:m}, \theta_{1:m}) = \sum_{i=1}^m \phi_i \int_0^{t_j} C_P(t_j - s) e^{-\theta_i s} ds.$$

What can we say?

8

Motivating Problem: Hypothesis Testing

Testing Example: Chi-Squared Test

- $T = \sum_{i=1}^k \frac{(O_k - E_k)^2}{E_k}$
- Asymptotic argument:
- $T \stackrel{d}{\approx} \chi_{k-1}^2$ under regularity conditions.

What if we *don't* have many observations of every category?

What if we want to know whether the *medians* of two populations are *significantly different*?

What if we don't know the form of their distributions?

Motivating Problem: Confidence Intervals

Constructing confidence intervals requires knowledge of sampling distributions.

Confidence Interval: Medians

- $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} f_X$.
- $X_{[1]} \leq X_{[2]} \leq \dots X_{[n]}$ are the associated order statistics.
- $T = X_{[(n+1)/2]}$ is the sample median
- How can we construct a confidence interval for the median of f_X ?
- What if we don't even know the form of f_X ?

Motivating Problem: Bayesian Inference

Bayesian statistics

- Data $\mathbf{y}_1, \dots, \mathbf{y}_n$ and model $f(\mathbf{y}_i|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is some parameter of interest.
- \rightsquigarrow Likelihood $l(\mathbf{y}_1, \dots, \mathbf{y}_n|\boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{y}_i|\boldsymbol{\theta})$
- In the Bayesian framework $\boldsymbol{\theta}$ is a random variable with prior distribution $f^{\text{prior}}(\boldsymbol{\theta})$. After observing $\mathbf{y}_1, \dots, \mathbf{y}_n$ the posterior density of f is

$$\begin{aligned} f^{\text{post}}(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}|\mathbf{y}_1, \dots, \mathbf{y}_n) \\ &= \frac{f^{\text{prior}}(\boldsymbol{\theta})l(\mathbf{y}_1, \dots, \mathbf{y}_n|\boldsymbol{\theta})}{\int_{\Theta} f^{\text{prior}}(\boldsymbol{\vartheta})l(\mathbf{y}_1, \dots, \mathbf{y}_n|\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}} \end{aligned}$$

- Often intractable \rightsquigarrow use of an approximation.

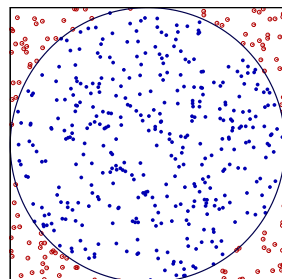
Simulation-based Methods

- Doing statistics backwards:

Representing the solution of a problem as a parameter of a hypothetical population, and using a random sequence of numbers to construct a sample of the population, from which statistical estimates of the parameter p values, confidence intervals, or other quantities of interest can be obtained.

Preliminary Example: Raindrop experiment for π

- Consider “uniform rain” on the square $[-1, 1] \times [-1, 1]$, i.e. the two coordinates $X, Y \stackrel{\text{iid}}{\sim} U[-1, 1]$.



- Probability that a rain drop falls in the circle is

$$\begin{aligned} \mathbb{P}(\text{drop within circle}) &= \frac{\text{area of the unit circle}}{\text{area of the square}} \\ &= \frac{\iint_{\{x^2+y^2 \leq 1\}} 1 \, dx dy}{\iint_{\{-1 \leq x, y \leq 1\}} 1 \, dx dy} = \frac{\pi}{2 \cdot 2} = \frac{\pi}{4}. \end{aligned}$$

13

Preliminary Example: Raindrop experiment for π

- Given π , we can compute $\mathbb{P}(\text{drop within circle}) = \frac{\pi}{4}$.
- Given n independent raindrops, the number of rain drops falling in the circle, Z_n is a binomial random variable:

$$Z_n \sim \text{Bin}\left(n, p = \frac{\pi}{4}\right).$$

- \leadsto we can estimate p with

$$\hat{p} = \frac{Z_n}{n}.$$

- and π by

$$\hat{\pi} = 4\hat{p} = 4 \cdot \frac{Z_n}{n}.$$

14

Preliminary Example: Raindrop experiment for π

- Result obtained for $n = 100$ raindrops: 77 points inside the circle.
- Resulting estimate of π is

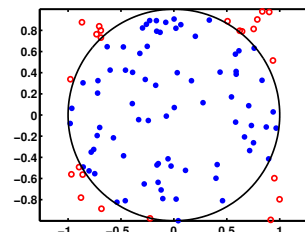
$$\hat{\pi} = \frac{4 \cdot Z_n}{n} = \frac{4 \cdot 77}{100} = 3.08,$$

(rather poor estimate)

- However: the *law of large numbers* **guarantees** that

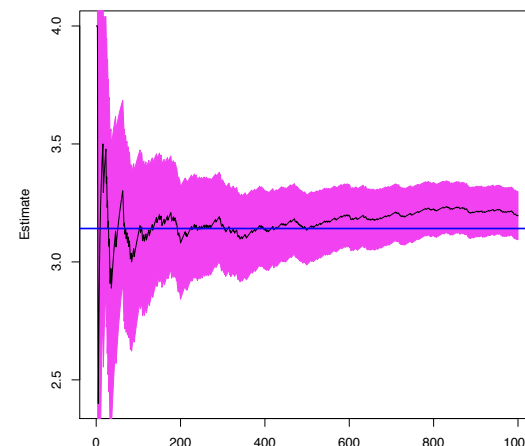
$$\hat{\pi}_n = \frac{4 \cdot Z_n}{n} \rightarrow \pi$$

almost surely for $n \rightarrow \infty$.



15

Preliminary Example: Raindrop experiment for π



16

Preliminary Example: Raindrop experiment for π

- How fast does $\hat{\pi}$ converge to π ?
Central limit theorem gives the answer.
- $(1 - 2\alpha)$ confidence interval for p ($\hat{p}_n = Z_n/n$):

$$\left[\hat{p}_n - z_{1-\alpha} \sqrt{\frac{\hat{p}_n(1 - \hat{p}_n)}{n}}, \hat{p}_n + z_{1-\alpha} \sqrt{\frac{\hat{p}_n(1 - \hat{p}_n)}{n}} \right]$$

- $(1 - 2\alpha)$ confidence interval for π ($\hat{\pi}_n = 4\hat{p}_n$):

$$\left[\hat{\pi}_n - z_{1-\alpha} \sqrt{\frac{\hat{\pi}_n(4 - \hat{\pi}_n)}{n}}, \hat{\pi}_n + z_{1-\alpha} \sqrt{\frac{\hat{\pi}_n(4 - \hat{\pi}_n)}{n}} \right]$$

- Width of the interval is $O(n^{-1/2})$, thus speed of convergence $O_{\mathbb{P}}(n^{-1/2})$.

17

Preliminary Example: Raindrop experiment for π

Recall the two core elements of this example:

- Writing the quantity of interest (here π) as an expectation:

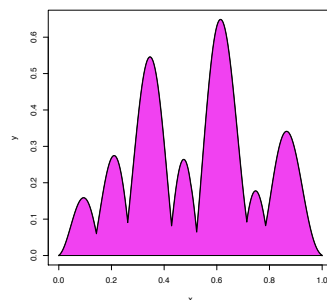
$$\pi = 4\mathbb{P}(\text{drop within circle}) = \mathbb{E}(4 \cdot \mathbb{I}_{\{\text{drop within circle}\}})$$

- Replaced this algebraic representation with a sample approximation.
 - SLLN guarantees that the sample approximation converges to the algebraic representation.
 - CLT gives information about the speed of convergence.

18

The Generalisation to Monte Carlo Integration

$$\begin{aligned} & \int_0^1 f(x) dx \\ &= \int_0^1 \int_0^{f(x)} 1 dt dx \\ &= \int \int_{\{(x,t): t \leq f(x)\}} 1 dt dx \\ &= \frac{\int \int_{\{(x,t): t \leq f(x)\}} 1 dt dx}{\int \int_{\{0 \leq x, t \leq 1\}} 1 dt dx} \end{aligned}$$



$$f : [0, 1] \rightarrow [0, 1]$$

19

Comparison of the speed of convergence

- Monte Carlo integration is $O_{\mathbb{P}}(n^{-1/2})$.
- Numerical integration of a *one-dimensional* function by Riemann sums is $O(n^{-1})$.
- Monte Carlo does not compare favourably for one-dimensional problems.
- However:
 - Monte Carlo estimates are often *unbiased*.
 - Order of convergence of Monte Carlo integration is *independent* of dimension.
 - Order of convergence of numerical integration techniques deteriorates with increasing dimension.

↪ Monte Carlo methods can be a good choice for high-dimensional integrals.

20

Views of Simulation-based Inference

Direct approximation of a quantity of interest.

- Careful construction of random experiment for particular task at hand.
- Justify with a dedicated argument in each case.

Approximation of integrals of interest.

- Represent quantity of interest as expectation wrt some f .
- Use sample average to approximate expectation.
- Appeal to SLLN and CLT.

Approximation of distributions of interest.

- Represent quantity of interest as a function of distribution f .
- Use empirical measure of sample to approximate f .
- Appeal to Glivenko-Cantelli theorem.

21

Theoretical Motivation of Sample Approximation

Theorem (Strong Law of Large Numbers)

Let $X_1, X_2, \dots \stackrel{iid}{\sim} f$, and let $\varphi : E \rightarrow \mathbb{R}$ with $\mathbb{E} [|\varphi(X_1)|] < \infty$ then:

$$\frac{1}{n} \sum_{i=1}^n \varphi(X_i) \xrightarrow{a.s.} \mathbb{E}_f [\varphi(X_1)].$$

Theorem (Central Limit Theorem)

Let $X_1, \dots \stackrel{iid}{\sim} f_X$ and let $\varphi : E \rightarrow \mathbb{R}^k$ with $\Sigma = \mathbb{V}ar [\varphi(X)] < \infty$, then as $n \rightarrow \infty$:

$$\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n \varphi(X_i) - \mathbb{E} [\varphi(X_1)] \right] \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Sigma).$$

22

Theoretical Motivation of Sample Approximation

Theorem (Glivenko-Cantelli)

Let $X_1, \dots \stackrel{iid}{\sim} f_X$ have cdf F_X .

Let

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, X_i]}(x)$$

then as $n \rightarrow \infty$

$$\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

23

Part 1— Section 2

Randomized Testing

Randomized Testing

- One simple example of computer intensive statistics.
- We'll revisit *how* we can implement these things later.
- Art of testing: find a set R_α such that

$$\mathbb{P}(T \in R_\alpha; H_0) = \alpha$$

and

$$\mathbb{P}(T \in R_\alpha; H_1) > \alpha$$

- What if we don't know f_T ?

25

Is a Die Fair?

- Given n rolls of a die, we want to establish whether it's fair.
- Canonical first-year example of a χ -squared test. . .
- Compute

$$T = \sum_{k=1}^K \frac{(O_k - E_k)^2}{E_k}$$

- $T \overset{\text{approx}}{\sim} \chi_{k-1}^2$ by asymptotic arguments.
- What if the asymptotics don't hold?

26

A Randomized Goodness of Fit Test

- Imagine we have 9 measured rolls (and can't easily obtain more):

Value	1	2	3	4	5	6
Count	0	1	0	2	2	4

- If the die is fair we *expect* 1.5 observations of each value.
- The test statistic is:

$$T = \frac{1.5^2 + 0.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 2.5^2}{1.5} = 7\frac{2}{3}$$

- The asymptotics *certainly* don't hold:

$$(O_k - E_k)^2 \in \{0.5^2, 1.5^2, 2.5^2, 3.5^2, 4.5^2, 5.5^2, 6.5^2, 7.5^2\}.$$

- But we can *simulate* from H_0 .

27

An R Implementation

Randomized Goodness of Fit Testing: Setup

```
p <- 1/6 * c(1,1,1,1,1,1)
n <- 9
r <- 10000
ob <- rmultinom(r,n,p)
ex <- n*p
T <- colSums((ob - ex)^2/ex)
```

How many elements in T are larger than the observed value?

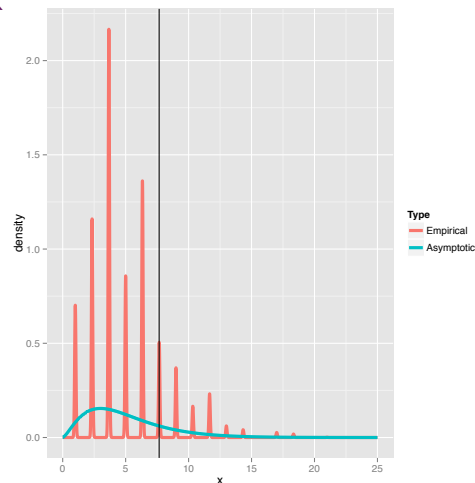
Randomized Goodness of Fit Testing: Comparison

```
t <- 7.66
m <- sum(T >= (t - 1E-9)) #T discrete
print(m/r)
```

28

A Startling (Anti)climax

Empirical p -value:
0.1848
Asymptotic p -value:
0.1860



29

Randomized Test in General

- Given a hypothesis, H_0 and an alternative, H_1 , and data \mathbf{x} which realises \mathbf{X} under H_0 :
 - Obtain a realisation \mathbf{u} of \mathbf{U} ($\mathbf{U}|\mathbf{X} \sim f_{\mathbf{U}|\mathbf{X}}$ from some known distribution).
 - Compute R_α such that $\mathbb{P}((\mathbf{X}, \mathbf{U}) \in R_\alpha; H_0) = \alpha$.
 - Reject H_0 if $(\mathbf{x}, \mathbf{u}) \in R_\alpha$.

Goodness of Fit Test in General Form

- Let $f_{\mathbf{U}|\mathbf{X}}(\mathbf{u}|\mathbf{x}) = \prod_{i=1}^r f_T(u_i; H_0)$.
By sampling $Z_i \stackrel{\text{iid}}{\sim} f_X(\cdot; H_0)$ and setting $U_i = g(Z_i)$.
- Let $R_\alpha = \{(\mathbf{x}, \mathbf{u}) : g(\mathbf{x}) > u_{[r(1-\alpha)]}\}$.
Where g is such that $T = g(\mathbf{X})$; $u_{[i]}$ is the i^{th} order statistic.

30

Are Those Medians Different (Part I)?

- Consider testing for different medians:

$$H_0 : X_1, \dots, X_{n_X} \stackrel{\text{iid}}{\sim} f_X(\cdot; m) \quad Y_1, \dots, Y_{n_Y} \stackrel{\text{iid}}{\sim} f_Y(\cdot; m)$$

$$H_1 : X_1, \dots, X_{n_X} \stackrel{\text{iid}}{\sim} f_X(\cdot; m) \quad Y_1, \dots, Y_{n_Y} \stackrel{\text{iid}}{\sim} f_Y(\cdot; m')$$

- Here, let's consider the two-sided version: $m' \neq m$.
- And we'll assume that we know the form of the two distributions:

$$f_X(x; m) = f_Y(x; m) = \frac{1}{2} \exp(-|x - m|)$$

- Letting $\tilde{X} = X_{[(n_X+1)/2]}$ and $\tilde{Y} = Y_{[(n_Y+1)/2]}$:

$$\begin{aligned} \tilde{X} - \tilde{Y} &= (\tilde{X} - m) - (\tilde{Y} - m) \\ &= (X - m)_{[(n_X+1)/2]} - (Y - m)_{[(n_Y+1)/2]} \end{aligned}$$

- So the distribution of $\tilde{X} - \tilde{Y}$ is independent of $m|H_0$.

31

- A Randomized test:
 - Let $T = \tilde{X} - \tilde{Y}$.
 - Draw $i = 1 : r$ copies of \mathbf{X} and \mathbf{Y} with $m = 0$:

$$X_{1, \dots, n_X}^{i,j} \stackrel{\text{iid}}{\sim} f_X(\cdot; 0),$$

$$Y_{1, \dots, n_Y}^{i,j} \stackrel{\text{iid}}{\sim} f_Y(\cdot; 0).$$

- Compute the difference between their medians:

$$i = 1, \dots, r : \quad T_i' = X_{[(n_X+1)/2]}^{i,i} - Y_{[(n_Y+1)/2]}^{i,i}$$

- Let $p = (1 + |\{i : T_i' \geq T\}|)/(r + 1)$.
- Reject H_0 if $p < \alpha$.

But surely this is cheating: what if we *don't* know so much?

32

Permutation Tests

- Consider the hypotheses:

$$H_0: X_1, \dots, X_{n_X} \stackrel{\text{iid}}{\sim} f_X(\cdot) \quad Y_1, \dots, Y_{n_Y} \stackrel{\text{iid}}{\sim} f_Y(\cdot)$$

$$F_X^{-1}(0.5) = F_Y^{-1}(0.5)$$

$$H_1: X_1, \dots, X_{n_X} \stackrel{\text{iid}}{\sim} f_X(\cdot) \quad Y_1, \dots, Y_{n_Y} \stackrel{\text{iid}}{\sim} f_Y(\cdot)$$

$$F_X^{-1}(0.5) \neq F_Y^{-1}(0.5)$$

where f_X and f_Y are unknown.

- Here, F_X^{-1} and F_Y^{-1} are assumed to exist.
- Sample medians are a natural test statistics, but:
 - We don't know their distribution under H_0 .
 - And can't sample from that distribution.
- What can we do?

33

The Permutation Approach

- If $\exists m$ s.t. $\mathbb{P}(X_i \leq m) = \mathbb{P}(Y_i \leq m) = 0.5$,
- and F_X^{-1} and F_Y^{-1} both exist,
- then $F_X^{-1}(0.5) = F_Y^{-1}(0.5) = m$
- and $F_X(m) = F_Y(m) = 0.5$
- so $\alpha F_X(m) + (1 - \alpha) F_Y(m) = 0.5$.
- In fact, under H_0 , the distribution of \tilde{X} and \tilde{Y} should be invariant under label permutations.

34

- Let $\mathbf{Z} = (X_1, \dots, X_{n_X}, Y_1, \dots, Y_{n_Y})$ be an $n = n_X + n_Y$ vector.
 - Now let
- $$T(\mathbf{Z}) = \text{median}(Z_1, \dots, Z_{n_X}) - \text{median}(Z_{n_X+1}, \dots, Z_n)$$
- And let $\pi \in \mathcal{P} \subset \{1, \dots, n\}^n$ denote a permutation, writing:

$$\pi \mathbf{Z} := (Z_{\pi_1}, Z_{\pi_2}, \dots, Z_{\pi_n})$$

- Now, under H_0 :
- $$\forall \pi \in \mathcal{P}: \quad T(\pi \mathbf{Z}) \stackrel{\mathcal{D}}{=} T(\mathbf{Z})$$
- So if $T(\mathbf{Z}) > T(\pi \mathbf{Z})$ for $100(1 - \alpha)\%$ of π we can reject H_0 .
 - We *just* need to compute $T(\pi \mathbf{Z})$ for every $\pi \in \mathcal{P}$...

35

A Randomized Permutation Test

- We can sample elements uniformly from \mathcal{P} :
 - Sample $\pi_1 \sim \mathcal{U}(1, \dots, n)$.
 - Sample $\pi_2 \sim \mathcal{U}(\{1, \dots, n\} \setminus \{\pi_1\})$.
 - \vdots
 - Sample $\pi_n \sim \mathcal{U}(\{1, \dots, n\} \setminus \{\pi_1, \dots, \pi_{n-1}\})$.
- We can do this many times to approximate the law of $T(\pi \mathbf{Z})$ when $\pi \sim \mathcal{U}(\mathcal{P})$:
 - Sample $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_k \stackrel{\text{iid}}{\sim} \mathcal{U}(\mathcal{P})$.
 - Compute $T_1 = T(\boldsymbol{\pi}_1 \mathbf{Z}), \dots, T_k = T(\boldsymbol{\pi}_k \mathbf{Z})$.
 - Use the empirical distribution of (T_1, \dots, T_k) to approximate the law of $T(\pi \mathbf{Z})$.
- This provides a general strategy for nonparametric testing.

36

Bootstrap Methods

Bootstrap Methods

- Randomized tests: use empirical distribution of T .
- Permutation tests: use *resampling*-based empirical distribution of T .
- Bootstrap methods: use *resampling*-based empirical distribution of $\hat{\theta}$ to characterise the sampling distribution of $\hat{\theta}$.

The Bootstrap Ansatz

If $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F_X$ and n is large then " $\hat{F}_X^n \approx F$ "
 \Rightarrow sampling from \hat{F}_X^n is "close" to sampling from F
 \Rightarrow samples from \hat{F}_X^n might be suitable for approximating F !

The Basis of the Bootstrap

- Given a simple random sample X_1, \dots, X_n
- Repeat the following for $b = 1 : B$:
 - Sample n times from $\hat{F}_X^n(x)$ i.e. sample n times uniformly with replacement from X_1, \dots, X_n to obtain $\hat{X}_1^b, \dots, \hat{X}_n^b$.
- Given a function $g : E^n \rightarrow \mathbb{R}$ approximate the distribution of g under F using the sample $g(\hat{X}_1^1, \dots, \hat{X}_n^1), \dots, g(\hat{X}_1^B, \dots, \hat{X}_n^B)$.
- Glivenko-Cantelli (and extensions) tells us that $\hat{F}_X^n(x) \xrightarrow{a.s.} F_X(x)$.

NB Regularity conditions must hold in order for this to work.

Approximating the Sampling Distribution of the Median

- Given X_1, \dots, X_n a simple random sample:
- Compute $T = \text{median}(X_1, \dots, X_n)$.
- For $b = 1 : B$
 - Sample n times with replacement from X_1, \dots, X_n to obtain $\hat{X}_1^b, \dots, \hat{X}_n^b$.
 - Compute $\hat{T}^b = \text{median}(\hat{X}_1^b, \dots, \hat{X}_n^b)$.
- Treat the empirical distribution of $\hat{T}^1, \dots, \hat{T}^B$ as a proxy for the sampling distribution of T .

Bootstrap Bias Correction

- Given x_1, \dots, x_n and,
- estimator $T : E^n \rightarrow \mathbb{R}$ of θ
- computer $t = T(x_1, \dots, x_n)$.
- For $b = 1 : B$
 - Sample n times with replacement from X_1, \dots, X_n to obtain $\hat{X}_1^b, \dots, \hat{X}_n^b$.
 - Compute $\hat{T}^b = T(\hat{X}_1^b, \dots, \hat{X}_n^b)$.
- Treat the empirical distribution of $\hat{T}^1 - t, \dots, \hat{T}^B - t$ as a proxy for the sampling distribution of $T(X_1, \dots, X_n) - \theta$.
- Obtain *bias-corrected* estimate:

$$t - \frac{1}{B} \sum_{b=1}^B (\hat{T}^b - t) = 2t - \frac{1}{B} \sum_{b=1}^B \hat{T}^b.$$

41

Naïve Bootstrap Confidence Intervals 1: The Asymptotic Approach

- For some T we might expect T to have an asymptotically normal distribution.
- So, estimate it's variance:

$$\hat{\sigma}_T^2 = \frac{1}{B-1} \sum_{b=1}^B \left(\hat{T}^b - \frac{1}{B} \sum_{b=1}^B \hat{T}^b \right)^2$$

- And use the normal confidence interval:

$$[T - z_{\alpha/2} \hat{\sigma}_T, T + z_{\alpha/2} \hat{\sigma}_T]$$

with approximate coverage α .

- Depends on asymptotic normality.
- Further approximation for finite samples.

42

Naïve Bootstrap Confidence Intervals 2: Bootstrap Percentile Confidence Intervals

- We could use the bootstrap distribution of T directly:

$$[\hat{T}^{[B(\alpha/2)]}, \hat{T}^{[B(1-\alpha/2)]}]$$

- These are known as *bootstrap percentile confidence intervals*.
- Depend on the *bootstrap* approximation; no additional approximations.

43

Bootstrap “pivotal” Confidence Intervals

- Using bootstrap approximations of (approximate) pivots can be more elegant.
- Assume that T is an estimator of some real population parameter, θ .
- Define $R = T - \theta$.
- Let F_R denote the cdf of R , then:

$$\begin{aligned} \mathbb{P}(L \leq \theta \leq U) &= \mathbb{P}(L - T \leq \theta - T \leq U - T) \\ &= \mathbb{P}(T - U \leq R \leq T - L) \\ &= F_R(T - L) - F_R(T - U). \end{aligned}$$

Suggests using:

$$[T - F_R^{-1}(1 - \alpha/2), T - F_R^{-1}(\alpha/2)]$$

- We can't use this interval directly because we don't know F_R and we certainly don't know F_R^{-1} .

44

Bootstrap “pivotal” Confidence Intervals

- We can invoke the bootstrap idea again:
- Compute $T = g(X_1, \dots, X_n)$.
- For $b = 1 : B$
 - Sample n times with replacement from X_1, \dots, X_n to obtain $\hat{X}_1^b, \dots, \hat{X}_n^b$.
 - Compute $\hat{T}^b = g(\hat{X}_1^b, \dots, \hat{X}_n^b)$.
- Claim that “ $\hat{T}^1, \dots, \hat{T}^B$ are to T as T is to θ ”.
- Set $\hat{R}^b = \hat{T}^b - T$.
- Use the empirical distribution, \hat{F}_R , of $\hat{R}^1, \dots, \hat{R}^B$ instead of F_R :

$$[T - \hat{F}_R^{-1}(1 - \alpha/2), T - \hat{F}_R^{-1}(\alpha/2)]$$

45

Summary of Part 1

- Motivation: Bayesian inference, Fisherian inference, ...
- Towards simulation-based inference (see later).
- Randomized Tests
- Permutation Tests
- Bootstrap Characterisation of Estimators.
- Bootstrap Confidence Intervals.
- Young, G. A. (1994) Bootstrap: More than a stab in the dark? *Statistical Science*, 9, 382–395.
- Davison, A. C., Hinkley, D. V. and Young, G. A. (2003) Recent developments in bootstrap methodology. *Statistical Science*, 18, 141–157.

46

Part 2

Simulation and the Monte Carlo Method

Simulation

- We've seen *motivation* of simulation for inference.
- We've seen *examples* of simulation-based methods.
- We *need* methods for addressing broad classes of problems.
- We *need* methods for obtaining the necessary samples.

48

The Monte Carlo Method

Monte Carlo Method

- A generic scheme for approximating expectations.
- To approximate $I = \mathbb{E}_f [\varphi(X)]$,
- Draw $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f$,
- Use $\hat{I}_{\text{mc}} = \frac{1}{n} \sum_{i=1}^n \varphi(X_i)$.
- Convergence follows from SLLN, CLT, ...

50

Recall: The Three Views of the Monte Carlo Method

Direct Approximation Design an experiment such that:

$$\varphi(X) \sim f_{\varphi(X)}$$

constructed such that it has the expectation of interest.

Integral Approximation We're interested in

$$\mathbb{E}_f [\varphi(X)]$$

and know how to approximate such.

Distributional Approximation We're interested in

$$\mathbb{E}_f [\varphi(X)]$$

so obtain an approximation of f which we can compute expectations with respect to.

51

Contrasting Views of Monte Carlo

- Usual explanation of the Monte Carlo Method, with $X_1, \dots \stackrel{\text{iid}}{\sim} f$ approximate the integral:

$$\frac{1}{n} \sum_{i=1}^n \varphi(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}_f [\varphi(X)]$$

- Another perspective, approximate the distribution:
 - let $\hat{f}^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$
 - if $\hat{f}^n \Rightarrow f$
 - then we automatically have that

$$\mathbb{E}_{\hat{f}^n} [\varphi(X)] \rightarrow \mathbb{E}_f [\varphi(X)]$$

for every continuous bounded φ .

52

Part 2— Section 5

PRNGs

Problem: (how) can computers produce random numbers?

von Neumann's perspective

Any one who considers arithmetical methods of reproducing random digits is, of course, in a state of sin. . . there is no such thing as a random number — there are only methods of producing random numbers, and a strict arithmetic procedure is of course not such a method.

As in so many other areas, von Neumann was completely correct.

Three Resolutions of this Philosophical Paradox

- 1 Use Exogeneous Randomness (TRNGs)
See www.random.org or http://en.wikipedia.org/wiki/Hardware_random_number_generator.
- 2 Pseudorandom Number Generators (PRNGs; cf. *Statistical Computing* module)
Sacrifice randomness whilst mimicking its *relevant statistical properties*.
- 3 Quasirandom Number Sequences (QRNSs)
Sacrifice randomness in exchange for *minimising discrepancy*.

All have advantages and disadvantages; we'll focus on PRNGs.

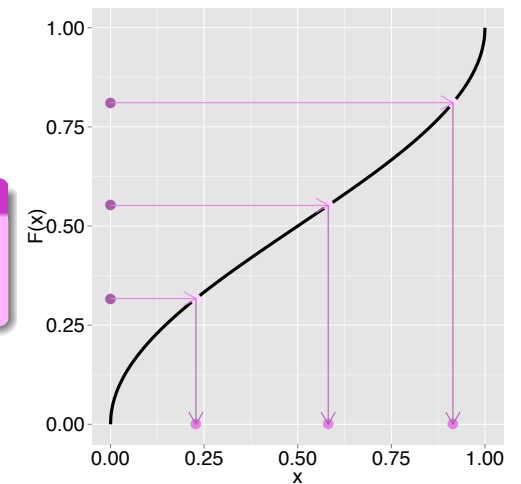
Part 2— Section 6

Sampling From Distributions

Transformation Methods

- Assume we have a *good* PRNG.
- How can we obtain (pseudo)samples from other distributions?
- General framework:
 - Treat output of PRNG as a stream of iid $U[0, 1]$ RVs.
 - Use laws of probability to transform these to obtain RVs with other distributions.
 - Treat transformed PRNG output as RVs of the target distribution.
- But, how?

Inversion Sampling



Inversion Sampling

The Inversion method

Let $U \sim \text{U}[0, 1]$ and F be an invertible CDF.
Then $F^{-1}(U)$ has the CDF F .

Inversion Sampling: A simple algorithm for drawing $X \sim F$

- 1 Draw $U \sim \text{U}[0, 1]$.
- 2 Set $X = F^{-1}(U)$.

Example: Exponential distribution

The exponential distribution with rate $\lambda > 0$ has the CDF ($x \geq 0$)

$$\begin{aligned} F_\lambda(x) &= 1 - \exp(-\lambda x) \\ F_\lambda^{-1}(u) &= -\log(1 - u)/\lambda. \end{aligned}$$

So we have a simple algorithm for drawing $X \sim \text{Exp}(\lambda)$:

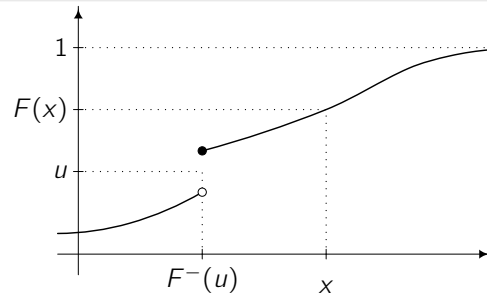
- 1 Draw $U \sim U[0, 1]$.
- 2 Set $X = -\frac{\log(1 - U)}{\lambda}$.

Actually, setting $X = -\frac{\log(U)}{\lambda}$ makes more sense.

The Generalised Inverse of the CDF

Generalised inverse of the CDF

$$F^{-}(u) := \inf\{x : F(x) \geq u\}$$



Replacing F^{-1} with F^{-} yields a generally-applicable inversion sampling algorithm — key is $F^{-}(u) \leq x \Leftrightarrow u \leq F(x)$.

61

Box-Muller: Fast Normally-Distributed Random Variables

- Consider (X_1, X_2) their polar representation (R, θ) :

$$X_1 = R \cdot \cos(\theta), \quad X_2 = R \cdot \sin(\theta)$$

- The following equivalence holds (with θ, R independent):

$$X_1, X_2 \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1) \iff \theta \sim \mathcal{U}[0, 2\pi] \text{ and } R^2 \sim \text{Expo}(1/2)$$

- Given $U_1, U_2 \stackrel{\text{iid}}{\sim} \mathcal{U}[0, 1]$ set

$$R = \sqrt{-2 \log(U_1)}, \quad \theta = 2\pi U_2,$$

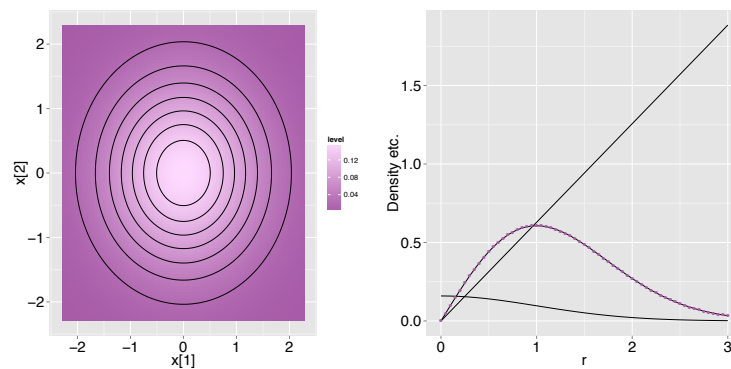
- By substitution

$$X_1 = \sqrt{-2 \log(U_1)} \cdot \cos(2\pi U_2)$$

$$X_2 = \sqrt{-2 \log(U_1)} \cdot \sin(2\pi U_2)$$

62

The Geometry of Box-Muller Transformation



63

Box-Muller: Algorithm

Box-Muller method

- Draw

$$U_1, U_2 \stackrel{\text{iid}}{\sim} \mathcal{U}[0, 1].$$

- Set

$$X_1 = \sqrt{-2 \log(U_1)} \cdot \cos(2\pi U_2),$$

$$X_2 = \sqrt{-2 \log(U_1)} \cdot \sin(2\pi U_2).$$

- Output $X_1, X_2 \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

64

The Limitations of Simple Transformations. . .

- When F^- is available and cheap to evaluate, inversion sampling is very efficient. But:
 - We often don't have access to F^- ;
 - if we do F^- may be difficult/impossible to obtain.
 - The multivariate case can be even harder.
- Clever custom transformations:
 - are costly to develop
 - require considerable ingenuity
 - are completely infeasible in complicated scenarios
- We need alternatives.

65

The Fundamental Theorem of simulation

Fundamental Theorem of Simulation

Sampling from a density f is equivalent to sampling uniformly from the area between f and the ordinal axes and discarding the "vertical" component.

- Follows from the identity

$$f(x) = \int_0^{f(x)} 1 \, du = \int_0^\infty \underbrace{1_{0 < u < f(x)}}_{=f(x,u)} \, du.$$

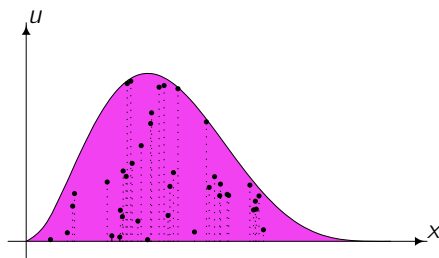
- i.e. $f(x)$ can be interpreted as the marginal density of a uniform distribution on the area under the density $f(x)$:

$$\{(x, u) : 0 \leq u \leq f(x)\}.$$

66

First element of rejection sampling

- We can sample from f by sampling from the area under the density.



- If $(X, U) \sim U(\{(x, u) : 0 \leq u \leq f(x)\})$ then $X \sim f$.

67

Second Element of Rejection Sampling

- Generally $\mathcal{G} = \{(x, u) : 0 \leq u \leq f(x)\}$ is complicated: we can't sample uniformly from it — at least not directly.
- Idea: Instead:
 - Sample from some $\mathcal{A} \supset \mathcal{G}$.
 - Keep only those points which lie within \mathcal{G} .
 - *Reject* the rest.

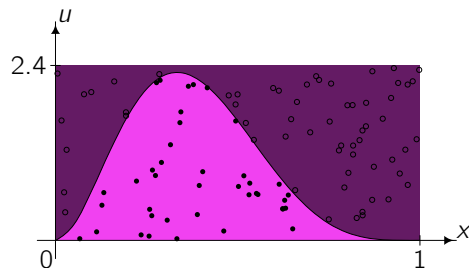
68

Example: Sampling from a Beta(3, 5) distribution (1)

- 1 Draw (X, U) from the dark rectangle, i.e.:

$$X \sim U(0, 1) \quad U \sim U(0, 2.4) \quad X \perp U.$$

- 2 Accept X as a sample from f if (X, U) lies under the density.

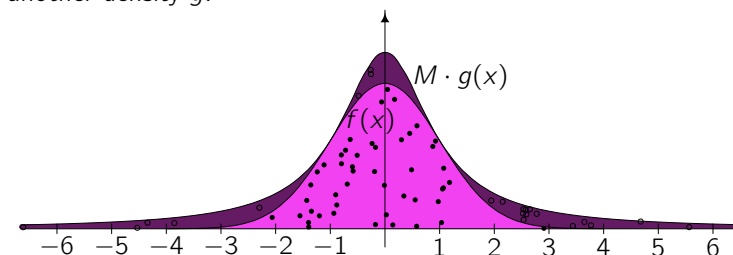


Step 2 is equivalent to: Accept X if $U < f(X)$,
i.e. accept X with probability $\mathbb{P}(U < f(X) | X = x) = f(X)/2.4$.

69

Example: Sampling from a Beta(3, 5) distribution (2)

- Algorithm:
 - 1 Draw $X \sim U(0, 1)$.
 - 2 Accept X as a sample from Beta(3, 5) w.p. $f(X)/2.4$.
- Not every density can be bounded by a box.
- Natural generalisation: replace M times $U[0, 1]$ with M times another density g .



70

A General Algorithm

Algorithm: Rejection sampling

Given two densities f, g with $f(x) < M \cdot g(x)$ for all x , we can generate a sample from f by

1. Draw $X \sim g$.
2. Accept X as a sample from f with probability

$$\frac{f(X)}{M \cdot g(X)},$$

otherwise go back to step 1.

For $f(x) < M \cdot g(x)$ to hold f cannot have heavier tails than g .

71

A Useful Trick

Avoiding Unknown Constants

If we know only $\tilde{f}(x)$ and $\tilde{g}(x)$, where $f(x) = C \cdot \tilde{f}(x)$, and $g(x) = D \cdot \tilde{g}(x)$ we can carry out rejection sampling using acceptance probability

$$\frac{\tilde{f}(X)}{M \cdot \tilde{g}(X)}$$

provided $\tilde{f}(x) < M \cdot \tilde{g}(x)$ for all x .

Can be useful in Bayesian statistics:

$$f^{\text{post}}(\theta) = \frac{f^{\text{prior}}(\theta) l(\mathbf{y}_1, \dots, \mathbf{y}_n | \theta)}{\int_{\Theta} f^{\text{prior}}(\vartheta) l(\mathbf{y}_1, \dots, \mathbf{y}_n | \vartheta) d\vartheta} = C \cdot f^{\text{prior}}(\theta) l(\mathbf{y}_1, \dots, \mathbf{y}_n | \theta)$$

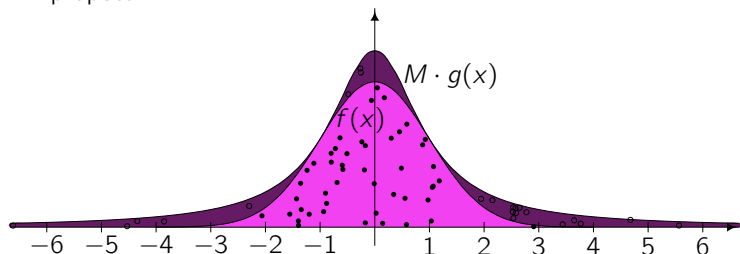
72

Example: Sampling from $N(0, 1)$

- Recall the following densities:

$$N(0, 1) \quad f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad \text{Cauchy} \quad g(x) = \frac{1}{\pi(1+x^2)}$$

- For $M = \sqrt{2\pi} \cdot \exp(-1/2)$ we have that $f(x) \leq Mg(x)$.
 \leadsto We can use rejection sampling targetting f using g as proposal.



73

Non-example: Sampling from a Cauchy Distribution

- We cannot sample from a Cauchy distribution (g) using a Normal (f) as instrumental distribution.
- The Cauchy distribution has heavier tails than the Normal distribution: there is no $M \in \mathbb{R}$ such that

$$\frac{1}{\pi(1+x^2)} < M \cdot \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{x^2}{2}\right).$$

74

An Alternative to Rejection

- Rejection sampling discards many samples.
- This seems wasteful.
- Couldn't we, instead, *weight* samples based on the acceptance probability?

75

The fundamental identities behind importance sampling

Assume that $g(x) > 0$ for (almost) all x with $f(x) > 0$:

$$\mathbb{P}(X \in \mathcal{X}) = \int_{\mathcal{X}} f(x) dx = \int_{\mathcal{X}} g(x) \underbrace{\frac{f(x)}{g(x)}}_{=:w(x)} dx = \int_{\mathcal{X}} g(x)w(x) dx$$

Assume that $g(x) > 0$ for (almost) all x with $f(x) \cdot \varphi(x) \neq 0$

$$\begin{aligned} \mathbb{E}_f(\varphi(X)) &= \int f(x)\varphi(x) dx = \int g(x) \underbrace{\frac{f(x)}{g(x)}}_{=:w(x)} \varphi(x) dx \\ &= \int g(x)w(x)\varphi(x) dx = \mathbb{E}_g(w(X) \cdot \varphi(X)), \end{aligned}$$

76

The fundamental identities behind importance sampling

- Consider $X_1, \dots, X_n \sim g$ and $\mathbb{E}_g |w(X) \cdot \varphi(X)| < +\infty$. Then

$$\frac{1}{n} \sum_{i=1}^n w(X_i) \varphi(X_i) \xrightarrow[n \rightarrow \infty]{a.s.} \mathbb{E}_g(w(X) \cdot \varphi(X))$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^n w(X_i) \varphi(X_i) \xrightarrow[n \rightarrow \infty]{a.s.} \mathbb{E}_f(\varphi(X)).$$

- Thus we can estimate $\mu := \mathbb{E}_f(\varphi(X))$ by

- 1 Sample $X_1, \dots, X_n \sim g$
- 2 $\tilde{\mu} := \frac{1}{n} \sum_{i=1}^n w(X_i) \varphi(X_i)$

77

The importance sampling algorithm

Algorithm: Importance Sampling

Choose g such that $\text{supp}(g) \supset \text{supp}(f \cdot \varphi)$.

- 1 For $i = 1, \dots, n$:

- 1 Generate $X_i \sim g$.
- 2 Set $w(X_i) = \frac{f(X_i)}{g(X_i)}$.

- 3 Return

$$\tilde{\mu} = \frac{\sum_{i=1}^n w(X_i) \varphi(X_i)}{n}$$

as an estimate of $\mathbb{E}_f(\varphi(X))$.

- Importance sampling does not yield realisations from f ,
 \rightsquigarrow but a *weighted sample* (X_i, W_i) ,
 \rightsquigarrow which can be used for estimating expectations $\mathbb{E}_f(\varphi(X))$,
 \rightsquigarrow or approximating f itself.

78

Basic properties of the importance sampling estimate

- We have already seen that $\tilde{\mu}$ is consistent if $\text{supp}(g) \supset \text{supp}(f \cdot \varphi)$ and $\mathbb{E}_g |w(X) \cdot \varphi(X)| < +\infty$, as

$$\tilde{\mu} := \frac{1}{n} \sum_{i=1}^n w(X_i) \varphi(X_i) \xrightarrow[n \rightarrow \infty]{a.s.} \mathbb{E}_f(\varphi(X))$$

- The expected value of the weights is $\mathbb{E}_g(w(X)) = 1$.
- $\tilde{\mu}$ is unbiased (see theorem below)

Theorem 2.2: Bias and Variance of Importance Sampling

$$\begin{aligned} \mathbb{E}_g(\tilde{\mu}) &= \mu \\ \text{Var}_g(\tilde{\mu}) &= \frac{\text{Var}_g(w(X) \cdot \varphi(X))}{n} \end{aligned}$$

79

Optimal proposals

Theorem (Optimal proposal)

The proposal distribution g that minimises the variance of $\tilde{\mu}$ is

$$g^*(x) = \frac{|\varphi(x)| f(x)}{\int |\varphi(t)| f(t) dt}.$$

- Theorem of little practical use: the optimal proposal involves $\int |\varphi(t)| f(t) dt$, which is the integral we want to estimate!
- Practical relevance:
Choose g such that it is close to $|\varphi(x)| \cdot f(x)$

80

Super-efficiency of importance sampling

- For the optimal g^* we have that

$$\text{Var}_f \left(\frac{\varphi(X_1) + \dots + \varphi(X_n)}{n} \right) > \text{Var}_{g^*}(\tilde{\mu}),$$

if φ is not almost surely constant.

Superefficiency of importance sampling

The variance of the importance sampling estimate can be *less* than the variance obtained when sampling directly from the target f .

- Intuition: Importance sampling allows us to choose a g that focuses on areas which contribute most to $\int \varphi(x)f(x) dx$.
- Even sub-optimal proposals can be super-efficient.

81

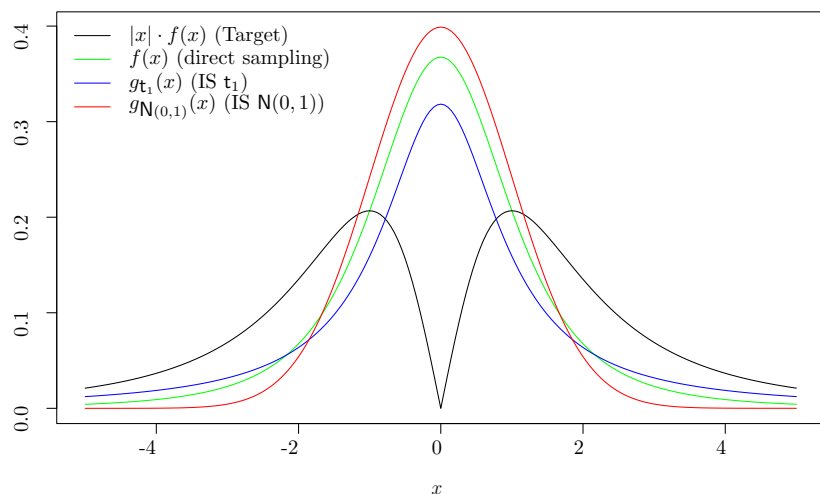
Importance Sampling Example 1: Setup

Compute $\mathbb{E}_f|X|$ for $X \sim t_3$ by ...

- sampling directly from t_3 .
- using a t_1 distribution as instrumental distribution.
- using a $N(0, 1)$ distribution as instrumental distribution.

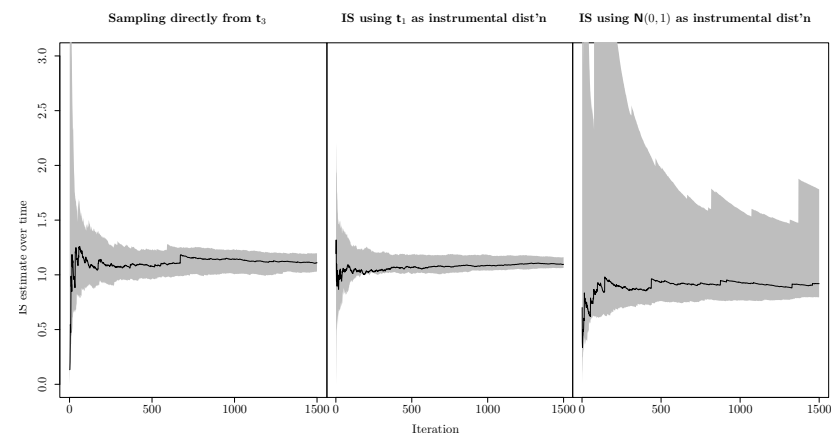
82

IS Example: Densities



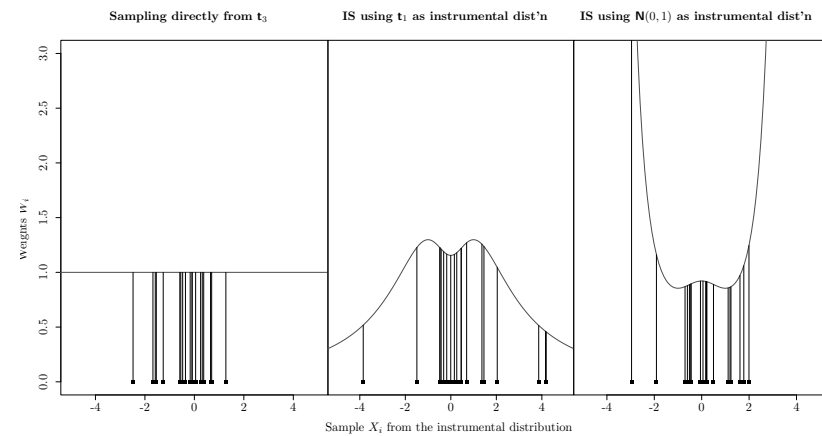
83

IS Example: Estimates obtained



84

IS Example: Weights



85

Another Example: Rare Events (1)

Consider

$$f(x, y) = \mathbf{N} \left(\begin{pmatrix} x \\ y \end{pmatrix}; \mu, \Sigma \right)$$

where

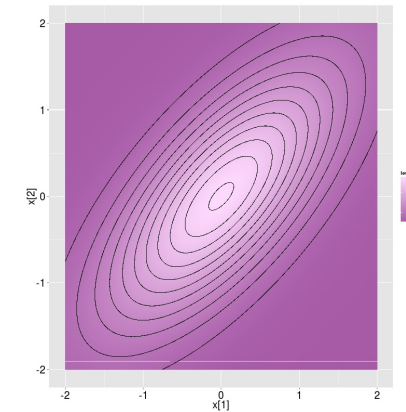
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

For

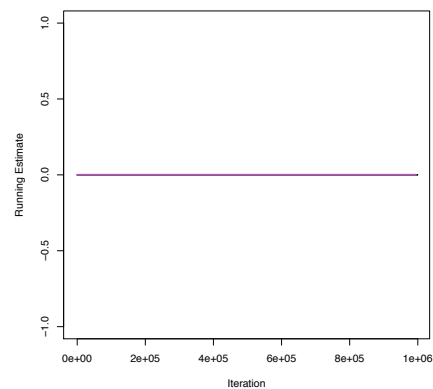
$$\varphi(x, y) = \mathbb{I}_{[4, \infty)}(x) \mathbb{I}_{[4, \infty)}(y)$$



86

Another Example: Rare Events (2)

Using simple Monte Carlo with 1,000,000 samples from f :

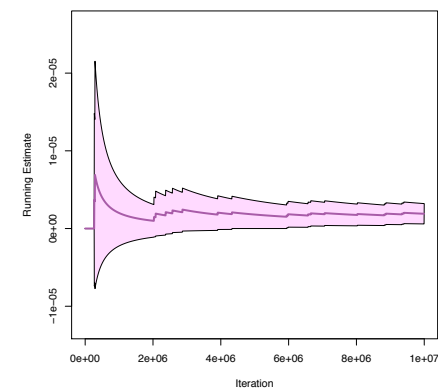


shaded region shows *estimated* 99.7% confidence interval.

87

Another Example: Rare Events (3)

Using simple Monte Carlo with 10,000,000 samples from f :

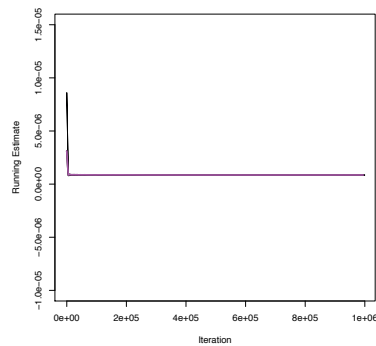


shaded region shows *estimated* 99.7% confidence interval.

88

Another Example: Rare Events (4)

Using importance sampling with 1,000,000 samples from
 $g(x, y) = \exp(-(x - 4) - (y - 4)) \mathbb{I}_{x \geq 4} \mathbb{I}_{y \geq 4}$:

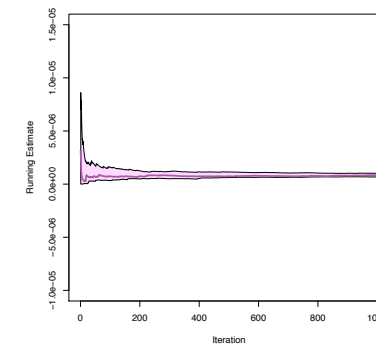


shaded region shows range of 100 replications.

89

Another Example: Rare Events (5)

Using importance sampling with 1,000 samples from
 $g(x, y) = \exp(-(x - 4) - (y - 4)) \mathbb{I}_{x \geq 4} \mathbb{I}_{y \geq 4}$:



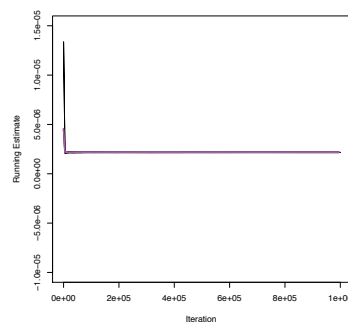
shaded region shows range of 100 replications.

90

Another Example: Rare Events (6)

Using importance sampling with 1,000,000 samples from

$$g(x, y) = 4N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \Sigma\right) \mathbb{I}_{x \geq 4} \mathbb{I}_{y \geq 4} :$$



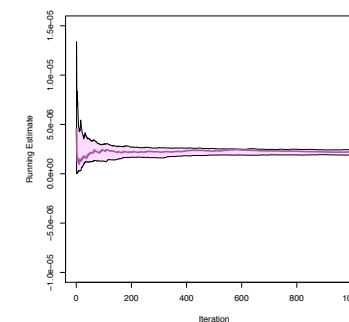
shaded region shows range of 100 replications.

91

Another Example: Rare Events (7)

Using importance sampling with 1,000 samples from

$$g(x, y) = 4N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \Sigma\right) \mathbb{I}_{x \geq 4} \mathbb{I}_{y \geq 4} :$$



shaded region shows range of 100 replications.

92

We only need f up to a multiplicative constant.

- Assume $f(x) = C\tilde{f}(x)$. Then

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^n w(X_i) \varphi(X_i) = \frac{1}{n} \sum_{i=1}^n \frac{C\tilde{f}(X_i)}{g(X_i)} \varphi(X_i)$$

$\leadsto C$ does not cancel out \leadsto knowing $\tilde{f}(\cdot)$ is not enough.

- Idea: Estimate C using the sample, via $\sum_{i=1}^n w(X_i)$, i.e. consider the *self-normalised estimator*

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n w(X_i) \varphi(X_i) / \frac{1}{n} \sum_{i=1}^n w(X_i)$$

- Now we have that

$$\hat{\mu} = \frac{\sum_{i=1}^n w(X_i) \varphi(X_i)}{\sum_{i=1}^n w(X_i)} = \frac{\sum_{i=1}^n \frac{\pi(X_i)}{g(X_i)} \varphi(X_i)}{\sum_{i=1}^n \frac{\pi(X_i)}{g(X_i)} w(X_i)},$$

$\leadsto \hat{\mu}$ does not depend on C

93

The importance sampling algorithm (2)

Algorithm: Importance Sampling using self-normalised weights

Choose g such that $\text{supp}(g) \supset \text{supp}(f)$.

- 1 For $i = 1, \dots, n$:
 - 1 Generate $X_i \sim g$.
 - 2 Set $w(X_i) = \frac{f(X_i)}{g(X_i)}$.
- 2 Return

$$\hat{\mu} = \frac{\sum_{i=1}^n w(X_i) \varphi(X_i)}{\sum_{i=1}^n w(X_i)}$$

as an estimate of $\mathbb{E}_f(\varphi(X))$.

94

Basic properties of the self-normalised estimate

- $\hat{\mu}$ is consistent as

$$\hat{\mu} = \underbrace{\frac{\sum_{i=1}^n w(X_i) \varphi(X_i)}{n}}_{\rightarrow \tilde{\mu} \rightarrow \mathbb{E}_f(\varphi(X))} \underbrace{\frac{\sum_{i=1}^n w(X_i)}{n}}_{\rightarrow 1} \xrightarrow{n \rightarrow \infty} \mathbb{E}_f(\varphi(X)),$$

(provided $\text{supp}(g) \supset \text{supp}(f)$ and $\mathbb{E}_g|w(X) \cdot \varphi(X)| < +\infty$)

Theorem: Bias and Variance (ctd.)

$$\begin{aligned} \mathbb{E}_g(\hat{\mu}) &= \mu + \frac{\mu \text{Var}_g(w(X)) - \text{Cov}_g[w(X), w(X) \cdot \varphi(X)]}{n} + O(n^{-2}) \\ \text{Var}_g(\hat{\mu}) &= \frac{\text{Var}_g(w(X) \cdot \varphi(X)) - 2\mu \text{Cov}_g[w(X), w(X) \cdot \varphi(X)]}{n} \\ &\quad + \frac{\mu^2 \text{Var}_g(w(X))}{n} + O(n^{-2}) \end{aligned}$$

95

Finite variance estimators

- Importance sampling estimate consistent for large choice of g .
- More important in practice: *finite variance estimators*, i.e.

$$\text{Var}(\tilde{\mu}) = \text{Var}\left(\frac{\sum_{i=1}^n w(X_i) \varphi(X_i)}{n}\right) < +\infty$$

- Sufficient (albeit restrictive) conditions for finite variance of $\tilde{\mu}$:
 - $f(x) < M \cdot g(x)$ and $\text{Var}_f(\varphi(X)) < \infty$, or
 - E is compact, f is bounded above on E , and g is bounded below on E .
- Note: If f has heavier tails than g , then the weights may have *infinite* variance!

96

Summary of Part 2

- Pseudorandom Number Generators (and alternatives)
- Transformation: Inversion Methods, Box-Muller
- Rejection Sampling
- Importance Sampling

97

Part 3

Markov chain Monte Carlo

Part 3— Section 7

Motivation and Basics

Why do we need other, more complicated methods?

- Transformation's great when it works.
- Rejection sampling's good when M is small.
- Importance sampling works well with good proposals.
- What do we do when we can't meet any of these requirements?

100

One Approach

Markov Chain Monte Carlo methods (MCMC)

- Key idea: Create a *dependent* sample, i.e. $X^{(t)}$ depends on the previous value $X^{(t-1)}$.
→ allows for “local” updates.
- Yields an “approximate sample” from the target distribution*.
- More mathematically speaking: yields a Markov chain with the target distribution f as stationary distribution.
- Under conditions, the realised chain provides approximations of $\mathbb{E}_f[\varphi(X)]$ and of f itself.

* I don't think this is the right way to think, but it's pervasive terminology and so I mention it here.

101

Markov Chains

Markov Chain (NB Terminology varies)

A *discrete time* Markov process taking values in a *general space*:

$$X^{(0)} \sim \mu_0 \quad \text{Initial Dist.}$$

$$X^{(t)} | X^{(0)} = x^{(0)}, \dots, X^{(t-1)} = x^{(t-1)} \sim K(x^{(t-1)}, \cdot) \quad \text{Kernel}$$

Stationary Distribution

f is a *stationary* or *invariant* distribution for a Markov Chain on E with kernel K if

$$\int_A \int_E f(x) K(x, y) dx dy = \int_A f(y) dy$$

for all measurable sets A [or $\int f(x) K(x, y) dx = f(y)$].

102

Heuristically Motivating MCMC

- If $X^{(0)}, \dots$ is an f -invariant Markov chain and $X^{(t)} \sim f$ for some t then $X^{(t+s)} \sim f \quad \forall s \in \mathbb{N}$.
- So if $X^{(t)}$ is “approximately independent” of $X^{(t+s)}$ for large enough s then
 - $X^{(t)}, X^{(t+s)}, \dots, X^{(t+ks)}, \dots$ is approximately $\overset{\text{iid}}{\sim} f$,
 - $X^{(t+1)}, X^{(t+s+1)}, \dots, X^{(t+ks+1)}, \dots$ is approximately $\overset{\text{iid}}{\sim} f$,
 - \vdots
 - $X^{(t+s-1)}, X^{(t+2s-1)}, \dots, X^{(t+ks-1)}, \dots$ is approximately $\overset{\text{iid}}{\sim} f$.
- We might conjecture that for such a chain, for some large s :

$$\frac{1}{n} \sum_{k=1}^n \varphi(X^{(t+ks)}) \rightarrow \mathbb{E}_f[\varphi(X)] \quad \text{and} \quad \frac{1}{n} \sum_{k=1}^n \varphi(X^{(k)}) \rightarrow \mathbb{E}_f[\varphi(X)]$$

103

Some Questions to Answer

- Can we formalise this heuristic argument?
→ ergodic theory
- How can we construct f -invariant Markov kernels?
→ various types of sampler
- What properties of these kernels are important?
→ more ergodic theory
- How do we initialise the chain?
→ transient phases and burning
- How do we know if it's working?
→ ergodic theory and convergence diagnostics

104

Aperiodicity

Definition: Period

A Markov chain has a period d if there exists some partition of the state space, E_1, \dots, E_d with the properties that:

- $\forall i \neq j : E_i \cap E_j = \emptyset$
- $\bigcup_{i=1}^d E_i = E$
- The chain moves deterministically between elements of the partition:

$$\forall i, j, t, s : \mathbb{P}(X_{t+s} \in E_j | X_t \in E_i) = \begin{cases} 1 & j = i + s \bmod d \\ 0 & \text{otherwise.} \end{cases}$$

A Markov chain is *aperiodic* if its period is 1.

105

Irreducibility

Definition: Irreducibility

Given a distribution, f , over E , a Markov chain is said to be *f-irreducible* if for all points $x \in E$ and all measurable sets A such that $f(A) > 0$ there exists some t such that:

$$\int_A K^t(x, y) dy > 0.$$

If this condition holds with $t = 1$, then the chain is said to be *strongly f-irreducible*.

$$K^t(x, y) := \int K(x, z) K^{t-1}(z, y) dz \quad K^1(x, y) = K(x, y)$$

106

Transience and Recurrence I

Consider sets $A \subset E$ for f -irreducible Markov chains.

Let $\eta_A := \sum_{k=1}^{\infty} \mathbb{I}_A(X^{(k)})$.

Transience and Recurrence of Sets

A set A is recurrent if:

$$\forall x \in A : \mathbb{E}_x[\eta_A] = \infty.$$

A set is *uniformly transient* if there exists some $M < \infty$ such that:

$$\forall x \in A : \mathbb{E}_x[\eta_A] \leq M.$$

A set, $A \subset E$, is *transient* if it may be expressed as a countable union of uniformly transient sets.

107

Transience and Recurrence II

Transience Recurrence of Markov Chains

A Markov chain is *recurrent* if the following hold:

- The chain is f -irreducible for some distribution f .
- For every measurable set $A \subset E$ such that $\int_A f(y) dy > 0$, $\mathbb{E}_x[\eta_A] = \infty$ for every $x \in A$.

It is *transient* if it is f -irreducible for some distribution f and the entire space is transient.

In the case of irreducible chains, transience and recurrence are properties of the chain rather than individual states.

108

A Motivating Convergence Result

Theorem (A Simple Ergodic Theorem)

If $(X_i)_{i \in \mathbb{N}}$ is an f -irreducible, f -invariant, recurrent \mathbb{R}^d -valued Markov chain then the following strong law of large numbers holds for any integrable function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^t \varphi(\xi_i) \stackrel{a.s.}{=} \int \varphi(x) f(x) dx.$$

for almost every starting value x .

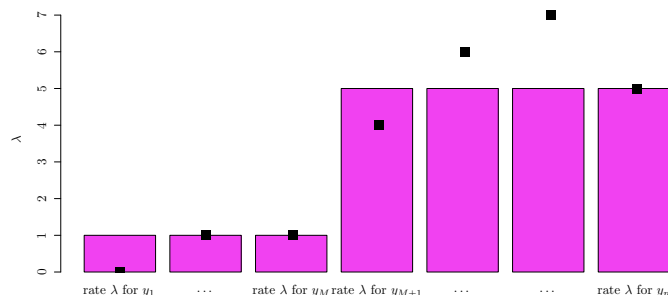
Note: this gives no rate of convergence.

109

Part 3— Section 8

The Gibbs Sampler

Example: Poisson change point model I



$$\begin{aligned} Y_i &\sim \text{Poi}(\lambda_1) & \text{for } i = 1, \dots, M \\ Y_i &\sim \text{Poi}(\lambda_2) & \text{for } i = M+1, \dots, n \end{aligned}$$

Objective: (Bayesian) inference about the parameters λ_1 , λ_2 , and M given observed data Y_1, \dots, Y_n .

111

Example: Poisson change point model II

- Prior distributions: $\lambda_j \sim \text{Gamma}(\alpha_j, \beta_j)$ ($j = 1, 2$), i.e.

$$f(\lambda_j) = \frac{1}{\Gamma(\alpha_j)} \lambda_j^{\alpha_j-1} \beta_j^{\alpha_j} \exp(-\beta_j \lambda_j).$$

(discrete uniform prior on M , i.e. $p(M) \propto 1$).

- Likelihood: $l(y_1, \dots, y_n | \lambda_1, \lambda_2, M)$

$$= \left(\prod_{i=1}^M \frac{\exp(-\lambda_1) \lambda_1^{y_i}}{y_i!} \right) \cdot \left(\prod_{i=M+1}^n \frac{\exp(-\lambda_2) \lambda_2^{y_i}}{y_i!} \right)$$

112

Example: Poisson change point model III

- Joint distribution $f(y_1, \dots, y_n, \lambda_1, \lambda_2, M)$

$$= l(y_1, \dots, y_n | \lambda_1, \lambda_2, M) \cdot f(\lambda_1) \cdot f(\lambda_2) \cdot p(M)$$

$$\propto \left(\prod_{i=1}^M \frac{\exp(-\lambda_1) \lambda_1^{y_i}}{y_i!} \right) \cdot \left(\prod_{i=M+1}^n \frac{\exp(-\lambda_2) \lambda_2^{y_i}}{y_i!} \right)$$

$$\cdot \frac{1}{\Gamma(\alpha_1)} \lambda_1^{\alpha_1-1} \beta_1^{\alpha_1} \exp(-\beta_1 \lambda_1) \cdot \frac{1}{\Gamma(\alpha_2)} \lambda_2^{\alpha_2-1} \beta_2^{\alpha_2} \exp(-\beta_2 \lambda_2)$$

- Joint posterior distribution $f(\lambda_1, \lambda_2, M | y_1, \dots, y_n)$

$$\propto \lambda_1^{\alpha_1-1+\sum_{i=1}^M y_i} \exp(-(\beta_1 + M) \lambda_1)$$

$$\cdot \lambda_2^{\alpha_2-1+\sum_{i=M+1}^n y_i} \exp(-(\beta_2 + n - M) \lambda_2)$$

Example: Poisson change point model IV

- Conditional on M (i.e. if M was known) we have

$$f(\lambda_1 | y_1, \dots, y_n, M) \propto \lambda_1^{\alpha_1-1+\sum_{i=1}^M y_i} \exp(-(\beta_1 + M) \lambda_1),$$

i.e.

$$\lambda_1 | Y_1, \dots, Y_n, M \sim \text{Gamma} \left(\alpha_1 + \sum_{i=1}^M y_i, \beta_1 + M \right)$$

$$\lambda_2 | Y_1, \dots, Y_n, M \sim \text{Gamma} \left(\alpha_2 + \sum_{i=M+1}^n y_i, \beta_2 + n - M \right).$$

- $p(M | \dots) \propto \lambda_1^{\sum_{i=1}^M y_i} \cdot \lambda_2^{\sum_{i=M+1}^n y_i} \cdot \exp((\lambda_2 - \lambda_1) \cdot M)$

Example: Poisson change point model V

This suggests an iterative algorithm:

1. Draw λ_1 from $\lambda_1 | Y_1, \dots, Y_n, M$, i.e. draw

$$\lambda_1 \sim \text{Gamma} \left(\alpha_1 + \sum_{i=1}^M y_i, \beta_1 + M \right)$$

2. Draw λ_2 from $\lambda_2 | Y_1, \dots, Y_n, M$, i.e. draw

$$\lambda_2 \sim \text{Gamma} \left(\alpha_2 + \sum_{i=M+1}^n y_i, \beta_2 + n - M \right)$$

3. Draw M from $M | Y_1, \dots, Y_n, \lambda_1, \lambda_2$, i.e. draw

$$p(M) \propto \lambda_1^{\sum_{i=1}^M y_i} \cdot \lambda_2^{\sum_{i=M+1}^n y_i} \cdot \exp((\lambda_2 - \lambda_1) \cdot M)$$

The systematic scan Gibbs sampler

Algorithm: (Systematic scan) Gibbs sampler

Starting with $(X_1^{(0)}, \dots, X_p^{(0)})$ iterate for $t = 1, 2, \dots$

1. Draw $X_1^{(t)} \sim f_{X_1 | X_{-1}}(\cdot | X_2^{(t-1)}, \dots, X_p^{(t-1)})$.

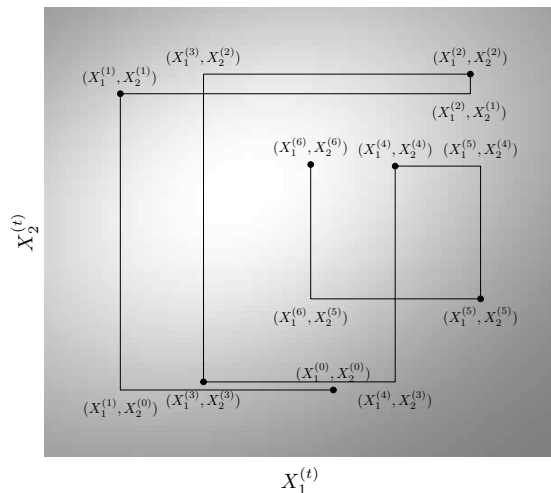
...

- j. Draw $X_j^{(t)} \sim f_{X_j | X_{-j}}(\cdot | X_1^{(t)}, \dots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \dots, X_p^{(t-1)})$.

...

- p. Draw $X_p^{(t)} \sim f_{X_p | X_{-p}}(\cdot | X_1^{(t)}, \dots, X_{p-1}^{(t)})$.

Illustration of the systematic scan Gibbs sampler



117

The random scan Gibbs sampler

Algorithm: (Random scan) Gibbs sampler

Starting with $(X_1^{(0)}, \dots, X_p^{(0)})$ iterate for $t = 1, 2, \dots$

- ➊ Draw an index j from a distribution on $\{1, \dots, p\}$ (e.g. uniform)
- ➋ Draw $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot | X_1^{(t-1)}, \dots, X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)}, \dots, X_p^{(t-1)})$, and set $X_\iota^{(t)} := X_\iota^{(t-1)}$ for all $\iota \neq j$.

118

Invariant distribution

Lemma (Kernel)

The transition kernel of the systematic scan Gibbs sampler is

$$K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) = f_{X_1|X_{-1}}(x_1^{(t)} | x_2^{(t-1)}, \dots, x_p^{(t-1)}) \\ \cdot f_{X_2|X_{-2}}(x_2^{(t)} | x_1^{(t)}, x_3^{(t-1)}, \dots, x_p^{(t-1)}) \\ \cdot \dots \\ \cdot f_{X_p|X_{-p}}(x_p^{(t)} | x_1^{(t)}, \dots, x_{p-1}^{(t)})$$

Proposition (Invariance)

The joint distribution $f(x_1, \dots, x_p)$ is indeed the invariant distribution of the Markov chain $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$ generated by the Gibbs sampler.

119

Proof (outline) I

Assume that $\mathbf{X}^{(t-1)} \sim f$, then

$$\mathbb{P}(\mathbf{X}^{(t)} \in \mathcal{X}) = \int_{\mathcal{X}} \int f(\mathbf{x}^{(t-1)}) K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) d\mathbf{x}^{(t-1)} d\mathbf{x}^{(t)}$$

We can expand the $K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)})$ of the integrand, and compute the $x_1^{(t-1)}$ -integral:

$$\underbrace{\int f(x_1^{(t-1)}, \dots, x_p^{(t-1)}) dx_1^{(t-1)} f_{X_1|X_{-1}}(x_1^{(t)} | x_2^{(t-1)}, \dots, x_p^{(t-1)})}_{=f(x_2^{(t-1)}, \dots, x_p^{(t-1)})} \\ = f(x_1^{(t)}, x_2^{(t-1)}, \dots, x_p^{(t-1)}) \\ f_{X_2|X_{-2}}(x_2^{(t)} | x_1^{(t)}, \dots, x_p^{(t-1)}) \dots f_{X_p|X_{-p}}(x_p^{(t)} | x_1^{(t)}, \dots, x_{p-1}^{(t)})$$

120

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ●○○○○○	○○○○○ ○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○

The Algorithm

Proof (outline) II

And we can then compute the $x_2^{(t-1)}$ integral:

$$\int \underbrace{f(x_1^{(t)}, x_2^{(t-1)}, \dots, x_p^{(t-1)})}_{=f(x_1^{(t)}, x_3^{(t-1)}, \dots, x_p^{(t-1)})} dx_2^{(t-1)} \underbrace{f_{X_2|X_{-2}}(x_2^{(t)}|x_1^{(t)}, x_3^{(t-1)}, \dots, x_p^{(t-1)})}_{=f(x_1^{(t)}, x_2^{(t)}, x_3^{(t-1)}, \dots, x_p^{(t-1)})} \dots f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)}, \dots, x_{p-1}^{(t)})$$

And so on until the $x_p^{(t-1)}$ -integral:

$$\int \underbrace{f(x_1^{(t)}, \dots, x_{p-1}^{(t)}, x_p^{(t-1)})}_{=f(x_1^{(t)}, \dots, x_{p-1}^{(t)})} dx_p^{(t-1)} \underbrace{f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)}, \dots, x_{p-1}^{(t)})}_{=f(x_1^{(t)}, \dots, x_p^{(t)})}$$

121

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ●○○○○○	○○○○○ ○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○

The Algorithm

Proof (outline) III

This just leaves the $\mathbf{x}^{(t)}$ -integrals:

$$\mathbb{P}(\mathbf{X}^{(t)} \in \mathcal{X}) = \int_{\mathcal{X}} f(x_1^{(t)}, \dots, x_p^{(t)}) d\mathbf{x}^{(t)}$$

Thus f is the density of $\mathbf{X}^{(t)}$ (if $\mathbf{X}^{(t-1)} \sim f$).

122

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ●○○○○○	○○○○○ ○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○

Examples

Recall our Poisson Changepoint Model

- Joint posterior distribution $f(\lambda_1, \lambda_2, M|y_1, \dots, y_n)$

$$\propto \lambda_1^{\alpha_1-1+\sum_{i=1}^M y_i} \exp(-(\beta_1 + M)\lambda_1) \cdot \lambda_2^{\alpha_2-1+\sum_{i=M+1}^n y_i} \exp(-(\beta_2 + n - M)\lambda_2)$$

- Full Posterior Distributions

$$\lambda_1|Y_1, \dots, Y_n, M \sim \text{Gamma}\left(\alpha_1 + \sum_{i=1}^M y_i, \beta_1 + M\right)$$

$$\lambda_2|Y_1, \dots, Y_n, M \sim \text{Gamma}\left(\alpha_2 + \sum_{i=M+1}^n y_i, \beta_2 + n - M\right).$$

- and $p(M|\dots) \propto \lambda_1^{\sum_{i=1}^M y_i} \cdot \lambda_2^{\sum_{i=M+1}^n y_i} \cdot \exp((\lambda_2 - \lambda_1) \cdot M)$

123

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ●○○○○○	○○○○○ ○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○

Examples

An R Implementation

```
cdist.M <- function(lambda1, lambda2) {
  dist.M.log <- cumsum(y[1:n-1]) * log(lambda1) +
    (sum(y) - cumsum(y[1:n-1])) * log(lambda2) +
    (lambda2 - lambda1) * (1:(n-1))
  dist.M <- exp(dist.M.log - mean(dist.M.log))
  dist.M <- dist.M / sum(dist.M)
}

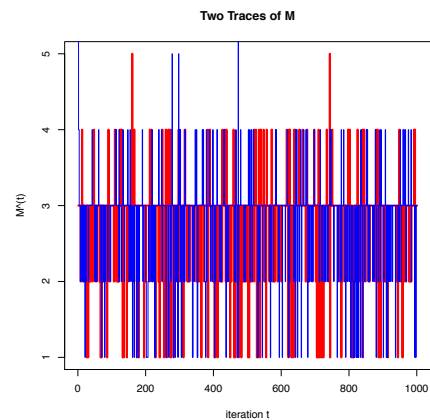
pmix.gibbs <- function(M, lambda1, lambda2, t) {
  r <- array(NA, c(t+1, 3))
  r[1,] <- c(M, lambda1, lambda2)
  for (i in 1:t) {
    #lambda1
    r[i+1, 2] <- rgamma(1, a1 + sum(y[1:r[i, 1]]), b1 + r[i, 1])
    #lambda2
    r[i+1, 3] <- rgamma(1, a2 + sum(y[(r[i, 1] + 1):n]), b2 + n - r[i, 1])
    #M
    r[i+1, 1] <- sample.int(n-1, 1, prob = cdist.M(r[i+1, 2], r[i+1, 3]))
  }
  r
}
```

124

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○●○○○○○○○○○	○○○○○ ○○○○○○○○○○○○○○○○○ ○○○	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Examples

Traces and Estimates: M



Consider two differently-initialised chains.

Chain 1:

$$(M, \lambda_1, \lambda_2)^{(0)} = (3, 1, 2)$$

Chain 2:

$$(M, \lambda_1, \lambda_2)^{(0)} = (6, 4, \frac{1}{2})$$

Estimated Posterior

Modes:

Chain 1: 3

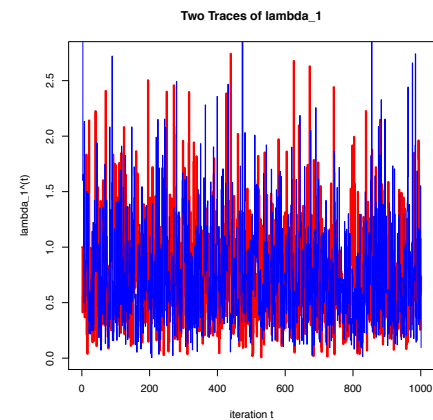
Chain 2: 3

125

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○●○○○○○○○○○	○○○○○ ○○○○○○○○○○○○○○○○○ ○○○	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Examples

Traces and Estimates: λ_1



Estimated Posterior

Means:

Chain 1: 0.76

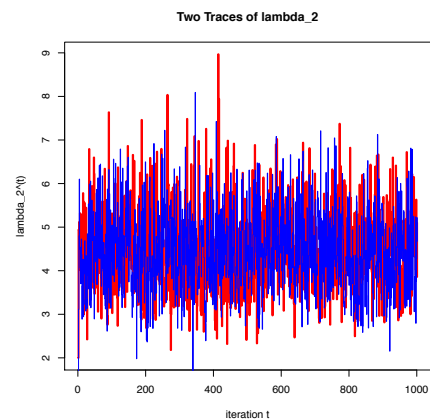
Chain 2: 0.78

126

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○●○○○○○○○○○	○○○○○ ○○○○○○○○○○○○○○○○○ ○○○	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Examples

Traces and Estimates: λ_2



Estimated Posterior

Means:

Chain 1: 4.51

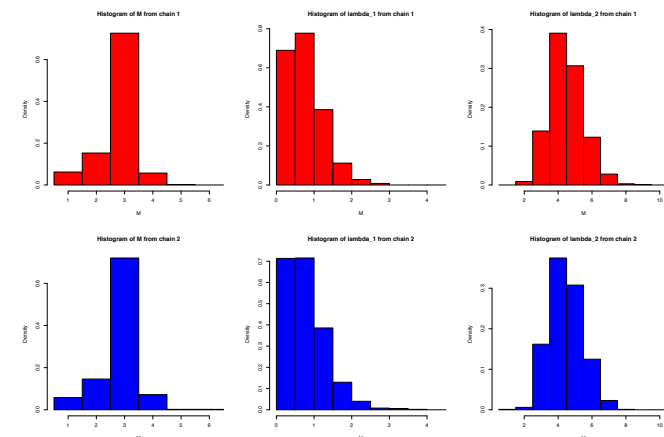
Chain 2: 4.47

127

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○●○○○○○○○○○	○○○○○ ○○○○○○○○○○○○○○○○○ ○○○	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Examples

Histograms: Approximations of the Posterior



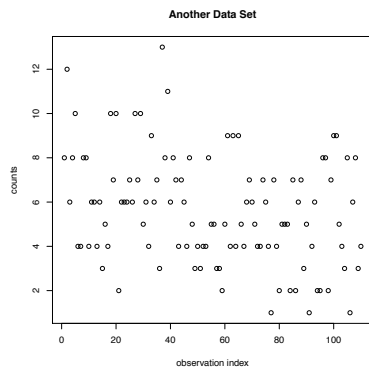
128

Motivation ○○○○○ ○○○	Gibbs Samplers ○ ○○○○○ ○○○○○○●○○○○○	Metropolis-Hastings ○○○○○ ○○○○○○○○○○○○○○○○ ○○○	Simulated Annealing ○○○○○○○○○ ○○○○○○○
----------------------------	--	---	---

Examples

Poisson Change-Point Model: More Challenging Data I

Consider the more realistic data:



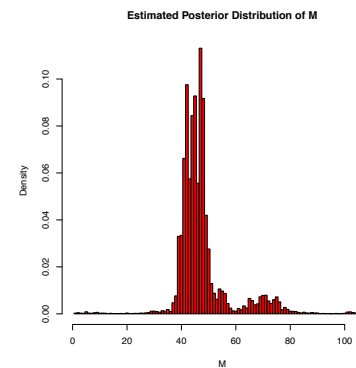
129

Motivation ○○○○○ ○○○	Gibbs Samplers ○ ○○○○○ ○○○○○○●○○○○○	Metropolis-Hastings ○○○○○ ○○○○○○○○○○○○○○○○ ○○○	Simulated Annealing ○○○○○○○○○ ○○○○○○○
----------------------------	--	---	---

Examples

Poisson Change-Point Model: More Challenging Data II

From a chain of length 100,000 we obtain the following histograms:



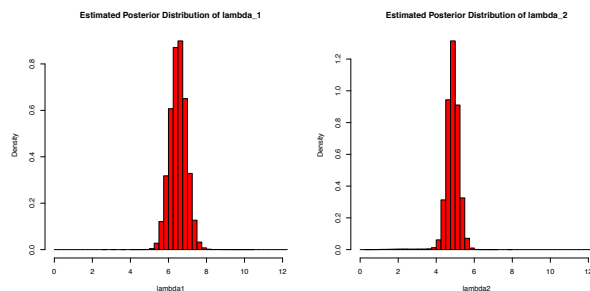
130

Motivation ○○○○○ ○○○	Gibbs Samplers ○ ○○○○○ ○○○○○○●○○○○○	Metropolis-Hastings ○○○○○ ○○○○○○○○○○○○○○○○ ○○○	Simulated Annealing ○○○○○○○○○ ○○○○○○○
----------------------------	--	---	---

Examples

Poisson Change-Point Model: More Challenging Data III

Data was generated with: $y \leftarrow c(rpois(40,7), rpois(70,5))$

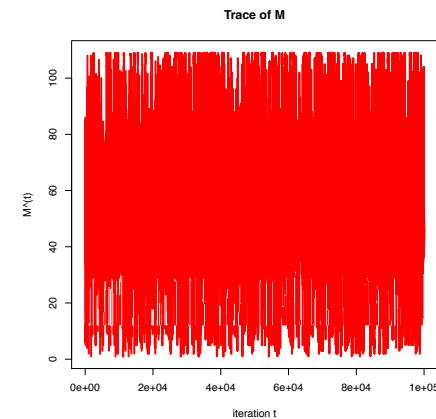


131

Motivation ○○○○○ ○○○	Gibbs Samplers ○ ○○○○○ ○○○○○○●○○○○○	Metropolis-Hastings ○○○○○ ○○○○○○○○○○○○○○○○ ○○○	Simulated Annealing ○○○○○○○○○ ○○○○○○○
----------------------------	--	---	---

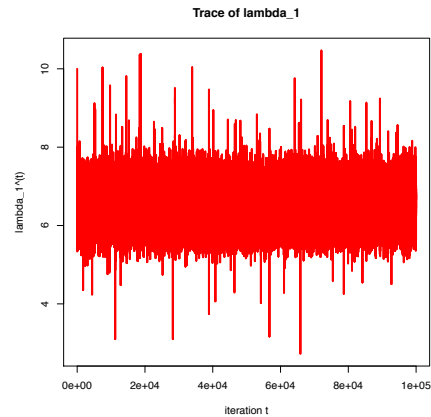
Examples

Poisson Change-Point Model: More Challenging Data IV



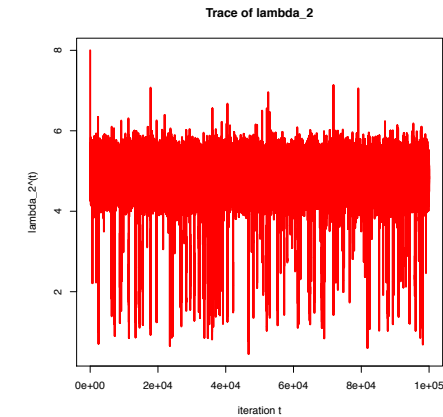
132

Poisson Change-Point Model: More Challenging Data V



133

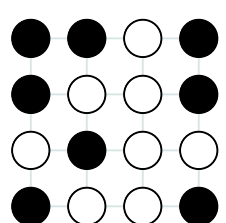
Poisson Change-Point Model: More Challenging Data VI



134

The Ising Model

The Ising model on $(\mathcal{V}, \mathcal{E})$ each $v_i \in \mathcal{V}$ has an associated $x_i \in \{-1, +1\}$:



$$\pi(x_1, \dots, x_m)$$

$$= \frac{1}{Z} \exp \left(J \sum_{(i,j) \in \mathcal{E}} x_i \cdot x_j \right)$$

$$= \frac{1}{Z} \exp(-J|\mathcal{E}|) \exp \left(2J \sum_{(i,j) \in \mathcal{E}} \mathbb{I}(x_i = x_j) \right)$$

$$= \frac{1}{Z'} \exp \left(2J \sum_{(i,j) \in \mathcal{E}} \mathbb{I}(x_i = x_j) \right)$$

$$\pi(x_j | x_{-j}) = \exp \left(J \sum_{i:(i,j) \in \mathcal{E}} x_i x_j \right) / \left[\exp \left(-J \sum_{i:(i,j) \in \mathcal{E}} x_i \right) + \exp \left(J \sum_{i:(i,j) \in \mathcal{E}} x_i \right) \right]$$

135

The Core Logic in R

```
tr <- list()
tr[[1]] <- x <- array(0, c(m, n))

for (t in 1:100) {
  for(i in 1:m) {
    for(j in 1:n) {
      ns <- neighbours(m, n, i, j)
      p1 <- 0
      for(k in 1:length(ns)) {
        p1 <- p1 + x[(ns[[k]])[1], (ns[[k]])[2]]
      }
      p0 <- length(ns) - p1
      pp <- c(exp(J*p0), exp(J*p1))
      pp <- pp / sum(pp)
      x[i, j] <- sample(c(0, 1), 1, prob=pp)
    }
  }
  tr[[t+1]] <- x
}
```

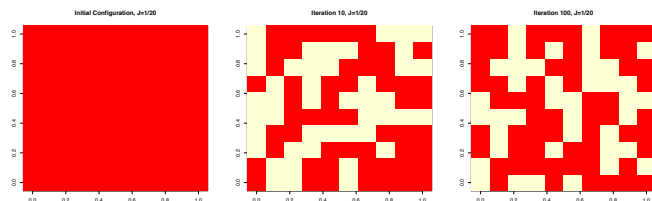
136

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○ ○○○○○	○ ○○○○○ ○○○○○●○○○○○	○○○○○ ○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○○○ ○○○○○○○○○

Examples

The Gibbs Sampler for Ising Models I

Samples 1, 10 and 100 with $J = 0.05$:



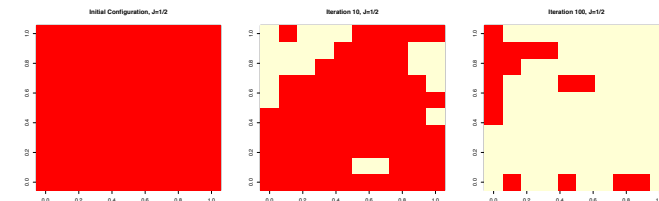
137

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○ ○○○○○	○ ○○○○○ ○○○○○●○○○○○	○○○○○ ○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○○○ ○○○○○○○○○

Examples

The Gibbs Sampler for Ising Models II

Samples 1, 10 and 100 with $J = 0.50$:



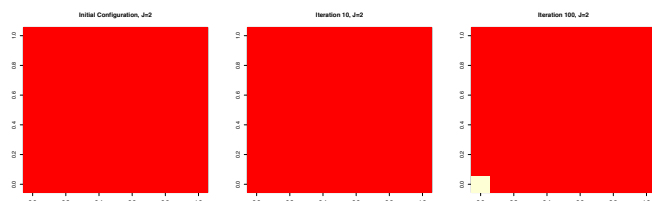
138

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○ ○○○○○	○ ○○○○○ ○○○○○●○○○○○	○○○○○ ○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○○○ ○○○○○○○○○

Examples

The Gibbs Sampler for Ising Models III

Samples 1, 10 and 100 with $J = 1.00$:



Solutions include the *Swendsen-Wang* algorithm (cf. assessment) or *perfect simulation*...

139

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○ ○○○○○	○ ○○○○○ ○○○○○●○○○○○	○○○○○ ○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○○○ ○○○○○○○○○

Examples

The Ising Model and Image Reconstruction

The Ising Model is widely used in statistics as a prior distribution.

- Consider image denoising: x an $m \times n$ image on $\mathcal{V} \subset \mathbb{Z}^2$ with obvious neighbourhood structure \mathcal{E} :
- Observe y where $y_v = x_v$ wp $1-\epsilon$.
- Prior: $X \sim \text{Ising}(J, \mathcal{V}, \mathcal{E})$.
- Likelihood: $l(y; x) = \prod_{v \in \mathcal{V}} [(1 - \epsilon)\mathbb{I}\{y_v = x_v\} + \epsilon\mathbb{I}\{y_v \neq x_v\}]$
- Posterior:

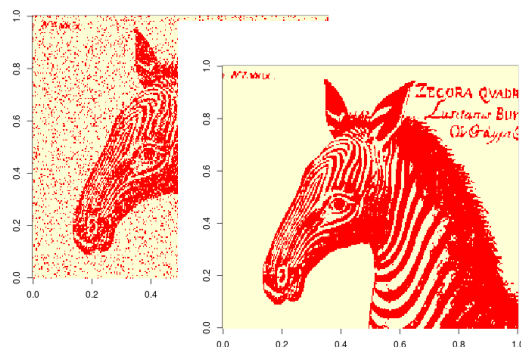
$$p(x|y) \propto \exp \left(2J \sum_{(i,j) \in \mathcal{E}} \mathbb{I}(x_i = x_j) \right) \cdot \prod_{v \in \mathcal{V}} [(1 - \epsilon)\mathbb{I}\{y_v = x_v\} + \epsilon\mathbb{I}\{y_v \neq x_v\}]$$

140

Examples

Ludolphus' Zebra

<https://upload.wikimedia.org/wikipedia/commons/a/af/ZebraLudolphus.jpg>



Noisy Image / Samples

Ground Truth

Examples

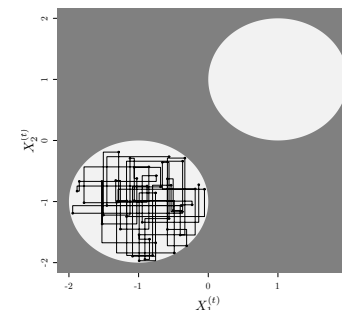
A Pathological Example: The Reducible Gibbs sampler

Consider Gibbs sampling from the uniform distribution

$$f(x_1, x_2) = \frac{1}{2\pi} \mathbb{I}_{C_1 \cup C_2}(x_1, x_2),$$

$$C_1 := \{(x_1, x_2) : \|(x_1, x_2) - (1, 1)\| \leq 1\}$$

$$C_2 := \{(x_1, x_2) : \|(x_1, x_2) + (1, 1)\| \leq 1\}$$



The resulting Markov chain is *reducible*:

It stays forever in either C_1 or C_2 .

142

Part 3— Section 9

The Metropolis-Hastings Algorithm

The Algorithm

The Metropolis-Hastings algorithm

Algorithm: Metropolis-Hastings

Starting with $\mathbf{X}^{(0)} := (X_1^{(0)}, \dots, X_p^{(0)})$ iterate for $t = 1, 2, \dots$

1 Draw $\mathbf{X} \sim q(\cdot | \mathbf{X}^{(t-1)})$.

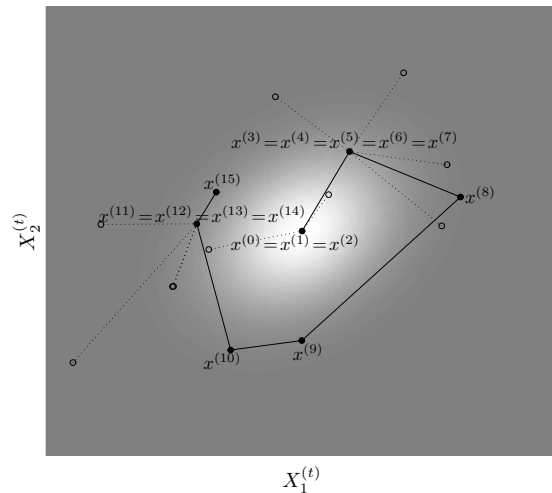
2 Compute

$$\alpha(\mathbf{X} | \mathbf{X}^{(t-1)}) = \min \left\{ 1, \frac{f(\mathbf{X}) \cdot q(\mathbf{X}^{(t-1)} | \mathbf{X})}{f(\mathbf{X}^{(t-1)}) \cdot q(\mathbf{X} | \mathbf{X}^{(t-1)})} \right\}.$$

3 With probability $\alpha(\mathbf{X} | \mathbf{X}^{(t-1)})$ set $\mathbf{X}^{(t)} = \mathbf{X}$, otherwise set $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$.

144

Illustration of the Metropolis-Hastings method



145

Basic properties of the Metropolis-Hastings algorithm

- The probability that a newly proposed value is accepted given $\mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}$ is

$$a(\mathbf{x}^{(t-1)}) = \int \alpha(\mathbf{x}|\mathbf{x}^{(t-1)})q(\mathbf{x}|\mathbf{x}^{(t-1)}) d\mathbf{x}.$$

- The probability of remaining in state $\mathbf{X}^{(t-1)}$ is
- $$\mathbb{P}(\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)} | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}) = 1 - a(\mathbf{x}^{(t-1)}).$$
- The probability of acceptance does not depend on the normalisation constant:
If $f(\mathbf{x}) = C \cdot \tilde{f}(\mathbf{x})$, then

$$\alpha(\mathbf{X}|\mathbf{X}^{(t-1)}) = \min \left(1, \frac{\tilde{f}(\mathbf{X}) \cdot q(\mathbf{X}^{(t-1)}|\mathbf{X})}{\tilde{f}(\mathbf{X}^{(t-1)}) \cdot q(\mathbf{X}|\mathbf{X}^{(t-1)})} \right)$$

146

Transition Kernel

Lemma (Transition Kernel of Metropolis-Hastings)

The transition kernel of the Metropolis-Hastings algorithm is

$$K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) = \alpha(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) + (1 - a(\mathbf{x}^{(t-1)}))\delta_{\mathbf{x}^{(t-1)}}(\mathbf{x}^{(t)}),$$

Lemma (Detailed Balance and Metropolis Hastings)

The Metropolis-Hastings kernel satisfies the detailed balance condition

$$K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)})f(\mathbf{x}^{(t-1)}) = K(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)})f(\mathbf{x}^{(t)}).$$

147

f -invariance of Metropolis-Hastings

Proposition (Detailed Balanced implies Invariance)

Any K which satisfies the detailed balance condition with respect to f ,

$$K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)})f(\mathbf{x}^{(t-1)}) = K(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)})f(\mathbf{x}^{(t)}),$$

is f -invariant.

Proof

Integrate both sides wrt $\mathbf{x}^{(t-1)}$.

Hence the Metropolis-Hastings algorithm is f -invariant.

148

Random-walk Metropolis: Idea

- In the Metropolis-Hastings algorithm the proposal is from $\mathbf{X} \sim q(\cdot | \mathbf{X}^{(t-1)})$.
- A popular choice for the proposal is $q(\mathbf{x} | \mathbf{x}^{(t-1)}) = g(\mathbf{x} - \mathbf{x}^{(t-1)})$ with g being a *symmetric* distribution, thus

$$\mathbf{X} = \mathbf{X}^{(t-1)} + \varepsilon, \quad \varepsilon \sim g.$$

- Probability of acceptance becomes

$$\min \left\{ 1, \frac{f(\mathbf{X}) \cdot g(\mathbf{X} - \mathbf{X}^{(t-1)})}{f(\mathbf{X}^{(t-1)}) \cdot g(\mathbf{X}^{(t-1)} - \mathbf{X})} \right\} = \min \left\{ 1, \frac{f(\mathbf{X})}{f(\mathbf{X}^{(t-1)})} \right\},$$

- We accept ...
 - every move to a more probable state with probability 1.
 - moves to less probable states with a probability $f(\mathbf{X})/f(\mathbf{x}^{(t-1)}) < 1$.

149

Random-walk Metropolis: Algorithm

Random-Walk Metropolis

Starting with $\mathbf{X}^{(0)} := (X_1^{(0)}, \dots, X_p^{(0)})$ and using a symmetric random walk proposal g , iterate for $t = 1, 2, \dots$

- Draw $\varepsilon \sim g$ and set $\mathbf{X} = \mathbf{X}^{(t-1)} + \varepsilon$.
- Compute

$$\alpha(\mathbf{X} | \mathbf{X}^{(t-1)}) = \min \left\{ 1, \frac{f(\mathbf{X})}{f(\mathbf{X}^{(t-1)})} \right\}.$$

- With probability $\alpha(\mathbf{X} | \mathbf{X}^{(t-1)})$ set $\mathbf{X}^{(t)} = \mathbf{X}$, otherwise set $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$.

Popular choices for g are (multivariate) Gaussians or t-distributions (the latter having heavier tails)

150

Example 3.4: Bayesian probit model (1)

- Medical study on infections resulting from birth by Cæsarean section
- 3 influence factors:
 - indicator whether the Cæsarian was planned or not (z_{i1}),
 - indicator of whether additional risk factors were present at the time of birth (z_{i2}), and
 - indicator of whether antibiotics were given as a prophylaxis (z_{i3}).
- Response variable: number of infections Y_i that were observed amongst n_i patients having the same covariates.

# births		planned	risk factors	antibiotics
infection	total			
y_i	n_i	z_{i1}	z_{i2}	z_{i3}
11	98	1	1	1
1	18	0	1	1
0	2	0	0	1
23	26	1	1	0
28	58	0	1	0
0	9	1	0	0
8	40	0	0	0

151

Example 3.4: Bayesian probit model (2)

- Model for Y_i :

$$Y_i \sim \text{Bin}(n_i, \pi_i), \quad \pi_i = \Phi(\mathbf{z}_i' \boldsymbol{\beta}),$$

where $\mathbf{z}_i = (1, z_{i1}, z_{i2}, z_{i3})$ and $\Phi(\cdot)$ being the CDF of a $N(0, 1)$.

- Prior on the parameter of interest $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbb{I}/\lambda)$.
- The posterior density of $\boldsymbol{\beta}$ is

$$f(\boldsymbol{\beta} | y_1, \dots, y_n) \propto \left(\prod_{i=1}^N \Phi(\mathbf{z}_i' \boldsymbol{\beta})^{y_i} \cdot (1 - \Phi(\mathbf{z}_i' \boldsymbol{\beta}))^{n_i - y_i} \right) \cdot \exp \left(-\frac{\lambda}{2} \sum_{j=0}^3 \beta_j^2 \right)$$

152

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○	○○○○○ ○○○○○●○○○○○○○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Example 3.4: Bayesian probit model (3)

Use the following “random walk Metropolis” algorithm.
Starting with any $\beta^{(0)}$ iterate for $t = 1, 2, \dots$:

1. Draw $\epsilon \sim N(\mathbf{o}, \Sigma)$ and set $\beta = \beta^{(t-1)} + \epsilon$.
2. Compute

$$\alpha(\beta|\beta^{(t-1)}) = \min \left\{ 1, \frac{f(\beta|Y_1, \dots, Y_n)}{f(\beta^{(t-1)}|Y_1, \dots, Y_n)} \right\}.$$

3. With probability $\alpha(\beta|\beta^{(t-1)})$ set $\beta^{(t)} = \beta$, otherwise set $\beta^{(t)} = \beta^{(t-1)}$.

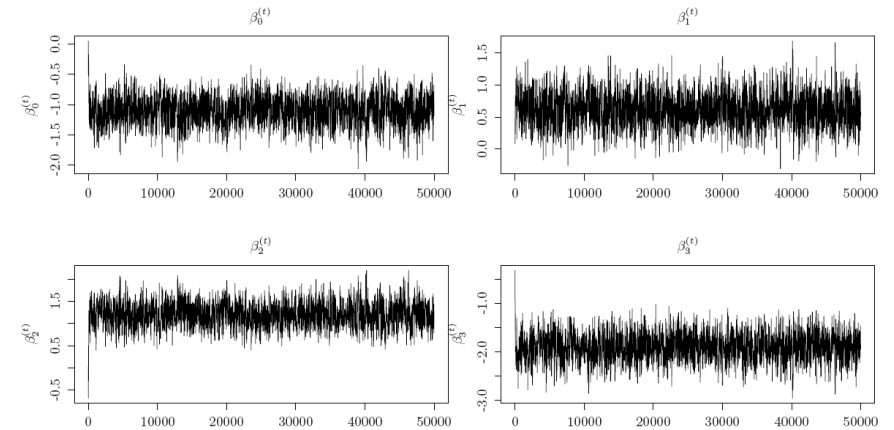
(for the moment we use $\Sigma = 0.08 \cdot \mathbb{I}$, and $\lambda = 10$).

153

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○	○○○○○ ○○○○○●○○○○○○○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Example 3.4: Bayesian probit model (4)



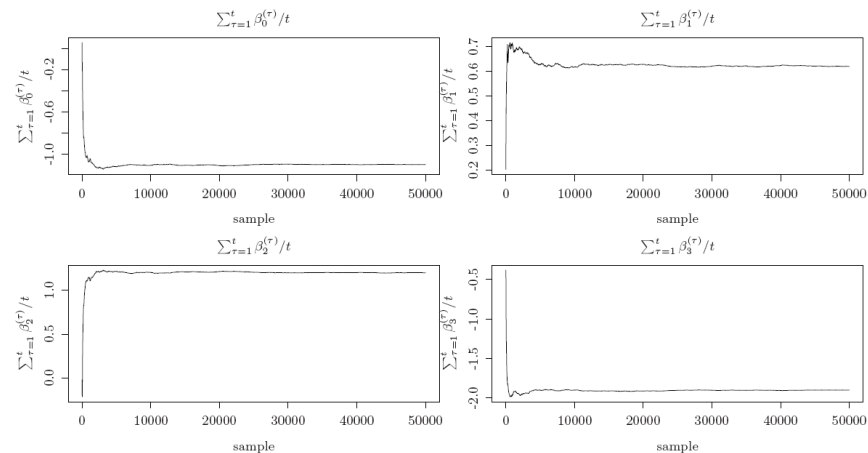
Convergence of the $\beta_j^{(t)}$ is to a distribution, not a value!

154

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○	○○○○○ ○○○○○●○○○○○○○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Example 3.4: Bayesian probit model (5)



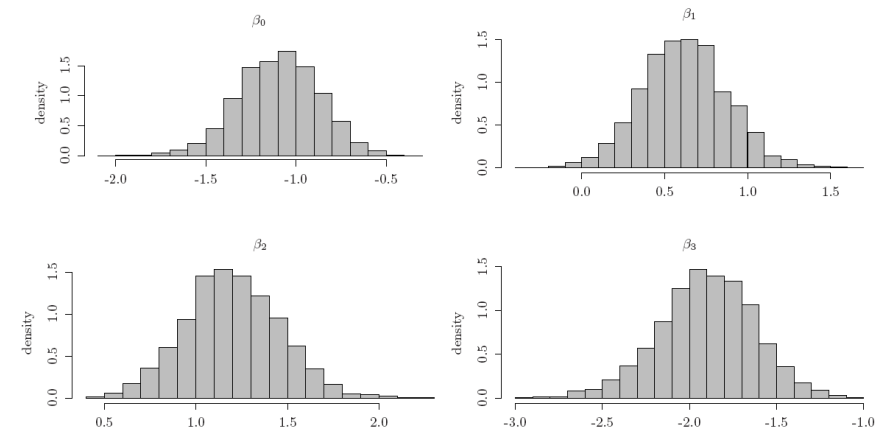
Convergence of cumulative averages $\sum_{\tau=1}^t \beta_j^{(\tau)} / t$ is to a value.

155

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○	○○○○○ ○○○○○●○○○○○○○○○○○ ○○○○○	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Example 3.4: Bayesian probit model (6)



156

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○ ○○○	○ ○○○○ ○○○○○○○○○○○○	○○○○ ○○○○○○○○●○○○○○○○ ○○○	○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Example 3.4: Bayesian probit model (7)

		Posterior mean	95% credible interval	
intercept	β_0	-1.0952	-1.4646	-0.7333
planned	β_1	0.6201	0.2029	1.0413
risk factors	β_2	1.2000	0.7783	1.6296
antibiotics	β_3	-1.8993	-2.3636	-1.471

157

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○ ○○○	○ ○○○○ ○○○○○○○○○○○○	○○○○ ○○○○○○○○○○●○○○○○○○ ○○○	○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Choosing a good proposal distribution

- Ideally: Markov chain with small correlation $\rho(\mathbf{X}^{(t-1)}, \mathbf{X}^{(t)})$ between subsequent values.
 \leadsto fast exploration of the support of the target f .
- Two sources for this correlation:
 - the correlation between the current state $\mathbf{X}^{(t-1)}$ and the newly proposed value $\mathbf{X} \sim q(\cdot | \mathbf{X}^{(t-1)})$
 (can be reduced using a proposal with high variance)
 - the correlation introduced by retaining a value $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$
 because the newly generated value \mathbf{X} has been rejected
 (can be reduced using a proposal with small variance)
- Trade-off for finding the ideal compromise between:
 - fast exploration of the space (good mixing behaviour)
 - obtaining a large probability of acceptance
- For multivariate distributions: covariance of proposal should reflect the covariance structure of the target.

158

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○ ○○○	○ ○○○○ ○○○○○○○○○○○○	○○○○ ○○○○○○○○○○●○○○○○○○ ○○○	○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Example: Choice of proposal (1)

- Target distribution, we want to sample from: $N(0, 1)$ (i.e. $f(\cdot) = \phi_{(0,1)}(\cdot)$)
- We want to use a random walk Metropolis algorithm with

$$\varepsilon \sim N(0, \sigma^2)$$

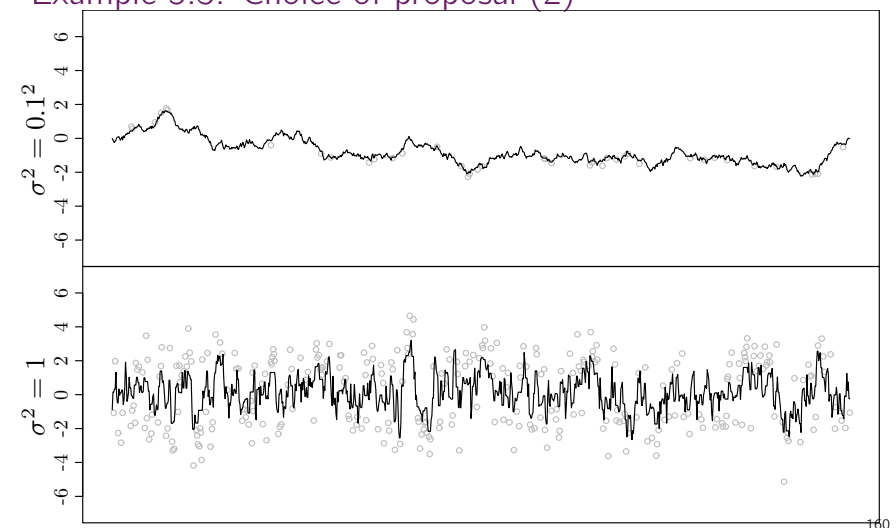
- What is the optimal choice of σ^2 ?
- We consider four choices $\sigma^2 = 0.1^2, 1, 2.38^2, 10^2$.

159

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○ ○○○	○ ○○○○ ○○○○○○○○○○○○	○○○○ ○○○○○○○○○○●○○○○○○○ ○○○	○○○○○○○○ ○○○○○○○ ○○○○○○○

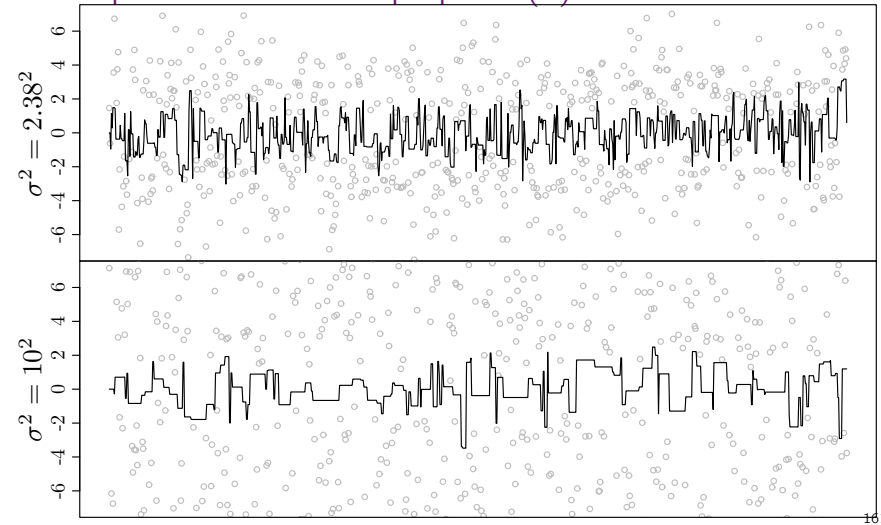
Random-walk Metropolis with Examples

Example 5.3: Choice of proposal (2)



160

Example 5.3: Choice of proposal (3)



161

Example 5.3: Choice of proposal (4)

	Autocorrelation $\rho(X^{(t-1)}, X^{(t)})$		Probability of acceptance $\alpha(X, X^{(t-1)})$	
	Mean	95% CI	Mean	95% CI
$\sigma^2 = 0.1^2$	0.9901	(0.9891, 0.9910)	0.9694	(0.9677, 0.9710)
$\sigma^2 = 1$	0.7733	(0.7676, 0.7791)	0.7038	(0.7014, 0.7061)
$\sigma^2 = 2.38^2$	0.6225	(0.6162, 0.6289)	0.4426	(0.4401, 0.4452)
$\sigma^2 = 10^2$	0.8360	(0.8303, 0.8418)	0.1255	(0.1237, 0.1274)

Suggests: Optimal choice is $2.38^2 > 1$.

162

Example 5.4: Bayesian probit model (revisited)

- So far we used: $\text{Var}(\epsilon) = 0.08 \cdot \mathbb{I}$.
- Better choice: Let $\text{Var}(\epsilon)$ reflect the covariance structure
- Frequentist asymptotic theory: $\text{Var}(\hat{\beta}^{\text{m.l.e}}) = (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$
 \mathbf{D} is a suitable diagonal matrix
- Better choice: $\text{Var}(\epsilon) = 2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$
- Increases rate of acceptance from 13.9% to 20.0% and reduces autocorrelation:

$\Sigma = 0.08 \cdot \mathbf{I}$	β_0	β_1	β_2	β_3
Autocorrelation $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$	0.9496	0.9503	0.9562	0.9532
$\Sigma = 2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}$	β_0	β_1	β_2	β_3
Autocorrelation $\rho(\beta_j^{(t-1)}, \beta_j^{(t)})$	0.8726	0.8765	0.8741	0.8792

(in this example $\det(0.08 \cdot \mathbb{I}) = \det(2 \cdot (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1})$)

163

Positron Emission Tomography I

Positron Emission Tomography

- Inject radioactive tracer into subject's bloodstream.
- Record tracer concentration in blood (at high speed).
- Record numbers of emissions from each volume element (voxel).
- Reconstruct brain activity from measurements.

Compartmental Modelling of Each Voxel Model each site as:

- A system of compartments. . .
- into which tracer flows from the blood
- between which tracer flows
- and from which tracer can flow back into the blood.

164

Positron Emission Tomography II

Consider a linear m -compartment model.

Vector $\mathbf{f}(t)$: element i element corresponds to concentration in compartment i at time t .

Similarly, $\mathbf{b}(t)$ describe all flow into the system from outside.

These models yields a set of ODEs:

$$\begin{aligned}\dot{\mathbf{f}}(t) &= \mathbf{A}\mathbf{f}(t) + \mathbf{b}(t), \\ \mathbf{f}(0) &= \boldsymbol{\xi},\end{aligned}$$

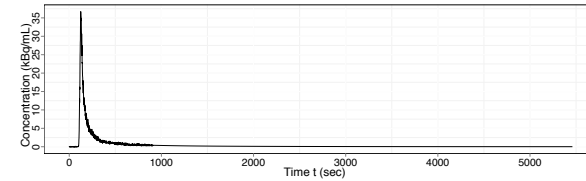
where $\boldsymbol{\xi}$ is the vector of initial concentrations and $\dot{\mathbf{f}}$ denotes the time derivative of \mathbf{f} .

The solution is:

$$\mathbf{f}(t) = e^{\mathbf{A}t}\boldsymbol{\xi} + \int_0^t e^{\mathbf{A}(t-s)}\mathbf{b}(s)ds,$$

Positron Emission Tomography III

We also have the measure *input* signal:



A plasma input model with m tissue compartments:

$$\begin{aligned}\dot{\mathbf{C}}_T(t) &= \mathbf{A}\mathbf{C}_T(t) + \mathbf{b}C_P(t) & \mathbf{C}_T(0) &= \mathbf{0}, \\ \mathbf{C}_T(t) &= \mathbf{1}^T \mathbf{C}_T(t)\end{aligned}$$

where:

Positron Emission Tomography IV

- $\mathbf{C}_T(t)$ are the compartmental “activities”
 - $C_P(t)$ is the input signal
 - \mathbf{A} is an $m \times m$ rate matrix
 - $\mathbf{b} = (K_1, 0, \dots, 0)^T$
 - $\mathbf{1}$ and $\mathbf{0}$ are m -vectors of ones and zeroes.
- The solution to this set of ODES is:

$$\begin{aligned}C_T(t) &= \int_0^t C_P(t-s)H_{TP}(s)ds \\ H_{TP}(t) &= \sum_{i=1}^m \phi_i e^{-\theta_i t},\end{aligned}$$

where the ϕ_i and θ_i parameters are functions of the rate constants.

Positron Emission Tomography V

The macro parameter of interest is the *volume of distribution*,

$$V_D := \int_0^\infty H_{TP}(t)dt = \sum_{i=1}^m \frac{\phi_i}{\theta_i}.$$

Combining this deterministic model with a measurement model:

$$\begin{aligned}C_T(t_j; \phi_{1:m}, \theta_{1:m}) &= \sum_{i=1}^m \phi_i \int_0^{t_j} C_P(s) e^{-\theta_i(t_j-s)} ds \\ y_j &= C_T(t_j; \phi_{1:m}, \theta_{1:m}) + \sqrt{\frac{C_T(t_j; \phi_{1:m}, \theta_{1:m})}{t_j - t_{j-1}}} \varepsilon_j,\end{aligned}$$

With normally-distributed errors, choosing an inverse gamma prior for σ^2 and uniform priors for the other parameters:

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○○○○○○○○○○	○○○○○ ○○○○○ ○○○○○○○○○○○○○○○●●	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Positron Emission Tomography VI

$$p(\phi_{1:m}, \theta_{1:m}, \lambda | \mathbf{y}) \propto \prod_{j=1}^n \sqrt{\lambda} \exp \left\{ -\frac{\lambda}{2} \left[\frac{t_j - t_{j-1}}{C_T(t_j; \phi_{1:m}, \theta_{1:m})} \right]^2 \right\} \\ \times \lambda^{\alpha-1} e^{-\beta\lambda} \times \prod_{i=1}^m I_{[\phi_i^a, \phi_i^b]}(\phi_i) I_{[\theta_i^a, \theta_i^b]}(\theta_i),$$

169

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○○○○○○○○○○	○○○○○ ○○○○○ ○○○○○○○○○○○○○○○●●	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Positron Emission Tomography VII

Some information is available from biology and physics: the following prior distributions are used to encode this information:

$$\begin{aligned} \phi_1 &\sim \mathcal{TN}_{[10^{-5}, .01]}(\cdot; 3 \times 10^{-3}, 10^{-3}) & \theta_1 | \phi_1 &\sim \mathcal{TN}_{[2 \times 10^{-4}, .01]}(\cdot; \frac{\phi_1}{15}, .01) \\ \phi_2 &\sim \mathcal{TN}_{[10^{-5}, .01]}(\cdot; 10^{-3}, 10^{-3}) & \theta_2 | \phi_2, \theta_1 &\sim \mathcal{TN}_{[\theta_1, 6 \times .01]}(\cdot; \frac{\phi_2}{4}, .01) \\ \phi_3 &\sim \mathcal{TN}_{[10^{-5}, .01]}(\cdot; 10^{-3}, 10^{-3}) & \theta_1 | \phi_3, \theta_2 &\sim \mathcal{TN}_{[\theta_2, 6 \times .01]}(\cdot; \phi_3, .01) \\ \lambda &\sim \text{Gamma}(\cdot, 10^{-3}) 10^{-3} \end{aligned}$$

170

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○○○○○○○○○○	○○○○○ ○○○○○ ○○○○○○○○○○○○○○○●●	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Positron Emission Tomography VIII

Algorithmically, a valid procedure is simply, let

$\psi = (\phi_1, \dots, \phi_3, \theta_1, \dots, \theta_3, \lambda)$:

- ➊ Initialize ψ with $\psi^{(0)} = \psi_0$, set $t = 0$. ψ_0 can be any value within the boundaries of the priors.
- ➋ Generate \mathbf{U}_t according to p -dimensional uniform distribution on $\prod_{i=1}^p [-s_i, s_i]$. Where s_i is the step size for ψ_i . Set $\boldsymbol{\eta}_t = \psi^{(t)} + \mathbf{U}_t$.
- ➌ Calculate $r_t = f(\boldsymbol{\eta}_t) / f(\psi^{(t)})$. Generate u_t according to uniform distribution on $[0, 1]$. If $u_t \leq r_t$, Set $\psi^{(t+1)} = \boldsymbol{\eta}_t$, otherwise set $\psi^{(t+1)} = \psi^{(t)}$. Increment t . If $t < N$ for some preset positive integer N , go to step (b), otherwise stop.

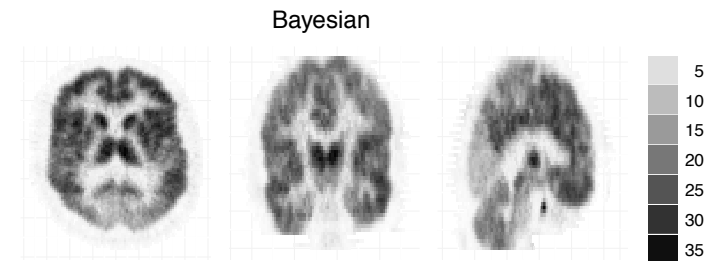
171

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○○○○○○○○○○	○○○○○ ○○○○○ ○○○○○○○○○○○○○○○●●	○○○○○○○○○ ○○○○○○○ ○○○○○○○

Random-walk Metropolis with Examples

Positron Emission Tomography IX

The following estimates of V_D were obtained using MCMC (three slices through the brain volume are shown):



See: Y. Zhou, J. A. D. Aston, and A. M. Johansen. Bayesian model comparison for compartmental models with applications in positron emission tomography. *Journal of Applied Statistics*, 40(5):993–1016, May 2013.

172

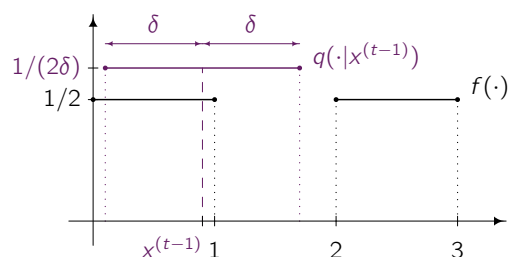
Pathological Example: Reducible Metropolis-Hastings

Consider the target distribution

$$f(x) = (\mathbb{I}_{[0,1]}(x) + \mathbb{I}_{[2,3]}(x))/2.$$

and the proposal distribution $q(\cdot|x^{(t-1)})$:

$$X|X^{(t-1)} = x^{(t-1)} \sim U[x^{(t-1)} - \delta, x^{(t-1)} + \delta]$$



Reducible if $\delta \leq 1$: the chain stays either in $[0, 1]$ or $[2, 3]$.

173

The Metropolis-Adjusted Langevin Algorithm

- Based on the Langevin diffusion:

$$d\mathbf{X}_t = -\frac{1}{2} \nabla \log(f(\mathbf{X}_t)) dt + d\mathbf{B}_t$$

which is f -invariant in continuous time.

- Given target f the MALA proposal mechanism samples:

$$\begin{aligned} \mathbf{X} &\leftarrow \mathbf{X}^{(t-1)} + \epsilon \\ \epsilon &\sim N\left(-\frac{\sigma^2}{2} \nabla \log f(\mathbf{X}^{(t-1)}), \sigma^2 I_p\right) \end{aligned}$$

at time t .

- Accepts X with the usual MH acceptance probability.

174

The Metropolised Independence Sampler

Independent proposals: choose $q(\cdot|x) = q(\cdot)$.

Algorithm 5.3 The Independence Sampler

Starting with $\mathbf{X}^{(0)} := (X_1^{(0)}, \dots, X_p^{(0)})$ iterate for $t = 1, 2, \dots$

1. Draw $\mathbf{X} \sim q(\cdot)$.
2. Compute

$$\alpha(\mathbf{X}|\mathbf{X}^{(t-1)}) = \min \left\{ 1, \frac{f(\mathbf{X}) \cdot q(\mathbf{X}^{(t-1)})}{f(\mathbf{X}^{(t-1)}) \cdot q(\mathbf{X})} \right\}.$$

3. With probability $\alpha(\mathbf{X}|\mathbf{X}^{(t-1)})$ set $\mathbf{X}^{(t)} = \mathbf{X}$, otherwise set $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$.

175

Acceptance Rate

Proposition (Acceptance Rate of Independence Sampler)

If $f(x)/q(x) \leq M < \infty$ the acceptance rate of the independence sampler is at least as high as that of the corresponding rejection sampler.

176

Gibbs Samplers Revisited

What about full conditionals as MH proposals?

- For $\mathbf{X} = (X_1, \dots, X_p)$:
- Consider $q(\mathbf{X}|\mathbf{x}^{(t-1)}) = \delta_{\mathbf{x}^{(t-1)}}(X_{-p})f_{X_p|X_{-p}}(X_p|X_{-p})$.

Remark

A Gibbs sampler step is a special case of the Metropolis-Hastings algorithm.

Part 3— Section 10

Simulated Annealing

Finding the mode of a distribution

- Our objective so far: estimate $\mathbb{E}(h(\mathbf{X}))$.
- A new objective: estimate (global) mode(s) of a distribution:

$$\{\xi : f(\xi) \geq f(\mathbf{x}) \forall \mathbf{x}\}$$

- Naïvely: Choose the $\mathbf{X}^{(t)}$ with maximal density $f(\mathbf{X}^{(t)})$.

Example: Naïvely Finding The Mode of a Normal Density

- Consider $f(\mathbf{x}) = \phi(\mathbf{x})$
- Use a Random Walk proposal $\mathbf{X} \sim N(\mathbf{X}^{(t-1)}, \sigma^2)$ with $\sigma^2 = 0.1^2, 1, 2.38^2, 10^2$.
- Run chains for various T , and pick for each:
 $\mathbf{X}^{\max} = \arg \max_{\mathbf{X} \in (\mathbf{X}^{(t)})_{t=1}^T} f(\mathbf{X})$

$N \sigma^2$	0.1^2	1.0^2	2.38^2	10^2
10	0.906	0.091	0.609	0.623
100	0.315	0.020	-0.063	-0.033
100b	-0.033	0.007	0.065	0.005
1000	0.001	0.001	-0.002	-0.002
1000b	0.015	0.001	-0.001	-0.001

- This approach seems to work here. . .

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○○○○○○○○○○	○○○○○ ○○○○○○○○○○○○○○○○○○○○	○○●○○○○○ ○○○○○○○

Finding the mode of a distribution

More Efficiently Finding the Mode

- Idea: Transform distribution such that it is more concentrated around the mode(s).
- Consider

$$f_{(\beta)}(x) \propto (f(x))^\beta$$
 for very large values of β .
- For $\beta \rightarrow +\infty$ the distribution $f_{(\beta)}(\cdot)$ will be concentrated on the (global) modes.

181

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○○○○○○○○○○	○○○○○ ○○○○○○○○○○○○○○○○○○○○	○○○●○○○○○ ○○○○○○○

Finding the mode of a distribution

Example: Normal distribution (1)

- Consider the $N(\mu, \sigma^2)$ distribution with density

$$f_{(\mu, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- Mode of the $N(\mu, \sigma^2)$ distribution is μ .
- For increasing β the distribution is more and more concentrated around its mode μ , as

$$\begin{aligned} (f_{(\mu, \sigma^2)}(x))^\beta &\propto \left(\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)^\beta \\ &= \exp\left(-\frac{(x-\mu)^2}{2\sigma^2/\beta}\right) \propto f_{(\mu, \sigma^2/\beta)}(x). \end{aligned}$$

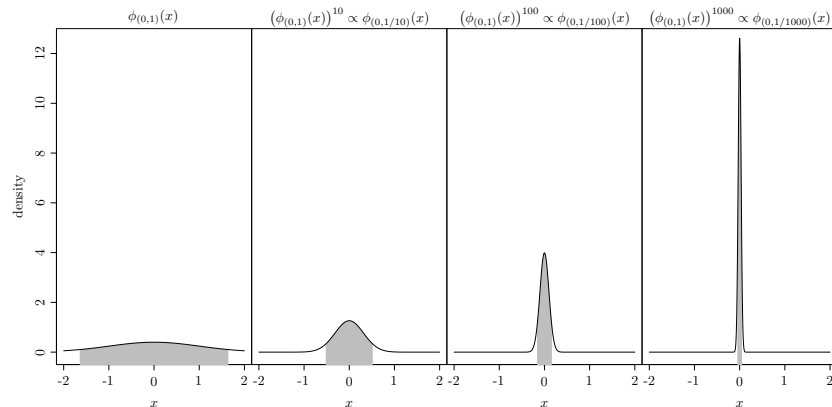
- Increasing β corresponds to reducing the variance.

182

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○○○○○○○○○○	○○○○○ ○○○○○○○○○○○○○○○○○○○○	○○○○○●○○○ ○○○○○○○

Finding the mode of a distribution

Example: Normal distribution (2)

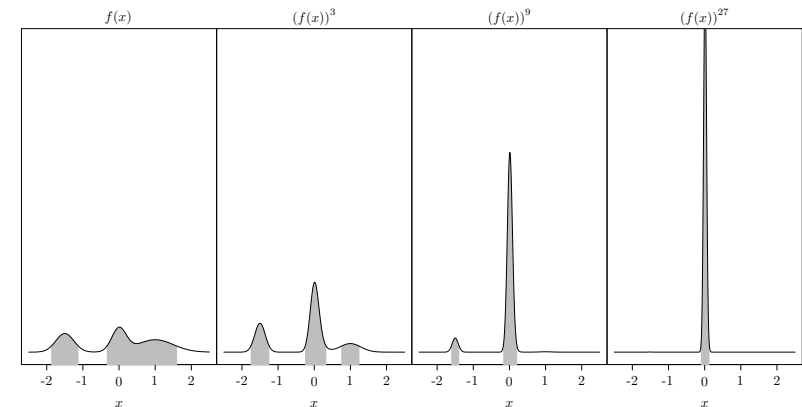


183

Motivation	Gibbs Samplers	Metropolis-Hastings	Simulated Annealing
○○○○○ ○○○○○	○ ○○○○○ ○○○○○○○○○○○○○○	○○○○○ ○○○○○○○○○○○○○○○○○○○○	○○○○○●○○○ ○○○○○○○

Finding the mode of a distribution

Another example



184

Sampling from $f_{(\beta)}(\cdot)$

- We can sample from $f_{(\beta)}(\cdot)$ using a random walk Metropolis algorithm.
- Probability of acceptance becomes

$$\min \left\{ 1, \frac{f_{(\beta)}(\mathbf{X})}{f_{(\beta)}(\mathbf{X}^{(t-1)})} \right\} = \min \left\{ 1, \left(\frac{f(\mathbf{X})}{f(\mathbf{X}^{(t-1)})} \right)^\beta \right\}.$$

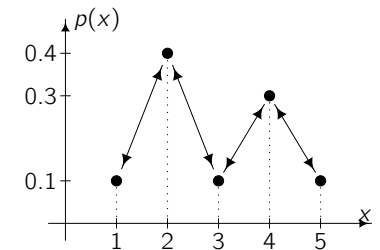
- For $\beta \rightarrow +\infty$ the probability of acceptance converges to ...
 - 1 if $f(\mathbf{X}) \geq f(\mathbf{X}^{(t-1)})$, and
 - 0 if $f(\mathbf{X}) < f(\mathbf{X}^{(t-1)})$.
- For large β the chain $(\mathbf{X}^{(t)})_t$ converges to a local maximum of $f(\cdot)$.
- Whether the chain can escape from local maxima of the density depends on whether it can reach the (global) mode within a single step.

185

Another Example

Assume we want to find the mode of

$$p(x) = \begin{cases} 0.4 & \text{for } x = 2 \\ 0.3 & \text{for } x = 4 \\ 0.1 & \text{for } x = 1, 3, 5. \end{cases}$$



using a random walk Metropolis algorithm that can only move one to the left or one to the right.

For $\beta \rightarrow +\infty$ the probability for accepting a move from 4 to 3 converges to 0, as $p(4) > p(3)$, thus the chain cannot escape from the local maximum at 4.

186

Sampling from $f_{(\beta)}(\cdot)$ is difficult

- For large β the distribution $f_{(\beta)}(\cdot)$ is increasingly concentrated around its modes.
- For large β sampling from $f_{(\beta)}$ gets increasingly difficult.
- Remedy: Start with a small β_0 and let β_t slowly increase.
- The sequence β_t determines whether local extrema are escaped.

187

Simulated Annealing: Minimising an arbitrary function

- More general objective: find global minima of a function $H : E \rightarrow \mathbb{R}_+$.
- Idea: Consider a distribution

$$f(x) \propto \exp(-H(x)) \text{ for } x \in E,$$

yielding

$$f_{(\beta_t)}(x) = (f(x))^{\beta_t} \propto \exp(-\beta_t \cdot H(x)) \text{ for } x \in E.$$

\rightsquigarrow back to the framework of the previous slides.

- In this context β_t is often referred to as *inverse temperature*.

188

Simulated Annealing: Algorithm

Algorithm: Simulated Annealing

Starting with $\mathbf{X}^{(0)} := (X_1^{(0)}, \dots, X_p^{(0)})$ and $\beta^{(0)} > 0$ iterate for $t = 1, 2, \dots$

1. Increase β_{t-1} to β_t .
2. Draw $\mathbf{X} \sim q(\cdot | \mathbf{X}^{(t-1)})$.
3. Compute

$$\alpha(\mathbf{X} | \mathbf{X}^{(t-1)}) = \min \left\{ 1, \exp \left(-\beta_t \left(H(\mathbf{X}) - H(\mathbf{X}^{(t-1)}) \right) \right) \right\} \cdot \frac{q(\mathbf{X}^{(t-1)} | \mathbf{X})}{q(\mathbf{X} | \mathbf{X}^{(t-1)})}.$$

4. With probability $\alpha(\mathbf{X} | \mathbf{X}^{(t-1)})$ set $\mathbf{X}^{(t)} = \mathbf{X}$, otherwise set $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$.

189

Annealing schedules

- As before $\mathbf{X}^{(t)}$ converges for $\beta_t \rightarrow \infty$ to a *local* minimum of $H(\cdot)$.
- Convergence to a *global* minimum depends on annealing schedule:

Logarithmic tempering $\beta_t = \frac{\log(1+t)}{\beta_0}$.

Good theoretical properties; practically irrelevant.

Geometric tempering $\beta_t = \alpha^t \cdot \beta_0$ for some $\alpha > 1$. Popular choice, no theoretical convergence results.

- In practise: expect simulated annealing to find a "good" *local* minimum, but don't expect it to find the *global* minimum!

190

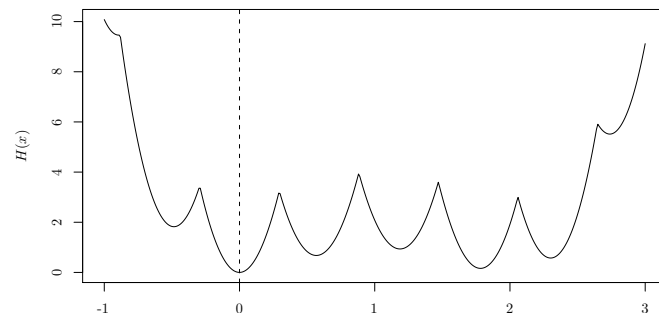
SA Example (1)

Minimise

$$H(x) = ((x-1)^2 - 1)^2 + 3 \cdot s(11.56 \cdot x^2)$$

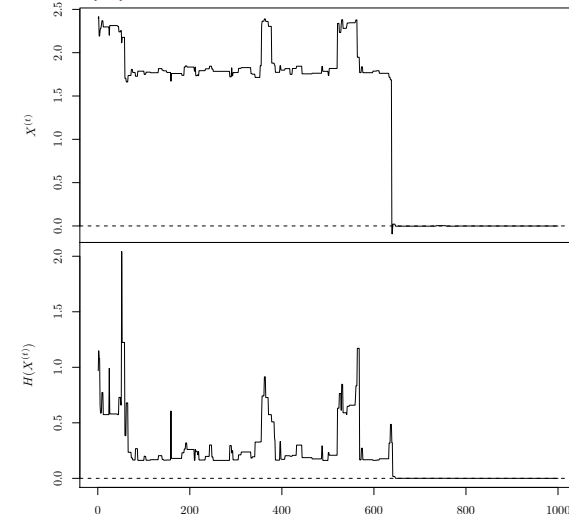
with

$$s(x) = \begin{cases} |x| \bmod 2 & \text{for } 2k \leq |x| \leq 2k+1, k \in \mathbb{N}_0 \\ 2 - |x| \bmod 2 & \text{for } 2k+1 \leq |x| \leq 2(k+1), k \in \mathbb{N}_0 \end{cases}$$



191

SA Example (2)

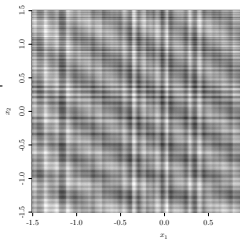


192

A More Challenging Example

- Consider:

$$f(x_1, x_2) = \exp(\sin(50x_1)) + \sin(60 \exp(x_2)) + \sin(70 \sin(x_1)) + \sin(\sin(80x_2)) - \sin(10(x_1 + x_2)) + \frac{1}{4}(x_1^2 + x_2^2)$$



- What is its minimum?
- This question was part of SIAM's 2002 hundred-dollar, hundred-digit challenge (*SIAM News*, Volume 35, Number 1).
- It is on the assessment.

193

Summary of Part 3

- Motivation
- MCMC
- Gibbs Samplers
- Metropolis-Hastings-type Algorithms
- Simulated Annealing

194

Part 4

Augmentation

Augmentation

- "Making the *space* bigger to make the problem easier."
- To target a distribution $f_X(\mathbf{x})$:
 - Construct some $f_{X,Z}(\mathbf{x}, \mathbf{z})$ on $\mathcal{X} \otimes \mathcal{Z}$
 - such that

$$f_X(\mathbf{x}) = \int_{\mathcal{Z}} f_{X,Z}(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$
 - and $f_{X,Z}$ is easy to sample from (when f_X is not).
- Versatile technique with many applications.

196

Composition Sampling

Composition Sampling

- Given a *mixture distribution*,

$$f_X(\mathbf{x}) = \sum_{i=1}^n w_i f_i(\mathbf{x})$$

- Define

$$f_{X,Z}(\mathbf{x}, z) = w_z \cdot f_z(\mathbf{x})$$

on $\mathcal{X} \otimes \{1, \dots, n\}$.

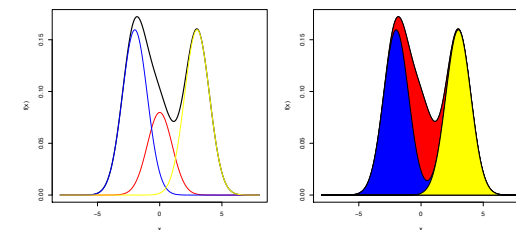
- Sample $Z \sim \sum_{i=1}^n w_i \delta_{\{i\}}(\cdot)$
- Sample $X \sim f_Z(\cdot)$

Normal Mixture: Composition Sampling in Detail I

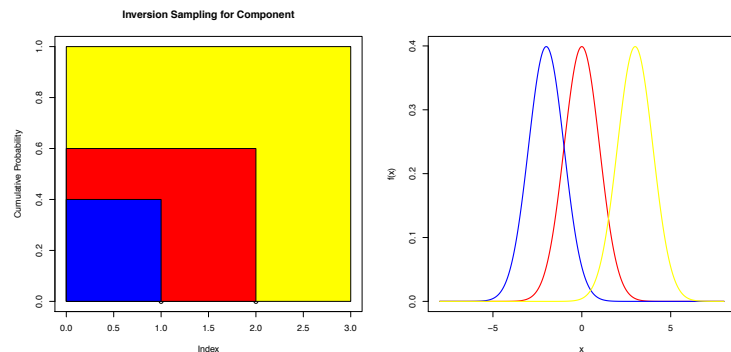
Example of Composition Sampling: Normal Mixture

- For $f(x) = 0.4\mathcal{N}(x; -2, 1) + 0.2\mathcal{N}(x; 0, 1) + 0.4\mathcal{N}(x; 3, 1)$
- Sample $U \sim \mathcal{U}[0, 1]$; set $I = 1$ if $U < 0.4$, $I = 2$ if $U \in [0.4, 0.6)$, $I = 3$ otherwise.
- Sample $X \sim f_I$ where $f_I = \mathcal{N}(\mu_I, 1)$ and $\mu = \{-2, 0, 3\}$.

Normal Mixture: Composition Sampling in Detail II



Normal Mixture: Composition Sampling in Detail III



201

Part 4— Section 12

Rejection Revisited

A Generic Augmentation Scheme

- Given any density $f(\mathbf{x})$, define

$$\bar{f}(\mathbf{x}, u) := f(\mathbf{x}) \cdot \bar{f}_{U|\mathbf{x}}(u|\mathbf{x})$$

- with

$$\bar{f}_{U|\mathbf{x}}(u|\mathbf{x}) = \frac{1}{f(\mathbf{x})} \mathbb{I}_{[0, f(\mathbf{x})]}(u)$$

- Then

$$\bar{f}(\mathbf{x}, u) = \mathbb{I}_{[0, f(\mathbf{x})]}(u).$$

203

Rejection Sampling Revisited

Proposition (Rejection Sampling Equivalence)

- Given $f(\mathbf{x})$, define

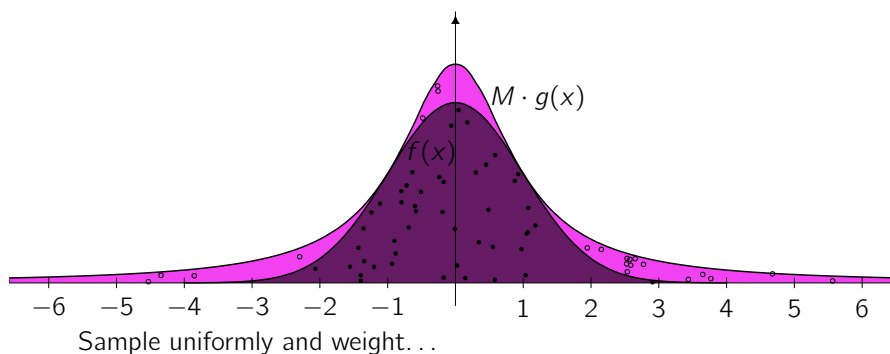
$$\bar{f}(\mathbf{x}, u) = \mathbb{I}_{[0, f(\mathbf{x})]}(u).$$

- Given proposal $g(\mathbf{x})$ and $M \geq \sup_{\mathbf{x}} f(\mathbf{x})/g(\mathbf{x})$, define

$$\bar{g}(\mathbf{x}, u) = \frac{1}{M} \mathbb{I}_{[0, M \cdot g(\mathbf{x})]}.$$

- Let $w(\mathbf{x}, u) = c \cdot \bar{f}(\mathbf{x}, u) / \bar{g}(\mathbf{x}, u)$
- The associated self-normalised importance sampling estimator of $\mathbb{E}_{\bar{f}}[\varphi(\mathbf{X})] \equiv \mathbb{E}_f[\varphi(\mathbf{X})]$ is the rejection sampling estimator.

204



Slice Sampling

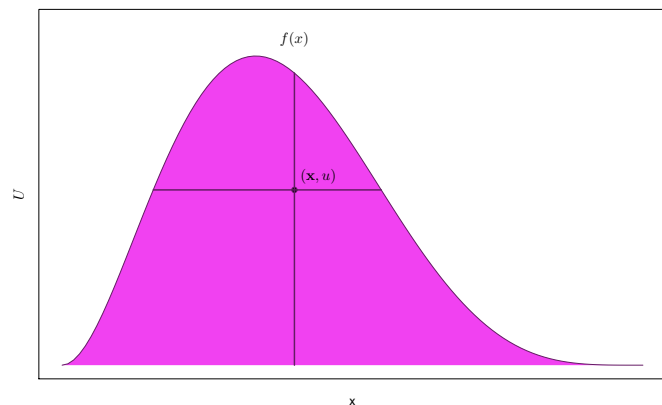
- Rejection sampling can be viewed as importance sampling with an extended target distribution...
- so can we apply other algorithms to that extended distribution?

Algorithm: The Slice Sampler

Starting with $(\mathbf{X}^{(0)}, U^{(0)})$ iterate for $t = 1, 2, \dots$

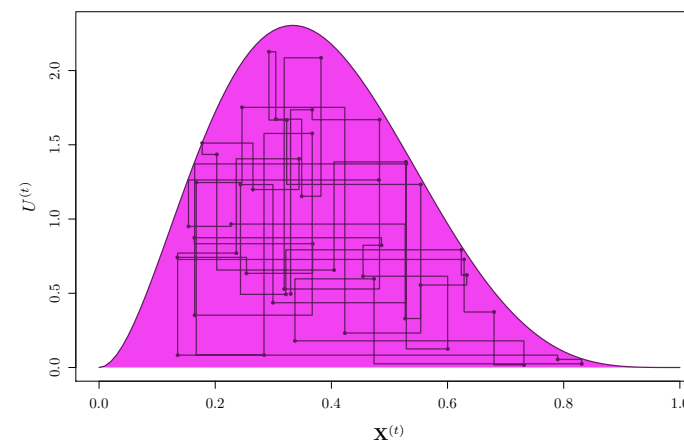
- 1 Draw $\mathbf{X}^{(t)} \sim \bar{F}_{\mathbf{X}|U}(\cdot | U^{(t-1)})$.
- 2 Draw $U^{(t)} \sim \bar{F}_{U|\mathbf{X}}(\cdot | \mathbf{X}^{(t)})$.

An Illustration of the Conditional Distributions



A Slice-Sampler Trajectory

Example: Sampling from a $\text{Beta}(3, 5)$ distribution



How Practical Is This?

- Sampling $U \sim U[0, f(\mathbf{X})]$ is *easy*.
- Sampling $\mathbf{X} \sim U(L(U))$ where

$$L(u) := \{\mathbf{x} : f(\mathbf{x}) \geq u\}$$

can be easy. . .

- but it might not be.
- Consider the bivariate density:

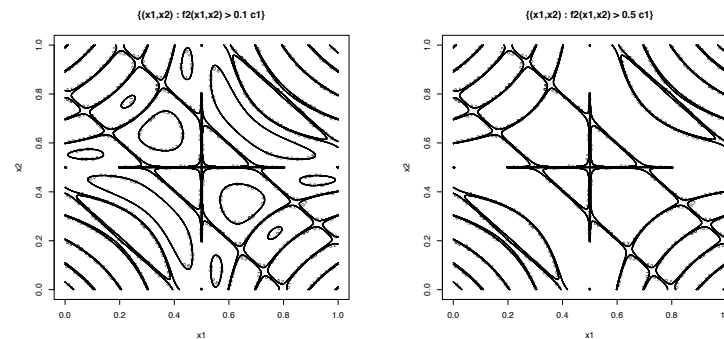
$$f_2(x_1, x_2) = c_1 \cdot \sin^2(x_1 \cdot x_2) \cdot \cos^2(x_1 + x_2) \cdot \exp\left(-\frac{1}{2}(|x_1| + |x_2|)\right)$$

209

The Trouble with Slice Sampling

Level sets of:

$$f_2(x_1, x_2) = c_1 \cdot \sin^2(x_1 \cdot x_2) \cdot \cos^2(x_1 + x_2) \cdot \exp\left(-\frac{1}{2}(|x_1| + |x_2|)\right)$$



here we could use rejection.

210

Algorithm: The Co-ordinate-wise Slice Sampler

Starting with $(X_1^{(0)}, \dots, X_p^{(0)}, U^{(0)})$ iterate for $t = 1, 2, \dots$

1. Draw $X_1^{(t)} \sim \bar{f}_{X_1|X_{-1}, U}(\cdot | X_{-1}^{(t-1)}, U^{(t-1)})$.
2. Draw $X_2^{(t)} \sim \bar{f}_{X_2|X_{-2}, U}(\cdot | X_1^{(t)}, X_3^{(t-1)}, \dots, X_p^{(t-1)}, U^{(t-1)})$.
- \vdots
- p. Draw $X_p^{(t)} \sim \bar{f}_{X_p|X_{-p}, U}(\cdot | X_{-p}^{(t)}, U^{(t-1)})$.
- p+1. Draw $U^{(t)} \sim \bar{f}_{U|X}(\cdot | \mathbf{X}^{(t)})$.

211

Algorithm: The Metropolised Slice Sampler

Starting with $(\mathbf{X}^{(0)}, U^{(0)})$ iterate for $t = 1, 2, \dots$

1. Draw $\mathbf{X} \sim \bar{q}(\cdot | \mathbf{X}^{(t-1)}, U^{(t-1)})$.
2. With probability

$$\min \left(1, \frac{\bar{f}(\mathbf{X}, U^{(t-1)}) q(\mathbf{X}^{(t-1)} | \mathbf{X}, U^{(t-1)})}{\bar{f}(\mathbf{X}^{(t-1)}, U^{(t-1)}) q(\mathbf{X} | \mathbf{X}^{(t-1)}, U^{(t-1)})} \right)$$

accept and set $\mathbf{X}^{(t)} = \mathbf{X}$.
Otherwise, set $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}$.

2. Draw $U^{(t)} \sim \bar{f}_{U|X}(\cdot | \mathbf{X}^{(t)})$.

212

Data Augmentation

Data Augmentation I

- *Latent variable models* are common: statistical models with:
 - parameters θ ,
 - observations y , and
 - latent variables, z .
- Typically, the joint distribution, $f_{y,z,\theta}$, is known,
- but integrating out the latent variables is not feasible.
- Without $f_{y,\theta}$ we can't implement an MCMC algorithm targetting $f_{\theta|y}$.
- The basis of data augmentation is to *augment* θ with z and to run an MCMC algorithm which targets $f_{\theta,z|y}$.
- This distribution has the correct marginal in θ .

214

Data Augmentation and Gibbs Samplers

- Gibbs sampling is only feasible when we can sample easily from the full conditionals.
- A technique that can help achieving full conditionals that are easy to sample from is *demarginalisation*:
Introduce a set of auxiliary random variables Z_1, \dots, Z_r such that f is the marginal density of $(X_1, \dots, X_p, Z_1, \dots, Z_r)$, i.e.

$$f(x_1, \dots, x_p) = \int f(x_1, \dots, x_p, z_1, \dots, z_r) d(z_1, \dots, z_r).$$

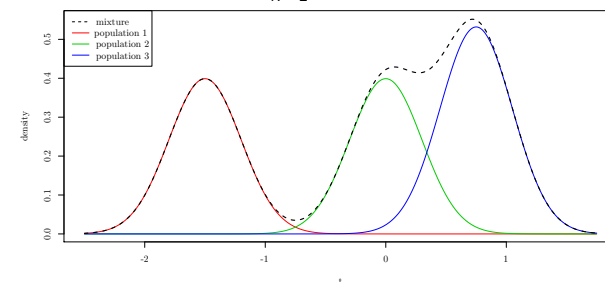
- In many cases there is a “natural choice” of the *completion* (Z_1, \dots, Z_r) .

215

Example: Mixture of Gaussians — Model

Consider the following K population mixture model for data Y_1, \dots, Y_n :

$$f(y_i) = \sum_{k=1}^K \pi_k \phi(\mu_k, 1/\tau)(y_i)$$



Objective: Bayesian inference for the parameters $(\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K)$.

216

Example: Mixture of Gaussians — Priors

- The number of components K is assumed to be known.
- The variance parameter τ is assumed to be known.
- $(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$, i.e.

$$f_{(\alpha_1, \dots, \alpha_K)}(\pi_1, \dots, \pi_K) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

- $(\mu_1, \dots, \mu_K) \sim \mathcal{N}(\mu_0, 1/\tau_0)$, i.e.

$$f_{(\mu_0, \tau_0)}(\mu_k) \propto \exp(-\tau_0(\mu_k - \mu_0)^2/2)$$

Example: Mixture of Gaussians — Joint distribution

$$f(\mu_1, \dots, \mu_K, \pi_1, \dots, \pi_K, y_1, \dots, y_n) \propto \left(\prod_{k=1}^K \pi_k^{\alpha_k - 1} \right) \cdot \left(\prod_{k=1}^K \exp(-\tau_0(\mu_k - \mu_0)^2/2) \right) \cdot \left(\prod_{i=1}^n \sum_{k=1}^K \pi_k \exp(-\tau(y_i - \mu_k)^2/2) \right)$$

The full conditionals do not seem to come from “nice” distributions.

Use data augmentation: include auxiliary variables Z_1, \dots, Z_n which indicate which population the i -th individual is from, i.e.

$$\mathbb{P}(Z_i = k) = \pi_k \quad \text{and} \quad Y_i | Z_i = k \sim \mathcal{N}(\mu_k, 1/\tau).$$

The marginal distribution of Y is as before, so Z_1, \dots, Z_n are indeed a completion.

Example: Mixture of Gaussians — Joint distribution

The joint distribution of the augmented system is

$$\begin{aligned} & f(y_1, \dots, y_n, z_1, \dots, z_n, \mu_1, \dots, \mu_K, \pi_1, \dots, \pi_K) \\ & \propto \left(\prod_{k=1}^K \pi_k^{\alpha_k - 1} \right) \cdot \left(\prod_{k=1}^K \exp(-\tau_0(\mu_k - \mu_0)^2/2) \right) \\ & \quad \cdot \left(\prod_{i=1}^n \pi_{z_i} \exp(-\tau(y_i - \mu_{z_i})^2/2) \right) \end{aligned}$$

The full conditionals now come from “nice” distributions.

Example: Mixture of Gaussians — Full conditionals

$$\begin{aligned} & \mathbb{P}(Z_i = k | Y_1, \dots, Y_n, \mu_1, \dots, \mu_K, \pi_1, \dots, \pi_K) \\ & = \frac{\pi_k \phi(\mu_k, 1/\tau)(y_i)}{\sum_{l=1}^K \pi_l \phi(\mu_l, 1/\tau)(y_i)} \\ & \mu_k | Y_1, \dots, Y_n, z_1, \dots, z_n, \pi_1, \dots, \pi_K \\ & \sim \mathcal{N} \left(\frac{\tau (\sum_{i: Z_i = k} Y_i) + \tau_0 \mu_0}{|\{i: Z_i = k\}| \tau + \tau_0}, \frac{1}{|\{i: Z_i = k\}| \tau + \tau_0} \right) \end{aligned}$$

$$\begin{aligned} & \pi_1, \dots, \pi_K | Y_1, \dots, Y_n, z_1, \dots, z_n, \mu_1, \dots, \mu_K \\ & \sim \text{Dirichlet}(\alpha_1 + |\{i: Z_i = 1\}|, \dots, \alpha_K + |\{i: Z_i = K\}|). \end{aligned}$$

Example

Example: Mixture of Gaussians — Gibbs sampler

Starting with initial values $\mu_1^{(0)}, \dots, \mu_K^{(0)}, \pi_1^{(0)}, \dots, \pi_K^{(0)}$ iterate the following steps for $t = 1, 2, \dots$

1. For $i = 1, \dots, n$:

Draw $Z_i^{(t)}$ from the discrete distribution on $\{1, \dots, K\}$ specified by

$$p(Z_i^{(t)}) = \left(\frac{\pi_k \phi(\mu_k^{(t-1)}, 1/\tau)(y_i)}{\sum_{\ell=1}^K \pi_\ell^{(t-1)} \phi(\mu_\ell^{(t-1)}, 1/\tau)(y_i)} \right).$$

2. For $k = 1, \dots, K$:

Draw

$$\mu_k^{(t)} \sim N \left(\frac{\tau \left(\sum_{i: Z_i^{(t)}=k} y_i \right) + \tau_0 \mu_0}{|\{i: Z_i^{(t)}=k\}| \tau + \tau_0}, \frac{1}{|\{i: Z_i^{(t)}=k\}| \tau + \tau_0} \right).$$

3. Draw

$$(\pi_1^{(t)}, \dots, \pi_K^{(t)}) \sim \text{Dirichlet} \left(\alpha_1 + |\{i: Z_i^{(t)}=1\}|, \dots, \alpha_K + |\{i: Z_i^{(t)}=K\}| \right).$$

221

Part 4— Section 14

Recent Innovations

Approximate Bayesian Computation

Bayesian Computation (Towards ABC)

- Consider a target distribution $f_{X|Y}(x|y)$ written as:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}.$$

- If both $f_X(x)$ and $f_{Y|X}(y|x)$ can be evaluated we're done.
- If we *cannot* evaluate $f_{Y|X}$ even pointwise, then we *can't* directly use the techniques which we've described previously.
- Consider the case in which Y is *discrete*.
- We can invoke a clever data augmentation trick which requires only that we can *sample* from $f_{Y|X}$.

223

Approximate Bayesian Computation

- We can define an extended distribution:

$$f_{X,Z|Y}(x, z|y) \propto f_{Y|X}(z|x) f_X(x) \delta_{y,z}$$

and note that it has, as a marginal distribution, our target:

$$\sum_z f_{X,Z|Y}(x, z|y) \propto \sum_z f_{Y|X}(z|x) f_X(x) \delta_{y,z} = f_{Y|X}(y|x) f_X(x).$$

- We can sample $(X, Z) \sim f_{Y|X}(z|x) f_X(x)$ and use this as a rejection sampling proposal for our target distribution, keeping samples with probability proportional to

$$f_{X,Z|Y}(x, z|y) / f_{Y|X}(z|x) f_X(x) \propto \delta_{y,z}$$

224

Approximate Bayesian Computation

- When data is not discrete / takes many values, exact matches have no or negligible probability.
- Instead, we keep samples for which $\|z - y\| \leq \epsilon$.
- This leads to a *different* target distribution:

$$f_{X,Z|Y}^{ABC}(x, y|z) \propto f_{Y|X}(z|x) f_X(x) \mathbb{I}_{B(y, \epsilon)}(z)$$

where $B(y, \epsilon) := \{x : |x - y| \leq \epsilon\}$, so

$$\begin{aligned} f_{x|Y}^{ABC} &\propto \int f_{Y|X}(z|x) f_X(x) \mathbb{I}_{B(y, \epsilon)}(z) dz \\ &\propto \int f_{Y|X}(z|x) \mathbb{I}_{B(y, \epsilon)}(z) dz f_X(x) \\ &\propto \int_{z \in B(y, \epsilon)} f_{Y|X}(z|x) dz f_X(x) \end{aligned}$$

this approximation amounts to a *smoothing* of the likelihood function.

225

Even More Approximate Bayesian Computation

- Often a further approximation is introduced by considering not the data itself but some low dimensional summary of the data: This leads to a *different* target distribution:

$$f_{X,Z|Y}^{ABC}(x, z|y) \propto f_{Y|X}(z|x) f_X(x) \mathbb{I}_{B(s(y), \epsilon)}(s(z))$$

- Unless the summary is a sufficient statistic (which it probably isn't) this introduces a difficult to understand approximation.
- Be very careful.

Pseudomarginal methods can also be considered as augmentation techniques, but we don't have enough time to do that here.

226

Part 5

Theory and Practice

Part 5— Section 15

Theoretical Considerations and Convergence Results

Some reassurance about Gibbs Samplers

Definition (Positivity condition)

A distribution with density $f(x_1, \dots, x_p)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the positivity condition if $f(x_1, \dots, x_p) > 0$ for all x_1, \dots, x_p with $f_{X_i}(x_i) > 0$.

Theorem (Hammersley-Clifford)

Let (X_1, \dots, X_p) satisfy the positivity condition and have joint density $f(x_1, \dots, x_p)$. Then for all $(\xi_1, \dots, \xi_p) \in \text{supp}(f)$

$$f(x_1, \dots, x_p) \propto \prod_{j=1}^p \frac{f_{X_j|X_{-j}}(x_j|x_1, \dots, x_{j-1}, \xi_{j+1}, \dots, \xi_p)}{f_{X_j|X_{-j}}(\xi_j|x_1, \dots, x_{j-1}, \xi_{j+1}, \dots, \xi_p)}$$

229

A Cautionary Example

Note the theorem does *not* guarantee the existence of a joint distribution for every set of “full conditionals”!

- Consider the following “model”

$$X_1|X_2 \sim \text{Expo}(\lambda X_2)$$

$$X_2|X_1 \sim \text{Expo}(\lambda X_1),$$

- Trying to apply the Hammersley-Clifford theorem, we obtain

$$\begin{aligned} f(x_1, x_2) &\propto \frac{f_{X_1|X_2}(x_1|\xi_2) \cdot f_{X_2|X_1}(x_2|x_1)}{f_{X_1|X_2}(\xi_1|\xi_2) \cdot f_{X_2|X_1}(\xi_2|\xi_1)} \\ &\propto \exp(-\lambda x_1 x_2) \end{aligned}$$

- $\int \int \exp(-\lambda x_1 x_2) dx_1 dx_2 = +\infty$
 \leadsto joint density cannot be normalised.
- There is no joint density with the above full conditionals.

230

Irreducibility and recurrence of Gibbs Samplers

Proposition

If the joint distribution $f(x_1, \dots, x_p)$ satisfies the positivity condition, the Gibbs sampler yields an f -irreducible, recurrent Markov chain.

Outline Proof

Given an \mathcal{X} such that $\int_{\mathcal{X}} f(x_1^{(t)}, \dots, x_p^{(t)}) d(x_1^{(t)}, \dots, x_p^{(t)}) > 0$.

$$\int_{\mathcal{X}} K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) d\mathbf{x}^{(t)} = \int_{\mathcal{X}} \underbrace{f_{X_1|X_{-1}}(x_1^{(t)}|x_2^{(t-1)}, \dots, x_p^{(t-1)}) \dots}_{>0} \underbrace{f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)}, \dots, x_{p-1}^{(t)})}_{>0} d\mathbf{x}^{(t)}$$

231

Ergodic theorem

Theorem (Ergodicity of the Gibbs Sampler)

If the Markov chain generated by the Gibbs sampler is irreducible and recurrent (which is e.g. the case when the positivity condition holds), then for any integrable function $\varphi : E \rightarrow \mathbb{R}$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \varphi(\mathbf{X}^{(t)}) \rightarrow \mathbb{E}_f(\varphi(\mathbf{X}))$$

for almost every starting value $\mathbf{X}^{(0)}$.

Thus we can approximate expectations $\mathbb{E}_f(\varphi(\mathbf{X}))$ by their empirical counterparts using a *single* Markov chain.

232

A Simple Example

- Consider

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$$

- Associated marginal distributions

$$X_1 \sim N(\mu_1, \sigma_1^2)$$

$$X_2 \sim N(\mu_2, \sigma_2^2)$$

- Associated full conditionals

$$X_1 | X_2 = x_2 \sim N(\mu_1 + \sigma_{12}/\sigma_2^2(x_2 - \mu_2), \sigma_1^2 - (\sigma_{12})^2/\sigma_2^2)$$

$$X_2 | X_1 = x_1 \sim N(\mu_2 + \sigma_{12}/\sigma_1^2(x_1 - \mu_1), \sigma_2^2 - (\sigma_{12})^2/\sigma_1^2)$$

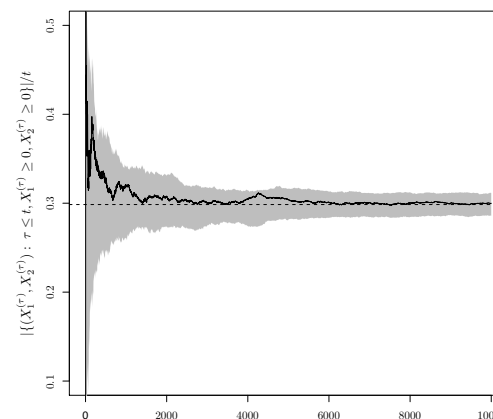
- Gibbs sampler consists of iterating for $t = 1, 2, \dots$

1. Draw $X_1^{(t)} \sim N(\mu_1 + \sigma_{12}/\sigma_2^2(X_2^{(t-1)} - \mu_2), \sigma_1^2 - (\sigma_{12})^2/\sigma_2^2)$

2. Draw $X_2^{(t)} \sim N(\mu_2 + \sigma_{12}/\sigma_1^2(X_1^{(t)} - \mu_1), \sigma_2^2 - (\sigma_{12})^2/\sigma_1^2)$

233

Using the ergodic theorem we can estimate $\mathbb{P}(X_1 \geq 0, X_2 \geq 0)$ by the proportion of samples $(X_1^{(t)}, X_2^{(t)})$ with $X_1^{(t)} \geq 0$ and $X_2^{(t)} \geq 0$:



234

Theoretical properties of Metropolis-Hastings

- The Markov chain $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$ is (strongly) irreducible if $q(\mathbf{x}|\mathbf{x}^{(t-1)}) > 0$ for all $\mathbf{x}, \mathbf{x}^{(t-1)} \in \text{supp}(f)$.
(see, e.g., (see Roberts & Tweedie, 1996) for weaker conditions)
- Such a chain is recurrent if it is irreducible.
(see e.g. Tierney, 1994)
- The chain is aperiodic if there is positive probability that the chain remains in the current state, i.e. $\mathbb{P}(\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)}) > 0$ (for a suitable group of “current states”).

235

Theorem (A Simple Ergodic Theorem)

If $(X_i)_{i \in \mathbb{N}}$ is an f -irreducible, f -invariant, recurrent \mathbb{R}^d -valued Markov chain then the following strong law of large numbers holds for any integrable function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^t \varphi(\xi_i) \stackrel{\text{a.s.}}{=} \int \varphi(x) f(x) dx.$$

for almost every starting value x .

236

Theorem (A Central Limit Theorem)

Under technical regularity conditions the following CLT holds for a recurrent, f -invariant Markov chain, and a function $\varphi : E \rightarrow \mathbb{R}$ which has at least two finite moments:

$$\lim_{t \rightarrow \infty} \sqrt{t} \left[\frac{1}{t} \sum_{i=1}^t \varphi(\xi_i) - \int \varphi(x) \mu(x) dx \right] \stackrel{\mathcal{D}}{=} N(0, \sigma^2(\varphi)),$$

$$\sigma^2(\varphi) = \mathbb{E}[(f(\xi_1) - \bar{\varphi})^2] + 2 \sum_{k=2}^{\infty} \mathbb{E}[(\varphi(\xi_1) - \bar{\varphi})(\varphi(\xi_k) - \bar{\varphi})],$$

where $\bar{\varphi} = \int \varphi(x) f(x) dx$.

237

Optimal Scaling

Much effort has gone into determining optimal scaling rules:

Diffusion Limits Under strong assumptions:

$$\lim_{p \rightarrow \infty} \frac{X_1^{(\lfloor tp \rfloor)}}{\sqrt{p}} \xrightarrow{d} \text{Diffusion}$$

where p is *dimension* and the *speed* of the diffusion depends upon proposal scale.

ESJD Seek to maximise:

$$\int f(x) K(x, y; \theta) (y - x)^2 dx dy$$

Rule of Thumb Optimal RWM Scaling depends upon dimension:

$p = 1$ Acceptance rate of around 0.44.

$p \geq 5$ Acceptance rate of around 0.234.

238

The Metropolis-Adjusted Langevin Algorithm

- Based on the Langevin diffusion:

$$d\mathbf{X}_t = \frac{1}{2} \nabla \log f(\mathbf{X}_t) dt + d\mathbf{B}_t$$

which is f -invariant in continuous time.

- Given target f the MALA proposal proposes:

$$\mathbf{X} \leftarrow \mathbf{X}^{(t-1)} + \epsilon$$

$$\epsilon \sim N\left(\frac{\sigma^2}{2} \nabla \log f(\mathbf{X}^{(t-1)}), \sigma^2 I_p\right)$$

at time t .

- Accepts X with the usual MH acceptance probability.
- Optimal acceptance rate (under similar strong conditions) now 0.574.

239

MALA Example: Normal (1)

Target $f(x) = N(0, 1)$

Proposal

$$q(X^{(t-1)}, X) = N\left(X^{(t-1)} - \frac{\sigma^2 X^{(t-1)}}{2}, \sigma^2\right)$$

Acceptance Probability

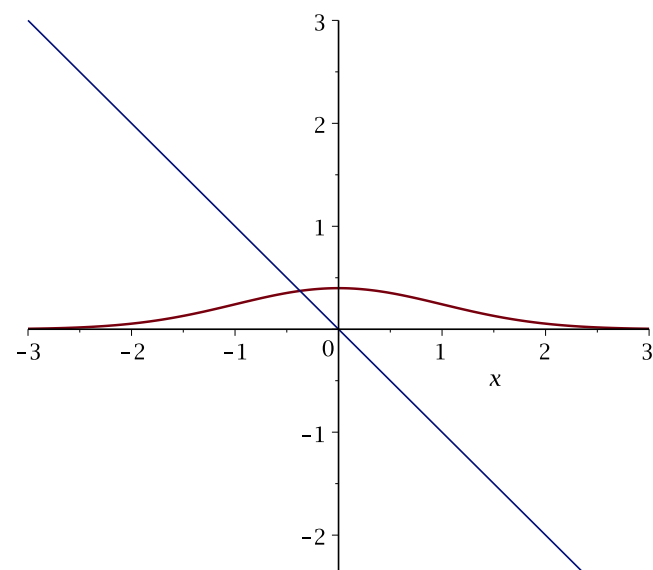
$$\begin{aligned} \alpha(X^{(t-1)}, X) &= 1 \wedge \frac{f(X)}{f(X^{(t-1)})} \frac{q(X, X^{(t-1)})}{q(X^{(t-1)}, X)} \\ &= 1 \wedge \exp\left(\frac{1}{2} \left[(X^{(t-1)})^2 - X^2 \right]\right) \times \\ &\quad \exp\left(\frac{1}{2\sigma^2} \left[\left\{ X - \mu(X^{(t-1)}) \right\}^2 - \left\{ X^{(t-1)} - \mu(X) \right\}^2 \right]\right) \end{aligned}$$

where $\mu(x) := x - \frac{x\sigma^2}{2}$.

240

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○●○○○	○○○○○○○○○○		○○○
	○○○○○○○		○○○○

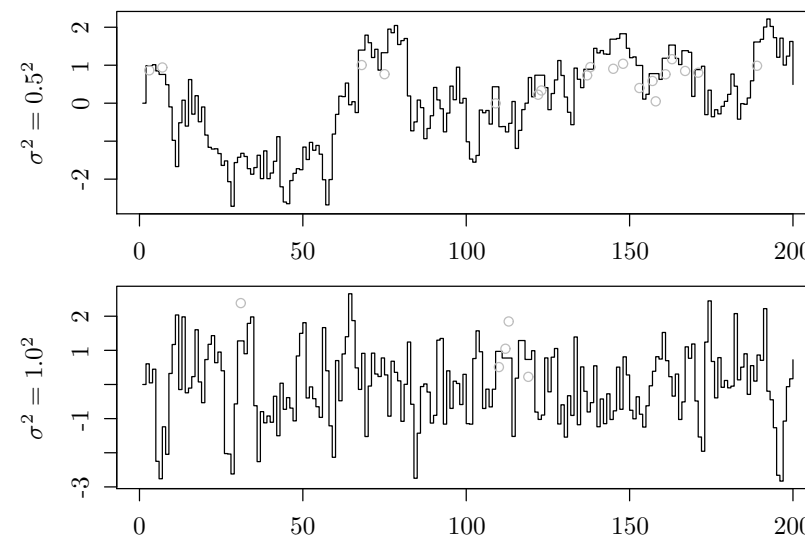
Scaling of Proposal Distributions



241

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○●○○○	○○○○○○○○○○		○○○
	○○○○○○○		○○○○

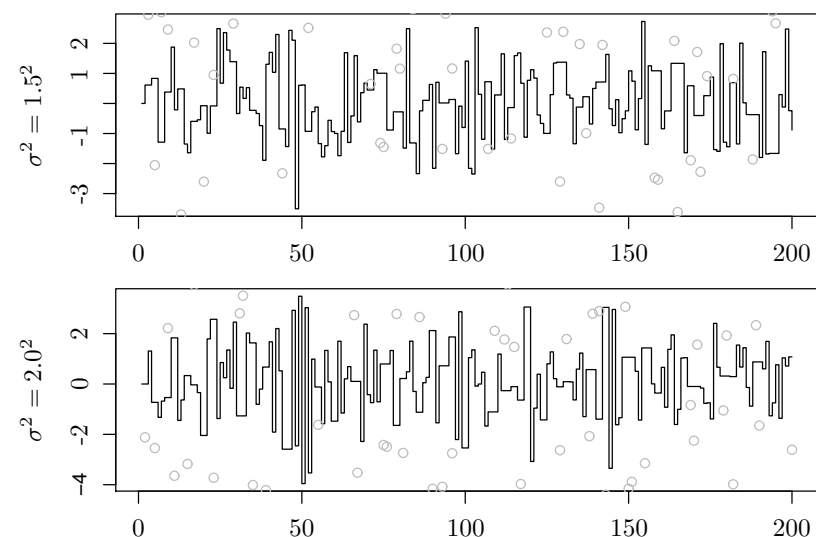
Scaling of Proposal Distributions



242

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○●○○○	○○○○○○○○○○		○○○
	○○○○○○○		○○○○

Scaling of Proposal Distributions



243

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○●○○○	○○○○○○○○○○		○○○
	○○○○○○○		○○○○

Scaling of Proposal Distributions

MALA Example: Normal (2)

RWM	Autocorrelation $\rho(X^{(t-1)}, X^{(t)})$	Probability of acceptance $\alpha(X, X^{(t-1)})$	ESJD
$\sigma^2 = 0.1^2$	0.9901	0.9694	0.010
$\sigma^2 = 1$	0.7733	0.7038	0.448
$\sigma^2 = 2.38^2$	0.6225	0.4426	0.742
$\sigma^2 = 10^2$	0.8360	0.1255	0.337

MALA	Autocorrelation $\rho(X^{(t-1)}, X^{(t)})$	Probability of acceptance $\alpha(X, X^{(t-1)})$	ESJD
$\sigma^2 = 0.5^2$	0.898	0.877	0.246
$\sigma^2 = 1$	0.492	0.961	1.293
$\sigma^2 = 1.5^2$	0.047	0.774	2.137
$\sigma^2 = 2.0^2$	0.011	0.631	4.119

244

Part 5— Section 16

Convergence Diagnostics

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo ooo ooooo	●oooo oooooooooooo ooooo	ooooo	oo ooo oooo

Motivation: The Need for Convergence Diagnostics

The need for convergence diagnostics

- Theory guarantees (under certain conditions) the convergence of the Markov chain $\mathbf{X}^{(t)}$ to the desired distribution.
- This does not imply that a *finite* sample from such a chain yields a good approximation to the target distribution.
- Validity of the approximation must be confirmed in practice.
- Convergence diagnostics help answering this question.
- Convergence diagnostics are *not* perfect and should be treated with a good amount of scepticism.

246

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo ooo ooooo	●oooo oooooooooooo ooooo	ooooo	oo ooo oooo

Motivation: The Need for Convergence Diagnostics

Different diagnostic tasks

Convergence to the target distribution Does $\mathbf{X}^{(t)}$ yield a sample from the target distribution?

- Has reached $(\mathbf{X}^{(t)})_t$ a stationary regime?
- Does $(\mathbf{X}^{(t)})_t$ cover the support of the target distribution?

Convergence of averages Is $\sum_{t=1}^T \varphi(\mathbf{X}^{(t)})/T \approx \mathbb{E}_f(\varphi(\mathbf{X}))$?

Comparison to i.i.d. sampling How much information is contained in the sample from the Markov chain compared to an i.i.d. sample?

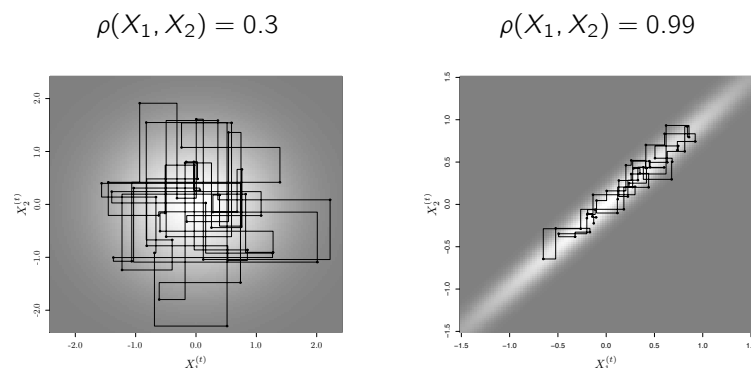
247

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo ooo ooooo	●oooo oooooooooooo ooooo	ooooo	oo ooo oooo

Motivation: The Need for Convergence Diagnostics

Pathological example 1: potentially slowly mixing

Gibbs sampler from a bivariate Gaussian with correlation $\rho(X_1, X_2)$



For correlations $\rho(X_1, X_2)$ close to ± 1 the chain can be poorly mixing

248

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○●○	○○○○○	○○
○○○	○○○○○○○○○○		○○○
○○○○○○○	○○○○○○○		○○○○

Motivation: The Need for Convergence Diagnostics

Pathological example 2: no central limit theorem

The following MCMC algorithm has the $\text{Beta}(\alpha, 1)$ distribution as stationary distribution:

Starting with any $X^{(0)}$ iterate for $t = 1, 2, \dots$

1. With probability $1 - X^{(t-1)}$, set $X^{(t)} = X^{(t-1)}$.
2. Otherwise draw $X^{(t)} \sim \text{Beta}(\alpha + 1, 1)$.

Markov chain converges very slowly (no central limit theorem applies).

249

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○●	○○○○○	○○
○○○	○○○○○○○○○○		○○○
○○○○○○○	○○○○○○○		○○○○

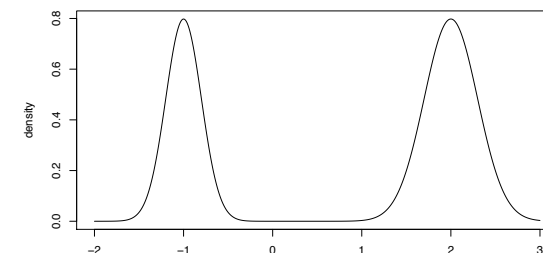
Motivation: The Need for Convergence Diagnostics

Pathological example 3: nearly reducible chain

Metropolis-Hastings sample from a mixture of two well-separated Gaussians, i.e. the target is

$$f(x) = 0.4 \cdot \phi_{(-1, 0.22)}(x) + 0.6 \cdot \phi_{(2, 0.32)}(x)$$

If the variance of the proposal is too small, the chain cannot move from one population to the other.



250

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○	●○○○○○○○○○		○○○
○○○○○○○	○○○○○○○		○○○○

Elementary Techniques for Assessing Convergence

Basic plots

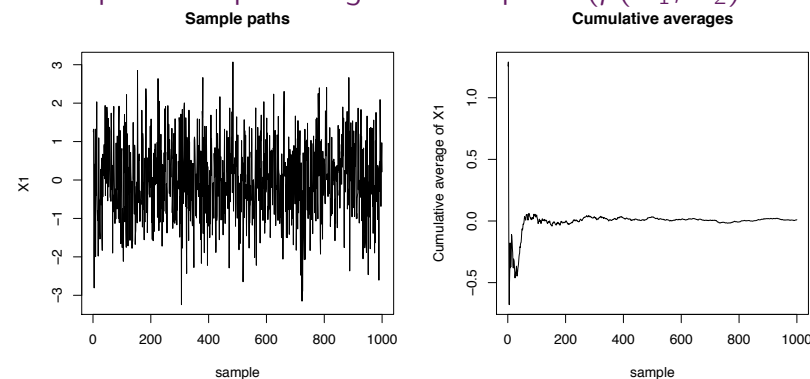
- Plot the sample paths $(X_j^{(t)})_t$.
should be oscillating very fast and show very little structure.
- Plot the cumulative averages $(\sum_{\tau=1}^t X_j^{(\tau)} / t)_t$.
should be converging to a value.
- Alternatively plot CUSUM $(\bar{X}_j - \sum_{\tau=1}^t X_j^{(\tau)} / t)_t$ with $\bar{X}_j = \sum_{\tau=1}^T X_j^{(\tau)} / T$.
should be converging to 0.
- Only very obvious problems visible in these plots.
- Difficult to assess multivariate distributions from univariate projections.

251

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○	●○○○○○○○○○		○○○
○○○○○○○	○○○○○○○		○○○○

Elementary Techniques for Assessing Convergence

Basic plots for pathological example 1 ($\rho(X_1, X_2) = 0.3$)



Looks OK.

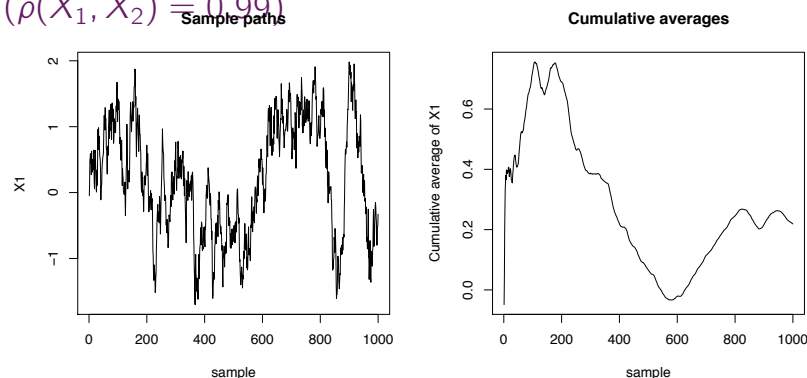
252

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○	○○●○○○○○○○	○○○○○	○○○
○○○○○○○	○○○○○○○	○○○○○○○	○○○○○

Elementary Techniques for Assessing Convergence

Basic plots for pathological example 1

$$(\rho(X_1, X_2) = 0.99)$$



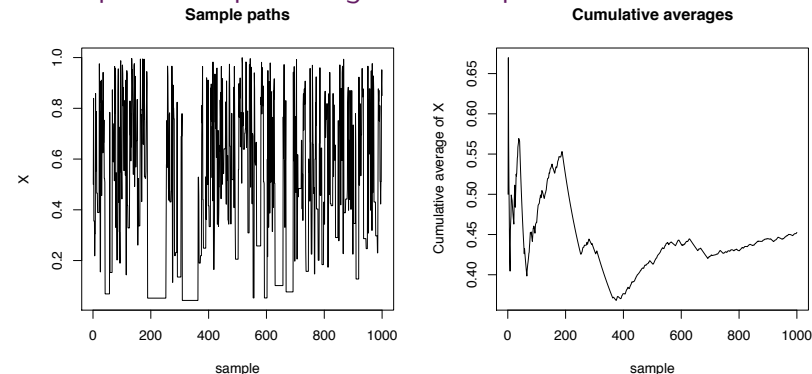
Slow mixing speed can be detected.

253

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○	○○●○○○○○○○	○○○○○	○○○
○○○○○○○	○○○○○○○	○○○○○○○	○○○○○

Elementary Techniques for Assessing Convergence

Basic plots for pathological example 2



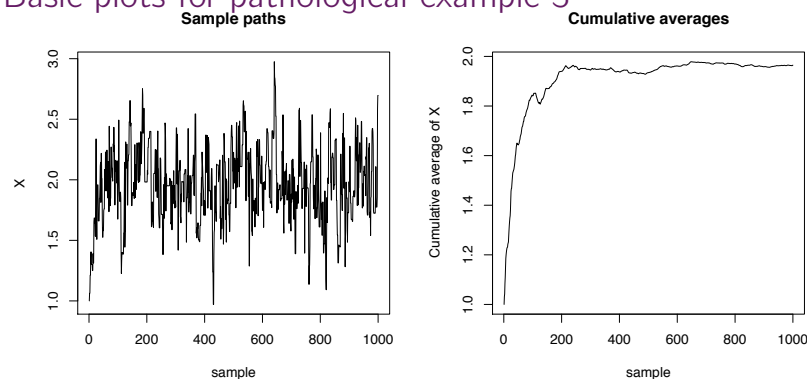
Slow convergence of the mean can be detected.

254

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○	○○●○○○○○○○	○○○○○	○○○
○○○○○○○	○○○○○○○	○○○○○○○	○○○○○

Elementary Techniques for Assessing Convergence

Basic plots for pathological example 3



We *cannot* detect that the sample only covers one part of the distribution.
("you've only seen where you've been")

255

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○	○○○○○	○○○○○	○○
○○○	○○●○○○○○○○	○○○○○	○○○
○○○○○○○	○○○○○○○	○○○○○○○	○○○○○

Elementary Techniques for Assessing Convergence

Non-parametric tests of convergence

- Partition chain in 3 blocks:
 - burn-in $(\mathbf{X}^{(t)})_{t=1, \dots, \lfloor T/3 \rfloor}$
 - first block $(\mathbf{X}^{(t)})_{t=\lfloor T/3 \rfloor + 1, \dots, 2\lfloor T/3 \rfloor}$
 - second block $(\mathbf{X}^{(t)})_{t=2\lfloor T/3 \rfloor + 1, \dots, T}$
- Distribution of $\mathbf{X}^{(t)}$ in both blocks should be identical.
- Idea: Use of a non-parametric test to test whether the two distributions are identical.
- Problem: Tests designed for i.i.d. samples.
 - \leadsto Resort to a (less correlated) thinned chain $\mathbf{Y}^{(t)} = \mathbf{X}^{(m \cdot t)}$.

256

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○●○○○ ○○○○○○○	○○○○○○	○○ ○○ ○○○

Elementary Techniques for Assessing Convergence

Kolmogorov-Smirnov test

- Two i.i.d. populations: $Z_{1,1}, \dots, Z_{1,n}$ and $Z_{2,1}, \dots, Z_{2,n}$
- Estimate empirical CDF in each population:

$$\hat{F}_k(z) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, z]}(Z_{k,i})$$

- Test statistic is the maximum difference between the two empirical CDFs: $K = \sup_{x \in \mathbb{R}} |\hat{F}_1(x) - \hat{F}_2(x)|$
- For $n \rightarrow \infty$ the CDF of $\sqrt{n} \cdot K$ converges to the CDF

$$R(k) = 1 - \sum_{i=1}^{+\infty} (-1)^{i-1} \exp(-2i^2 k^2)$$

257

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○●○○○ ○○○○○○○	○○○○○○	○○ ○○ ○○○

Elementary Techniques for Assessing Convergence

Kolmogorov-Smirnov test

- In our case the two populations are
 thinned first block $(\mathbf{Y}^{(t)})_{t=\lfloor T/(3 \cdot m) \rfloor + 1, \dots, 2\lfloor T/(3 \cdot m) \rfloor}$
 thinned second block $(\mathbf{X}^{(t)})_{t=2\lfloor T/(3 \cdot m) \rfloor + 1, \dots, \lfloor T/m \rfloor}$
- Even the thinned chain $(\mathbf{Y}^{(t)})_t$ is autocorrelated
 \leadsto test invalid from a formal point of view.
- Standardised test statistic $\sqrt{\lfloor T/(3 \cdot m) \rfloor} \cdot K$ can still be used a heuristic tool.

258

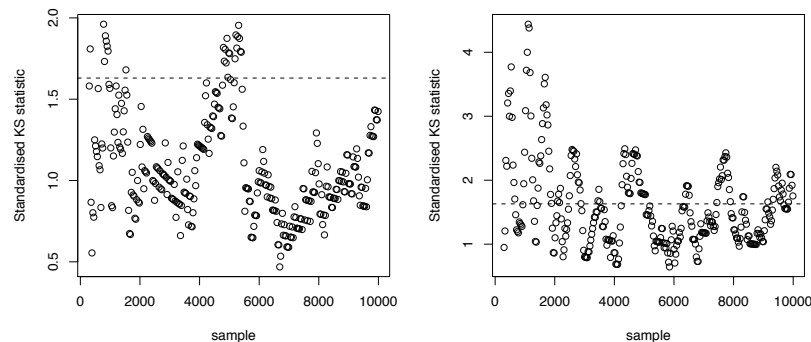
Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○●○○○ ○○○○○○○	○○○○○○	○○ ○○ ○○○

Elementary Techniques for Assessing Convergence

KS test for pathological example 1

$$\rho(X_1, X_2) = 0.3$$

$$\rho(X_1, X_2) = 0.99$$



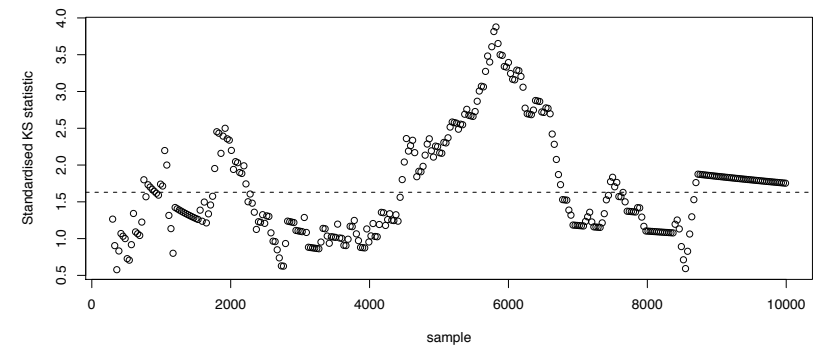
Slow mixing speed can be detected for the highly correlated chain.

259

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○●○○○ ○○○○○○○	○○○○○○	○○ ○○ ○○○

Elementary Techniques for Assessing Convergence

KS test for pathological example 2



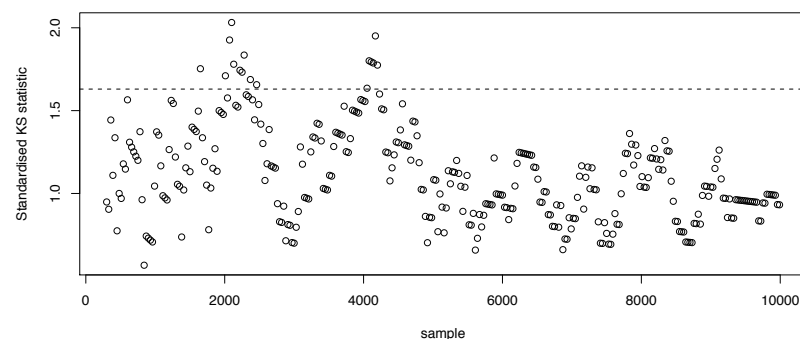
Problems can be detected.

260

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○○○● ○○○○○○○	○○○○○○○	○○ ○○○ ○○○○

Elementary Techniques for Assessing Convergence

KS test for pathological example 3



We *cannot* detect that the sample only covers one part of the distribution.
("you've only seen where you've been")

261

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○○○● ●○○○○○	○○○○○○○	○○ ○○○ ○○○○

Further Convergence Diagnostics

Comparing multiple chains

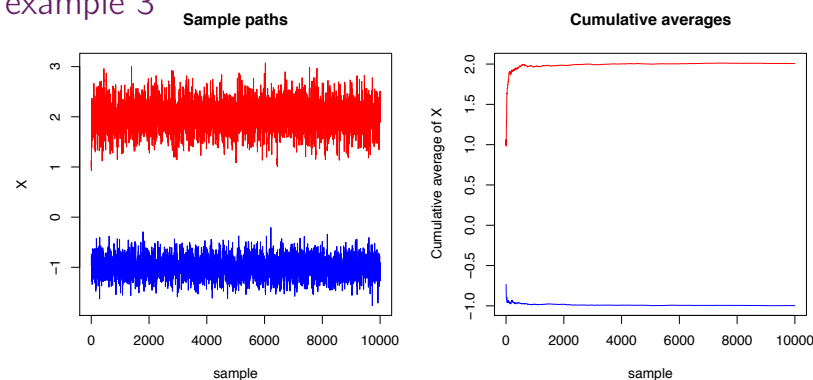
- Compare $L > 1$ chains $(\mathbf{X}^{(1,t)})_t, \dots, (\mathbf{X}^{(L,t)})_t$.
- Initialised using overdispersed starting values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$.
- Idea: Variance and range of each chain $(\mathbf{X}^{(l,t)})_t$ should equal the range and variance of all chains pooled together.
- Compare basic plots for the different chains.
- Quantitative measure:
 - Compute distance $\delta_\alpha^{(l)}$ between α and $(1 - \alpha)$ quantile of $(X_k^{(l,t)})_t$.
 - Compute distance $\delta_\alpha^{(\cdot)}$ between α and $(1 - \alpha)$ quantile of the pooled data.
 - The ratio $\hat{S}_\alpha^{\text{interval}} = \frac{\sum_{l=1}^L \delta_\alpha^{(l)} / L}{\delta_\alpha^{(\cdot)}}$ should be around 1.
- Alternative: compare variance within each chain with the pooled variance estimate.
- Choosing suitable initial values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$ difficult.

262

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○○○● ●○○○○○	○○○○○○○	○○ ○○○ ○○○○

Further Convergence Diagnostics

Comparing multiple chains plots for pathological example 3



$$\hat{S}_\alpha^{\text{interval}} = 0.2703 \ll 1$$

We can detect that the sample only covers one part of the

263

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○○○● ●○○○○○	○○○○○○○	○○ ○○○ ○○○○

Further Convergence Diagnostics

Riemann sums and control variates

- Consider order statistic $X^{[1]} \leq \dots \leq X^{[T]}$.
- Provided $(X^{[t]})_t = 1 \dots T$ covers the support of the target, the Riemann sum

$$\sum_{t=2}^T (X^{[t]} - X^{[t-1]}) f(X^{[t]})$$

converges to

$$\int f(x) dx = 1.$$

- Thus if $\sum_{t=2}^T (X^{[t]} - X^{[t-1]}) f(X^{[t]}) \ll 1$, the Markov chain has failed to explore all the support of the target.
- Requires that target density f is available inclusive of normalisation constants.
- Only effective in 1D.

264

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo	oooo	ooooo	oo
ooo	oooooooooooo		oo
oooooooo	oo●oooo		ooo

Further Convergence Diagnostics

Riemann sums for pathological example 3

For the chain stuck in the population with mean 2 we obtain

$$\sum_{t=2}^T (X^{[t]} - X^{[t-1]}) f(X^{[t]}) = 0.598 \ll 1,$$

so we can detect that we have not explored the whole distribution.

265

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo	oooo	ooooo	oo
ooo	oooooooooooo		oo
oooooooo	oooo●oo		ooo

Further Convergence Diagnostics

Effective sample size

- MCMC algorithms yield a positively correlated sample $(\mathbf{X}^{(t)})_{t=1,\dots,T}$.
- How much less useful is an MCMC sample of size T than an i.i.d. sample of size T ?
- Approximate $(\varphi(\mathbf{X}^{(t)}))_{t=1,\dots,T}$ by an $AR(1)$ process, i.e.:

$$\rho(\varphi(\mathbf{X}^{(t)}), \varphi(\mathbf{X}^{(t+\tau)})) = \rho^{|\tau|}.$$

- Variance of the estimator is

$$\text{Var} \left(\frac{1}{T} \sum_{t=1}^T \varphi(\mathbf{X}^{(t)}) \right) \approx \frac{1+\rho}{1-\rho} \cdot \frac{1}{T} \text{Var}(\varphi(\mathbf{X}^{(t)}))$$

- Same variance as an i.i.d. sample of the size $T \cdot \frac{1-\rho}{1+\rho}$.
- Thus define $T \cdot \frac{1-\rho}{1+\rho}$ as *effective sample size*.

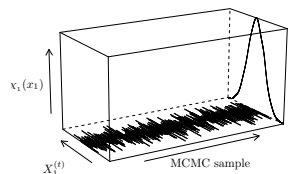
266

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo	oooo	ooooo	oo
ooo	oooooooooooo		oo
oooooooo	oooo●oo		ooo

Further Convergence Diagnostics

Effective sample for pathological example 1

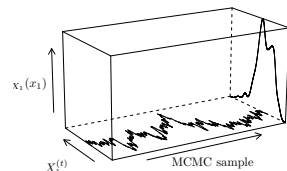
Rapidly mixing chain
 $(\rho(X_1, X_2) = 0.3)$
 10,000 samples



$$\rho(X_1^{(t-1)}, X_1^{(t)}) = 0.078$$

ESS for estimating $\mathbb{E}_f(X_1)$ is 8,547.

Slowly mixing chain
 $(\rho(X_1, X_2) = 0.99)$
 10,000 samples



$$\rho(X_1^{(t-1)}, X_1^{(t)}) = 0.979$$

ESS for estimating $\mathbb{E}_f(X_1)$ is 105.

267

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo	oooo	ooooo	oo
ooo	oooooooooooo		oo
oooooooo	oooo●oo		ooo

Further Convergence Diagnostics

What Else Can We Do?

- More sophisticated convergence diagnostics:
 - Geweke's method based on spectral analysis
 - Raftery's binary-chain method
 - ...
- Theoretical Computations
 - Convergence rates
 - Mixing times
 - Confidence intervals
- Perfect Simulation
 - Processes with "ordered transitions".
 - Certain spatial processes.

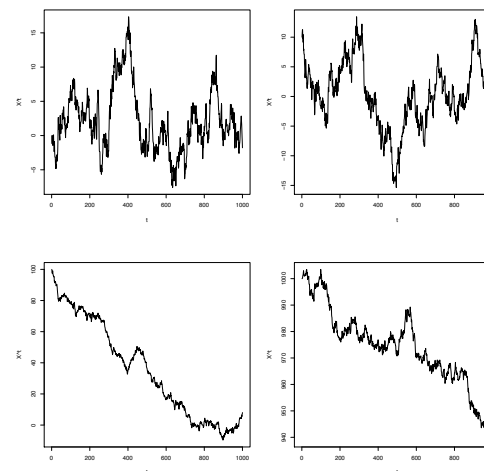
268

Part 5— Section 17

Practical Considerations

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○ ○○○ ○○○○○○○	○○○○ ○○○○○○○○○○ ○○○○○○○	○○○○○	○○ ○○○ ○○○○

Where do we start?



RWM Traces.

Target:

$$f(x) = e^{-|x|/5}/10$$

Starting values:

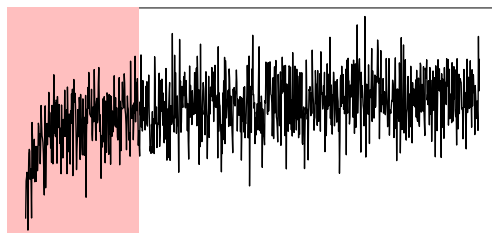
- $X^{(1)} = 0$
- $X^{(1)} = 10$
- $X^{(1)} = 100$
- $X^{(1)} = 1,000$

270

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○ ○○○ ○○○○○○○	○○○○ ○○○○○○○○○○ ○○○○○○○	○○○○○	○○ ○○○ ○○○○

Practical considerations: Burn-in period

- Theory (ergodic theorems) allows for the use of the entire chain $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$.
- However distribution of $(\mathbf{X}^{(t)})$ for small t might still be far from the stationary distribution f .
- Can be beneficial to discard the first iterations $\mathbf{X}^{(t)}$, $t = 1, \dots, T_0$ (burn-in period).
- Optimal T_0 depends on mixing properties of the chain.



271

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○ ○○○ ○○○○○○○	○○○○ ○○○○○○○○○○ ○○○○○○○	●○○○○○	○○ ○○○ ○○○○

Reducing Correlation

Practical considerations: Multiple Starts?

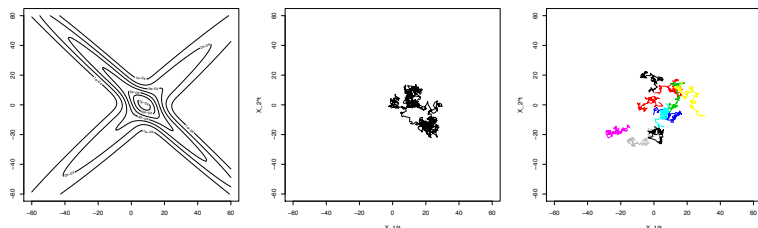
- Should we use “multiple overdispersed initialisations”?
- Advantages:
 - Exploring different parts of the space.
 - May be useful for assessing convergence.
 - Trivial to parallelize.
- Disadvantages:
 - We need to specify many starting values.
 - What does overdispersed mean, anyway?
 - Every chain needs to reach stationarity.
 - Multiple burn-in periods may be expensive.

272

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○○○○○ ○○○○○○○	●○○○○○	○○ ○○○ ○○○○

Reducing Correlation

One Chain vs. Many: 1000 or 10×100

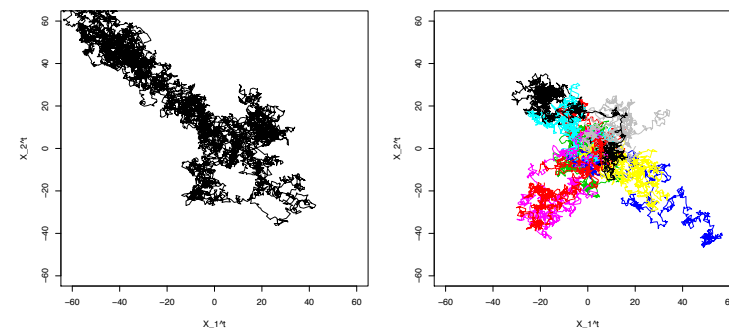


273

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○○○○○ ○○○○○○○	○○●○○○	○○ ○○○ ○○○○

Reducing Correlation

One Chain vs. Many: 10,000 or 10×1000

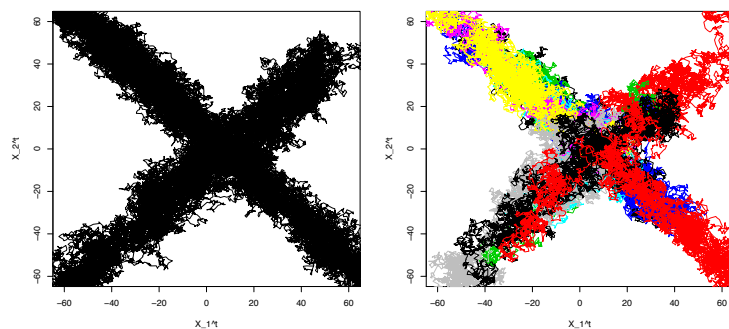


274

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○○○○○ ○○○○○○○	○○○●○○	○○ ○○○ ○○○○

Reducing Correlation

One Chain vs. Many: 100,000 or $10 \times 10,000$



275

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○○○ ○○○ ○○○○○○○	○○○○○ ○○○○○○○○○○○ ○○○○○○○	○○○○●○	○○ ○○○ ○○○○

Reducing Correlation

Practical considerations: Thinning (1)

- MCMC methods typically yield positively correlated chain: $\rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)})$ large for small τ .
- Idea: keeping only every m -th value: $(\mathbf{Y}^{(t)})_{t=1, \dots, \lfloor T/m \rfloor}$ with $\mathbf{Y}^{(t)} = \mathbf{X}^{(m \cdot t)}$ instead of $(\mathbf{X}^{(t)})_{t=1, \dots, T}$ (thinning).
- $(\mathbf{Y}^{(t)})_t$ exhibits less autocorrelation than $(\mathbf{X}^{(t)})_t$, i.e.

$$\rho(\mathbf{Y}^{(t)}, \mathbf{Y}^{(t+\tau)}) = \rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+m \cdot \tau)}) < \rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)}),$$

if the correlation $\rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)})$ decreases monotonically in τ .

- Price: length of $(\mathbf{Y}^{(t)})_{t=1, \dots, \lfloor T/m \rfloor}$ is only $(1/m)$ -th of the length of $(\mathbf{X}^{(t)})_{t=1, \dots, T}$.

276

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○ ○○ ○○○○○	○○○○ ○○○○○○○○○○ ○○○○○○○	○○○○○●	○○ ○○ ○○○

Reducing Correlation

Practical considerations: Thinning (2)

- If $\mathbf{X}^{(t)} \sim f$ and corresponding variances exist,

$$\text{Var} \left(\frac{1}{T} \sum_{t=1}^T \varphi(\mathbf{X}^{(t)}) \right) \leq \text{Var} \left(\frac{1}{\lfloor T/m \rfloor} \sum_{t=1}^{\lfloor T/m \rfloor} \varphi(\mathbf{Y}^{(t)}) \right),$$

i.e. thinning cannot be justified when objective is estimating $\mathbb{E}_f(\varphi(\mathbf{X}))$.

- Thinning can be a useful concept
 - if computer has insufficient memory.
 - for convergence diagnostics: $(\mathbf{Y}^{(t)})_{t=1, \dots, \lfloor T/m \rfloor}$ is closer to an i.i.d. sample than $(\mathbf{X}^{(t)})_{t=1, \dots, T}$.

277

Part 5— Section 18

Some Current Directions

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○ ○○ ○○○○○	○○○○ ○○○○○○○○○○ ○○○○○○○	○○○○○	●○ ○○ ○○○

Ensembles and Momentum

Ensemble-based Methods

- Allow (weighted) collections of “particles” to evolve in time.
- Can combine importance sampling and MCMC ideas.
- Well-suited to estimating normalising constants.
- Facilitate adaptation.
- Often computationally costly.

Cf. Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential Monte Carlo samplers. Journal of the Royal Statistical Society B, 63, 411–436.

279

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
○○○○○ ○○ ○○○○○	○○○○ ○○○○○○○○○○ ○○○○○○○	○○○○○	●○ ○○ ○○○

Ensembles and Momentum

Hamiltonian / Hybrid Monte Carlo

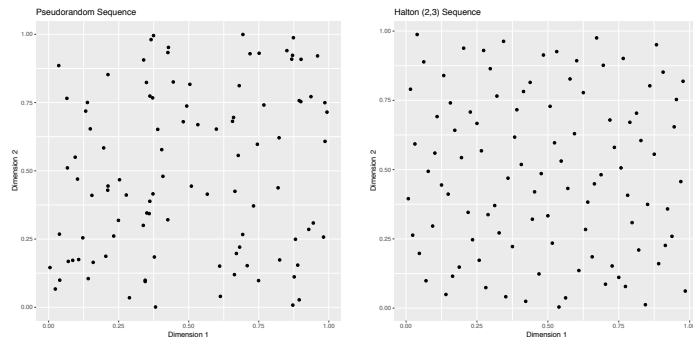
- Mimics a conservative physical system by introducing momentum.
- Approximate continuous measure-preserving flow using (symplectic) numerical integration.
- Use Metropolis-Hastings accept/reject correction.
- Can mix *much* faster than random walk algorithms.
- Difficulties with multi-modal targets and can be expensive.

Cf. Neal (2011) MCMC using Hamiltonian dynamics. In Brooks et al. (2011), 113–162. [Brooks, Gelman, Jones, and Meng (eds.) (2011) Handbook of Markov Chain Monte Carlo. CRC Press.]

280

Quasi Monte Carlo

- Why use “random” numbers?
- Wouldn't “regular” numbers be better?



281

Low Discrepancy Sequences

Definition (Discrepancy)

Given $P = \{x_1, \dots, x_n\} \subset [0, 1]^d$, the discrepancy and star discrepancy are:

$$D_N(P) = \sup_{J \in \mathcal{J}} \left| \frac{\#P \cap J}{N} - \lambda(J) \right|$$

$$D_N^*(P) = \sup_{J \in \mathcal{J}^*} \left| \frac{\#P \cap J}{N} - \lambda(J) \right|$$

where \mathcal{J} are sets of the form $\prod_{i=1}^d [a_i, b_i)$ and \mathcal{J}^* are $\prod_{i=1}^d [0, b_i)$.

- QMC: why not approximate integrals with low discrepancy (not random) sequences?
- The *Koksma-Hlawka Inequality* controls approximation error.

282

Quasi Monte Carlo

Advantages

- Can (dramatically) beat Monte Carlo's \sqrt{n} -convergence rate.
- Reduces dependency on random numbers.

Challenges

- Constructing minimum discrepancy sequences.
- Sequence extensibility.
- Transformations (& preserving discrepancy)

Cf. Niederreiter, H. (1992) Random Number Generation and Quasi-Monte Carlo Methods. Society for Industrial and Applied Mathematics.

283

Exact Approximate Methods

- Replacing appropriate quantities with unbiased estimators within other algorithms. E.g. $\pi(x)$ with $\hat{\pi}(x)$ in MH-acceptance probability.
- Justified via extended state space arguments.

Cf. Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37, 697–725; Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo. *Journal of the Royal Statistical Society B*, 72, 269–342.

284

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo	oooo	ooooo	oo
ooo	oooooooooooo		oo
ooooooo	ooooooo		ooo
			oo●o

Approximations

Approximate Approximate Methods

- Replacing difficult quantities with *biased* estimators within other algorithms.
- It's hard to control the approximation error theoretically.
- Much more speculative than pseudomarginal methods.

Cf. Medina-Aguayo, Lee, and Roberts (2016). Stability of noisy Metropolis-Hastings. *Statistics and Computing* 26(6):1187–1211.
 Alquier, Friel, Everitt. and Boland. (2016) Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26, 29–47; Everitt, Johansen, Rowing, and Evdemon-Hogan (2017). Bayesian model selection with un-normalised likelihoods. *Statistics and Computing* 27(2):403-422.

285

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo	oooo	ooooo	oo
ooo	oooooooooooo		oo
ooooooo	ooooooo		ooo
			oo●o

Approximations

Dealing with Big Data

- Distribution: sub-posteriors; consensus methods; medians of medians.
- Subsampling: unadjusted Langevin; zig-zag & bouncy particles; similar.
- A whole lot of computer science.

Cf. Bardenet, Doucet and Holmes (2015). On Markov chain Monte Carlo methods for tall arXiv:1505.02827; Scott (In Press). Comparing Consensus Monte Carlo Strategies for Distributed Bayesian Computation (*Brazilian Journal of Probability and Statistics*).

286

Theoretical Considerations	Convergence Diagnostics	Practical Considerations	Current Directions
ooooo	oooo	ooooo	oo
ooo	oooooooooooo		oo
ooooooo	ooooooo		ooo●

Approximations

Thank you!

287