Imperial College London

Advanced Computational Methods in Statistics: Lecture 3 - MCMC

Axel Gandy

Department of Mathematics
Imperial College London
www2.imperial.ac.uk/~agandy

London Taught Course Centre
for PhD Students in the Mathematical Sciences
Autumn 2014

Outline

MCMC methods

- Markov Chain Monte Carlo
- Main idea:
  - Want to simulate from a density f or compute functionals of f such as the mean: $\mathrm{E}\,X = \int x f(x)dx$.
  - Construct a Markov Chain whose stationary distribution is f.

  Note: Usually f need only be known up to a normalising constant.

Most of the material in this lecture is from Robert & Casella (2004).

MCMC and Bayesian Models

- MCMC is the main tool used in (applied) Bayesian statistics!
- Observation y
- Model: $Y \sim g(\cdot|\theta)$, $\theta \sim \pi$
- Mainly interested in the a-posteriori density:

$$\pi(\theta|y) = \frac{g(y|\theta)\pi(\theta)}{m(y)},$$

  where $m(y) = \int g(y|\theta)\pi(\theta)d\theta$.
- If θ is high-dimensional - hard to report $\pi(\theta|y)$
  $\rightarrow$ report e.g. the posterior mean

$$E(\theta|y) = \int \theta\pi(\theta|y)dy.$$

- MCMC: construct Markov chain $X_1, X_2, \dots$ with stationary distribution $\pi(\theta|y)$ (evaluation of m is not needed)
  run Markov chain for n steps; then $E(\theta|y) \approx \frac{1}{n}\sum_{i=1}^{n} X_i$

Outline

Definitions

- A sequence $X_0, X_1, X_2, \dots$ of random variables (random objects) is a Markov chain if for all A and $n \in \mathbb{N}$:

$$P(X_{n+1} \in A | X_n, \dots, X_0) = P(X_{n+1} \in A | X_n).$$

  In words: only the distribution of the current state is relevant for the distribution of the state at the next time.
  Note: discrete time, potentially continuous state.
- It is called (time) homogeneous if for all $t_0 \leq t_1 \leq \cdots \leq t_k$:

$$(X_{t_k}, X_{t_{k-1}}, \dots, X_{t_1}) | X_{t_0} \sim (X_{t_k - t_0}, X_{t_{k-1} - t_0}, \dots, X_{t_1 - t_0}) | X_0$$

  The Markov-chains we encounter will be time-homogeneous.
  Example: $k = 2$, $t_2 = 10$, $t_1 = 8$, $t_0 = 7$. For a time homogeneous chain, $(X_{10}, X_8) | X_7 \sim (X_3, X_1) | X_0$.
- transition kernel (corresponding to transition matrix):

$$K(x, B) = P(X_{n+1} \in B | X_n = x)$$

  Note: $\forall x : K(x, \cdot)$ is a probability measure.

Irreducibility, Recurrence

$\mathcal{X}$ finite: Irreducibility, Recurrence about reaching individual points.
Here: modification for $\mathcal{X}$ continuous.

- $\mathcal{X}$ state space of the Markov chain (X_n)
- $\tau_A = \inf\{n \geq 1 : X_n \in A\}$ (first hitting time of A)
- Let ϕ be a measure.
  (X_n) is ϕ-irreducible if
  $\forall A$ with $\phi(A) > 0$: $\mathrm{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$.
- $\eta_A = \sum_{n=1}^{\infty} 1_A(X_n)$ (number of passages of X_n through A)
- (X_n) is recurrent if
  1. $\exists$ measure ϕ s.t. (X_n) is ϕ-irreducible
  2. $\forall$ A with $\phi(A) > 0$: $\mathrm{E}_x(\eta_A) = \infty$ $\forall$ $x \in A$.
- (X_n) is Harris recurrent if
  1. $\exists$ a measure ϕ s.t. (X_n) is ϕ-irreducible
  2. $\forall A$ with $\phi(A) > 0$: $\mathrm{P}_x(\eta_A = \infty) = 1$ $\forall$ $x \in A$.

(P_x = Prob measure of Markov chain started at x,
E_x = expectation taken w.r.t. P_x)

Irreducibility, Recurrence

$\mathcal{X}$ finite: Irreducibility, Recurrence about reaching individual points.
Here: modification for $\mathcal{X}$ continuous.

- $\mathcal{X}$ state space of the Markov chain (X_n)
- $\tau_A = \inf\{n \geq 1 : X_n \in A\}$ (first hitting time of A)
- Let ϕ be a measure.
  (X_n) is ϕ-irreducible if
  $\forall A$ with $\phi(A) > 0$: $\mathrm{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$.
- $\eta_A = \sum_{n=1}^{\infty} 1_A(X_n)$ (number of passages of X_n through A)
- (X_n) is recurrent if
  1. $\exists$ measure ϕ s.t. (X_n) is ϕ-irreducible
  2. $\forall$ A with $\phi(A) > 0$: $\mathrm{E}_x(\eta_A) = \infty$ $\forall$ $x \in A$.
- (X_n) is Harris recurrent if
  1. $\exists$ a measure ϕ s.t. (X_n) is ϕ-irreducible
  2. $\forall A$ with $\phi(A) > 0$: $\mathrm{P}_x(\eta_A = \infty) = 1$ $\forall$ $x \in A$.

(P_x = Prob measure of Markov chain started at x,
E_x = expectation taken w.r.t. P_x)

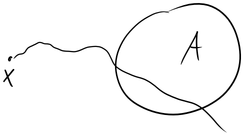
Irreducibility, Recurrence

$\mathcal{X}$ finite: Irreducibility, Recurrence about reaching individual points.
Here: modification for $\mathcal{X}$ continuous.

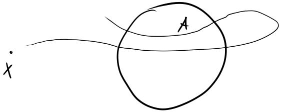
- $\mathcal{X}$ state space of the Markov chain (X_n)
- $\tau_A = \inf\{n \geq 1 : X_n \in A\}$ (first hitting time of A)
- Let ϕ be a measure.
  (X_n) is ϕ-irreducible if
  $\forall A$ with $\phi(A) > 0$: $\mathrm{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$.
- $\eta_A = \sum_{n=1}^{\infty} 1_A(X_n)$ (number of passages of X_n through A)
- (X_n) is recurrent if
  1. $\exists$ measure ϕ s.t. (X_n) is ϕ-irreducible
  2. $\forall$ A with $\phi(A) > 0$: $\mathrm{E}_x(\eta_A) = \infty$ $\forall$ $x \in A$.
- (X_n) is Harris recurrent if
  1. $\exists$ a measure ϕ s.t. (X_n) is ϕ-irreducible
  2. $\forall A$ with $\phi(A) > 0$: $\mathrm{P}_x(\eta_A = \infty) = 1$ $\forall$ $x \in A$.

(P_x = Prob measure of Markov chain started at x,
E_x = expectation taken w.r.t. P_x)

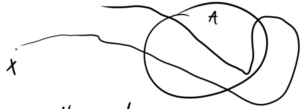
ϕ-irreducible: All sets A with $\phi(A)>0$ are reached from anywhere.



ϕ-recurrent: The expected number of times a set A with $\phi(A)>0$ is reached is infinite.



Harris-recurrent: Every set A with $\phi(A)>0$ is reached infinitely often.



Harris recurrence is much stronger than ϕ-recurrence:

X r.v. with $P(X>t)=\frac{1}{t}$ for $t>1$.

Then $P(X=\infty)=0$ but $E(X)=\int_1^\infty \frac{1}{t}\,dt=\log(\infty)-\log(1)=\infty$.

Ergodic Theorems

- Ergodic Theorems= convergence results equivalent to the law of large numbers in the iid case.
- A σ-finite measure π is invariant for the transition kernel $K(\cdot,\cdot)$ (and for the associated chain) if

$X_n \sim \pi$ the $P(X_{n+1} \in B) =$

$$\pi(B) = \int_{\mathcal{X}} K(x,B)\pi(dx), \forall B \in \mathcal{B}(\mathcal{X})$$

In other words: $\boxed{X_n \sim \pi \implies X_{n+1} \sim \pi}$

- Ergodic Theorem: If (X_n) has a σ-finite invariant measure π then the following two statements are equivalent:
  1. If $f, \cancel{g} \in L^1(\pi)$ with $\cancel{\int g(x) d\pi(x) \neq 0}$ then
  $$\frac{\frac{1}{n}\sum_{i=1}^n f(X_i)}{\cancel{\frac{1}{n}\sum_{i=1}^n g(X_i)}} \to \frac{\int f(x)\pi(dx)}{\cancel{\int g(x)\pi(dx)}} \quad (n \to \infty)$$
  2. (X_n) is Harris recurrent

Theorem (Convergence to the Stationary Distribution)

If (X_n) is Harris recurrent and aperiodic with invariant probability measure π then

$$\lim_{n\to\infty} \left\| \overbrace{\int K^n(x,\cdot)\mu(dx)}^{\text{dist. of chain after } n \text{ steps}} - \pi \right\|_{TV} = 0,$$

for every initial distribution μ, where

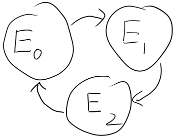
K^n *is the n step transition kernel and*

$\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \sup_A |\mu_1(A) - \mu_2(A)|$ *is the total variation norm.*

(X_n) is periodic if there exist $d \geq 2$ and nonempty disjoint sets $E_0, \ldots, E_{d-1}$ s.t. for all $i = 0, \ldots, d-1$ and all $x \in E_i$:

$$K(x, E_j) = 1 \quad \text{for } j = i + 1 \pmod{d}$$

Otherwise (X_n) is aperiodic.



periodic with $d=3$:

$P(X_{n+1} \in E_1 \mid X_n \in E_0) = 1$

" " E_2 " E_1 "

" " E_0 " E_2 "

Outline

Metropolis-Hastings algorithm

- (target) distribution f
- conditional density q (proposal of new position).

Let X^1 be arbitrary.
For $t = 1, 2, \dots$:

- Let $Y^t \sim q(X^t, \cdot)$
- Let

$$X^{t+1} = \begin{cases} Y^t \text{ with prob } \rho(X^t, Y^t) \\ X^t \text{ with prob } 1 - \rho(X^t, Y^t) \end{cases}$$

where $\rho(x, y) = \min\left(\frac{f(y)q(y,x)}{f(x)q(x,y)}, 1\right)$

Notes:

- f is only needed up to a normalising constant.
- the terms involving q cancel if proposal is symmetric around the current position.

Example - Space-Shuttle O-ring

- Explosion of the Space-shuttle Challenger caused by the failure of an *O-ring* (a ring of rubber used as a sealant)
- Caused by unusually low temperatures (31° F)
- Data from previous flights:

Failure	1	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
Temp	53	57	58	63	66	67	67	67	68	69	70	70	70	70	72	73	75	75	76	76	78	79	81

- Failure= blowby or erosion (diagnosed after the flight)
- More details: see Dalal et al. (1989).

Example - Space-Shuttle O-ring - Model

- Logistic model:

$$P(Y=1) = \frac{\exp(\alpha + x\beta)}{1+\exp(\alpha + x\beta)}$$

x =temperature

- prior:

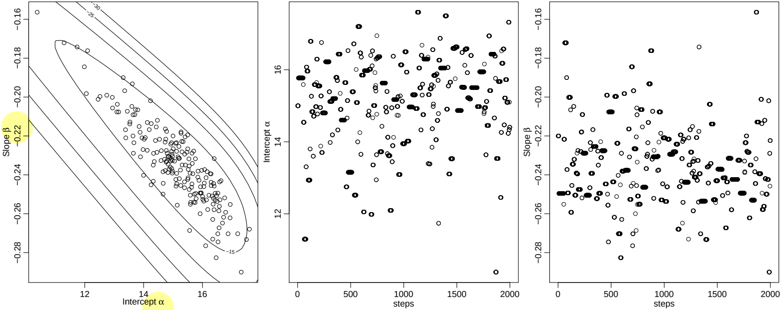
$$\pi(\alpha,\beta) = \frac{1}{b} e^{\alpha} e^{-e^{\alpha}/b}$$

(flat prior on β, exponential on $\log(\alpha)$)
choose b st $\mathrm{E}\,\alpha$=MLE of α.

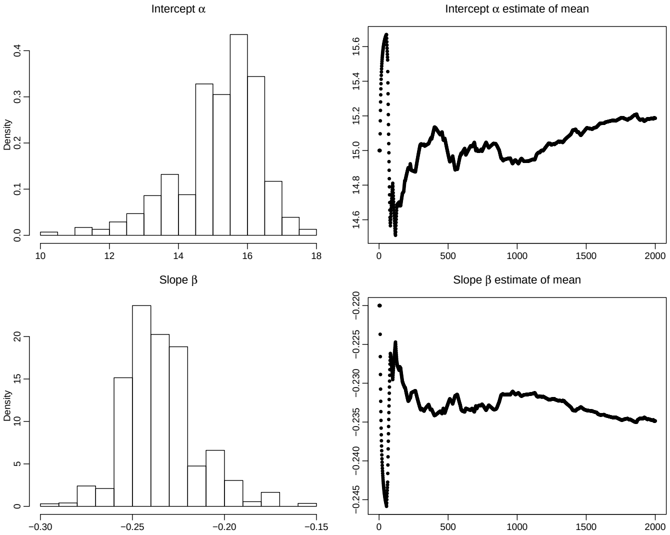
Space-Shuttle O-ring - Independent Proposal

Proposal for the Metropolis Hastings Algorithm

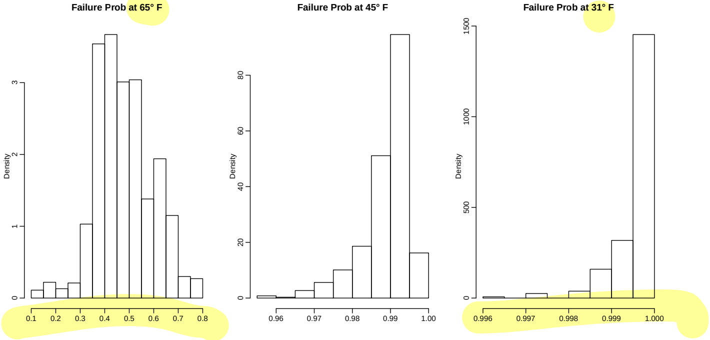
- $\exp(\alpha_{prop}) \sim \text{Exponential}(1/b)$
- $\beta_{prop} \sim N(-0.2322, 0.1082)$
- Realisation of the Markov chain:



Posterior Distribution, Mean of posterior



Prediction of Failure Probability



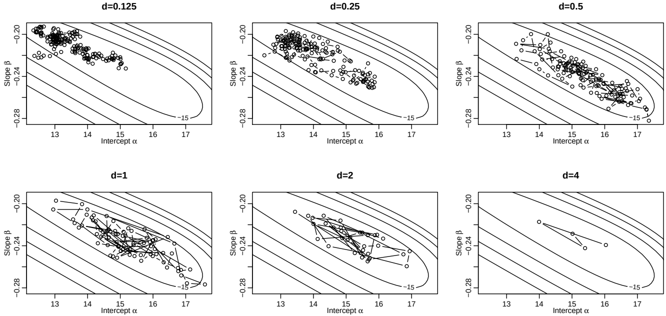
Space-Shuttle O-ring - Random Walk Proposal

Proposal for the Metropolis Hastings Algorithm → Gareth Roberts

- $\alpha_{prop} = \alpha + Z_a, \quad Z_a \sim N(0, \sqrt{0.02d})$
- $\beta_{prop} = \beta + Z_b, \quad Z_b \sim N(0, \sqrt{d})$

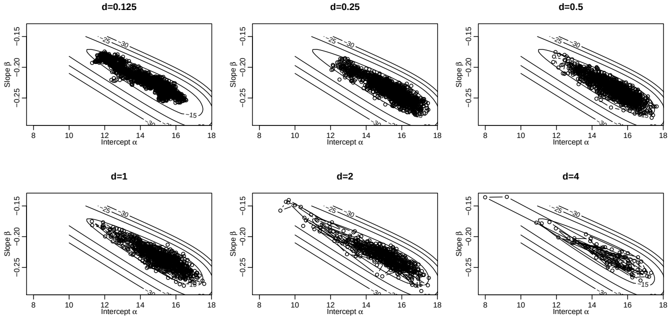
Acceptance prob simplifies: $\rho(x, y) = \min\left(\frac{f(y)\cancel{q(y,x)}}{f(x)\cancel{q(x,y)}}, 1\right)$

First 200 steps:



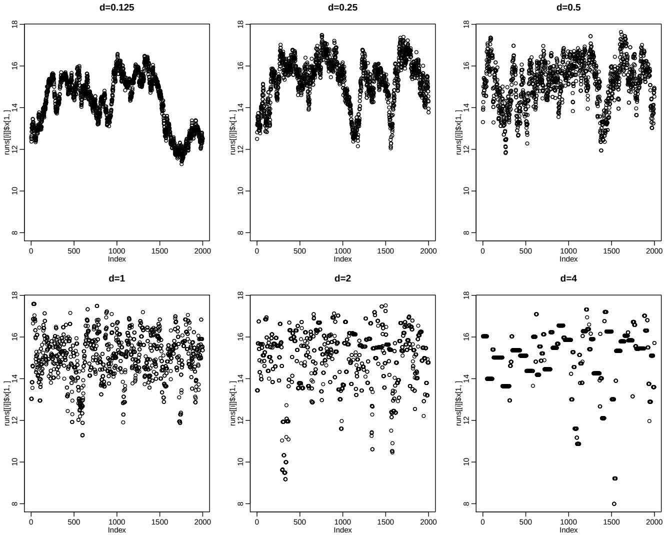
Space-Shuttle O-ring - Random Walk Proposal (cont)

First 2000 steps:

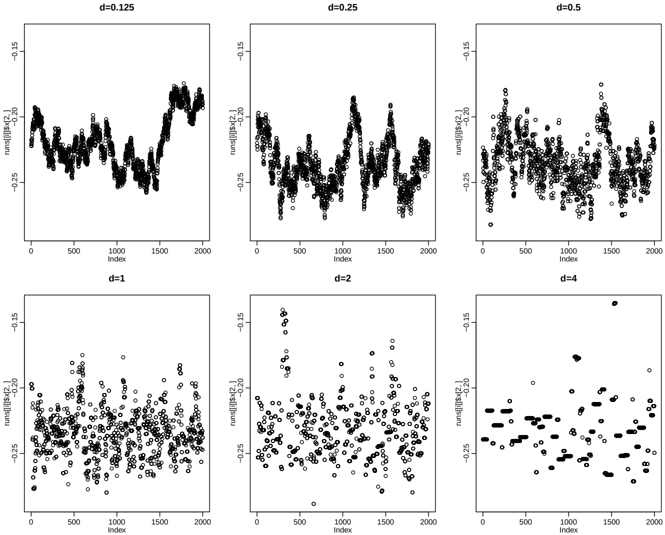


d	0.125	0.25	0.5	1	2	4
Acceptance Rate	0.88	0.762	0.5645	0.3275	0.1655	0.0425

Space-Shuttle O-ring - Random Walk Proposal - Intercept



Space-Shuttle O-ring - Random Walk Proposal - Slope



Sufficient Condition for Stationary Densities

Definition

A Markov chain with transition kernel K satisfies the detailed balance condition with the probability density function f if

$$K(x,y)f(x) = K(y,x)f(y) \quad \forall x,y$$

Remarks

- $K(x,y)f(x)$=mass flowing from x to y.
  $K(y,x)f(y)$=mass flowing from y to x.
- Detailed balance is (up to measure theoretic complications) equivalent to "reversibility":
  A stationary Markov chain (X_n) is *reversible* if $(X_{n+1}|X_{n+2}=x) \sim (X_{n+1}|X_n = x)$.

Sufficient Condition for Stationary Densities

Definition

A Markov chain with transition kernel K satisfies the detailed balance condition with the probability density function f if

$$K(x,y)f(x) = K(y,x)f(y) \quad \forall x, y$$

Theorem

Suppose a Markov chain satisfies the detailed balance condition with the pdf f. Then f is the invariant density of the chain.

Proof.

Let $X_n \sim f$. Then $\forall B$:

$$\begin{aligned} \mathrm{P}(X_{n+1} \in B) &= \int_{\mathcal{X}} K(y,B)f(y)dy = \int_{\mathcal{X}} \int_B K(y,x)f(y)dxdy \\ &= \int_{\mathcal{X}} \int_B K(x,y)f(x)dxdy = \int_B \underbrace{\int_{\mathcal{X}} K(x,y)dy}_{=1} f(x)dx = \mathrm{P}(X_n \in B) \end{aligned}$$

Stationary Distribution of the Metropolis-Hastings Alg.

every region is proposed

Theorem

Suppose $\bigcup_{x \in \operatorname{supp} f} \operatorname{supp} q(x, \cdot) \supset \operatorname{supp} f$. *Then f is a stationary distribution of the chain.*

Proof.

Will verify the detailed balance condition

$K(x, y) f(x) = K(y, x) f(y) \quad \forall x, y.$

Here,

$$K(x, y) = \overbrace{\rho(x, y) q(x, y)}^{\text{accept step}} + \overbrace{(1 - r(x)) \delta_x(y)}^{\text{rejecting step}}.$$

where $r(x) = \int \rho(x, y) q(x, y) dy$ is the overall acceptance probability at x and δ_x is the Dirac measure at x. Suffices to check

(a) $\rho(x, y) q(x, y) f(x) = \rho(y, x) q(y, x) f(y)$

(b) $(1 - r(x)) \delta_x(y) f(x) = (1 - r(y)) \delta_y(x) f(y)$

Both sides of (b)=0 for $x \neq y$;

To see (a): $\rho(x, y) = 1$ or $\rho(y, x) = 1$

(Recall: $\rho(x, y) = \min\left(\frac{f(y)q(y,x)}{f(x)q(x,y)}, 1\right)$)

Suppose $\rho(y, x) = 1$

then $\rho(x, y) = \dfrac{f(y)\, q(y, x)}{f(x)\, q(x, y)}$

Ergodicity of the Metropolis Hastings Algorithm

Let (X^t) be the Markov chain of a Metropolis Hastings algorithm.

- (X^t) is f-irreducible if

$$q(x, y) > 0 \text{ for every } (x, y)$$

  Then (X^t) is Harris-recurrent and the Ergodic theorem applies, i.e. $\forall h \in L^1(f)$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h(X^t) = \int h(x)f(x)dx \quad \text{a.s.}$$

- If (X^t) is also aperiodic then

$$\lim_{n\to\infty} \| \int K^n(x, \cdot)\mu(dx) - f\|_{\text{TV}} = 0,$$

  for every initial distribution μ, where K^n denotes the n step transition kernel.
- (X^t) is aperiodic if the probability of rejecting a step is positive (i.e. $\text{P}(X^t = X^{t+1}) > 0$).

What is a good acceptance rate?

- Independent Proposal Distribution:
  As close to 1 as possible
  (ideally, I would like the proposal distribution to equal the distribution to be simulated)
- Random Walk:
  - too high: support of f is not explored quickly
    In particular if the density is multimodal
  - too low: waste of simulations (proposals outside the range of f)
  - Heuristic: acceptance rate of $1/4$ for high-dimensional models and of $1/2$ for models of dimension 1 or 2.
    See Roberts et al. (1997).

Adaptive Schemes

- Unrealistic to hope for a generic MCMC sampler that works in every possible setting
- Problems: High dimension, disconnected support
- Problems of adaptive schemes (prior states of the Markov Chain are used to tune e.g. the proposal distribution): Markov property gets lost $\rightarrow$ loss of theoretical underpinning
- Article on theoretical underpinning of adaptive MCMC: e.g. Andrieu & Moulines (2006)
- To be on the safe side:
  - Use a burn-in period to tune parameters such as the proposal distribution.
  - The burn-in period should not contribute to expectations/quantiles of the target distribution.

Outline

Gibbs Sampler - Introduction

- Origin of the name "Gibbs sampling":
  Geman & Geman (1984), who brought Gibbs sampling into statistics, used the method for a Bayesian study of Gibbs random fields, which have their name from the physicist Gibbs (1839-1903)
- Main idea:
  - update components of the Markov Chain individually
  - by sampling the component to be updated conditional on the value of the other components.

The Gibbs Sampler

Want to sample from the density $f : \mathbb{R}^p \to [0, \infty)$
f_j=conditional density of $X_j | \{X_i, i \neq j\}$
Let X^0 be some starting value.
For $t = 0, 1, 2, \dots$:

- $X_1^{t+1} \sim f_1(x_1 | X_2^t, \dots, X_p^t)$
- $X_2^{t+1} \sim f_2(x_2 | X_1^{t+1}, X_3^t, \dots, X_p^t)$
- $\dots$
- $X_p^{t+1} \sim f_p(x_p | X_1^{t+1}, \dots, X_{p-1}^{t+1})$

Example - Truncated Normal

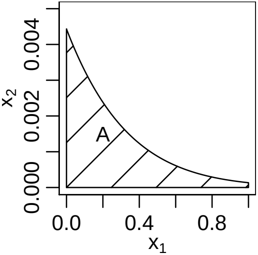
Want to sample from $N(-3,1)$ truncated to $[0,1]$, i.e.

$$f(x) \propto \exp\left(-\frac{(x+3)^2}{2}\right) \mathrm{I}(0 \leq x \leq 1)$$

Consider the uniform distribution g on

$$A = \left\{(x_1, x_2)' : x_1 \in [0,1], 0 \leq x_2 \leq f(x_1)\right\}$$

f is the marginal density of the first component.



Gibbs sampler for g

- $g_1(x_1|x_2) \propto \mathrm{I}(0 \leq x_1 \leq \min(1, -3 + \sqrt{-2\log x_2}))$
- $g_2(x_2|x_1) \propto \mathrm{I}(0 \leq x_2 \leq f(x_1))$

Example - Truncated Normal

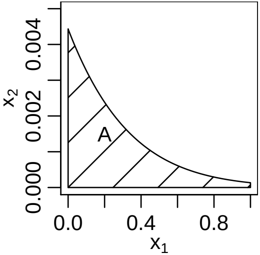
Want to sample from $N(-3, 1)$ truncated to $[0, 1]$, i.e.

$$f(x) \propto \exp\left(-\frac{(x+3)^2}{2}\right) \mathbb{I}(0 \leq x \leq 1)$$

Consider the uniform distribution g on

$$A = \left\{(x_1, x_2)' : x_1 \in [0, 1], 0 \leq x_2 \leq f(x_1)\right\}$$

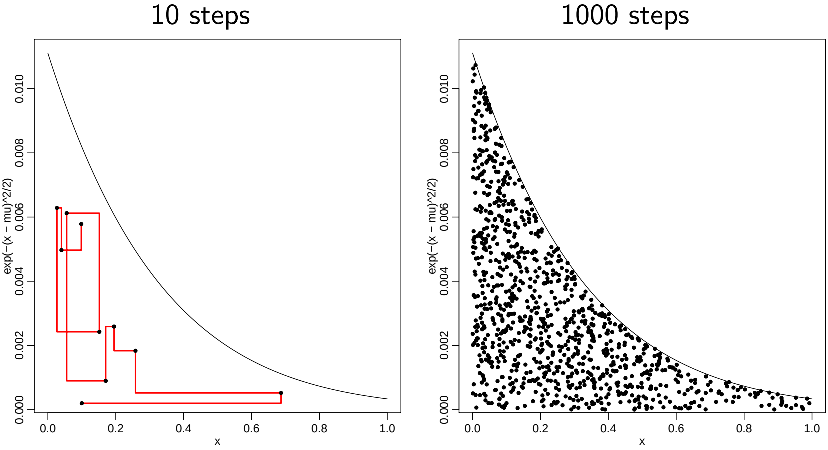
f is the marginal density of the first component.



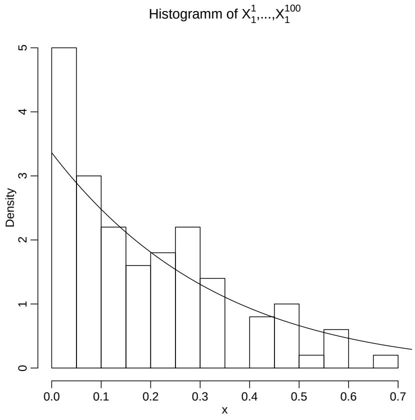
Gibbs sampler for g

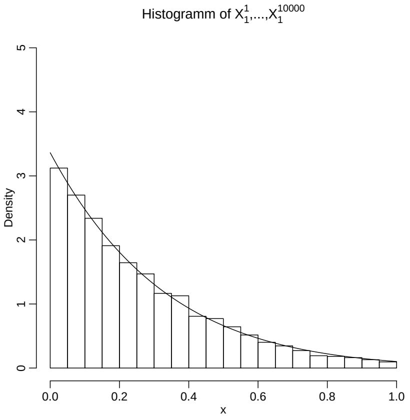
- $g_1(x_1|x_2) \propto \mathbb{I}(0 \leq x_1 \leq \min(1, -3 + \sqrt{-2 \log x_2}))$
- $g_2(x_2|x_1) \propto \mathbb{I}(0 \leq x_2 \leq f(x_1))$

Example - Truncated Normal



Example - Truncated Normal





Gibbs-Sampler- Stationary Distribution

- Will show that f is stationary for each of the p steps
- WLOG consider the first step
- Need to show: If $(X_1, X_2, \dots, X_p) \sim f$ and $\tilde{X}_1 \sim f_1(x_1|X_2, \dots, X_p)$ then $(\tilde{X}_1, X_2, \dots, X_p) \sim f$
- Let $X_{-1} = (X_2, \dots, X_p)$, $x_{-1} = (x_2, \dots, x_p)$.
- Let $p_A := \mathrm{P}((\tilde{X}_1, X_2, \dots, X_p) \in A)$.

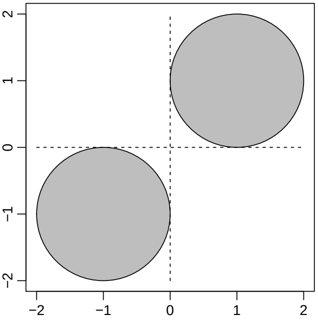
$$p_A = \int \int \mathtt{I}((\tilde{x}_1, x_{-1}) \in A) f_1(\tilde{x}_1|x_{-1}) d\tilde{x}_1 f(x) dx$$

- Using $\int f(x) dx_1 = \int f_1(x_1|x_{-1}) f_{-1}(x_{-1}) dx_1 = f_{-1}(x_{-1})$,

$$\begin{aligned} p_A &= \int \int \mathtt{I}((\tilde{x}_1, x_{-1}) \in A) f_1(\tilde{x}_1|x_{-1}) d\tilde{x}_1 f_{-1}(x_{-1}) dx_{-1} \\ &= \int \int \mathtt{I}((\tilde{x}_1, x_{-1}) \in A) f(\tilde{x}_1, x_{-1}) d\tilde{x}_1 dx_{-1} = \int \mathtt{I}(x \in A) f(x) dx \end{aligned}$$

Gibbs-Sampler- Disconnected Support - Example

- Let D_1 and D_2 be discs in $\mathbb{R}^2$ with radius 1 and and centres $(1,1)$ and $(-1,-1)$
- Consider the uniform distribution on $D_1 \cup D_2$
- Gibbs Sampler is not an irreducible chain (remains concentrated in the disc it is started in)
- (transformation of coordinates to $x_1 + x_2$ and $x_2 - x_1$ would solve the problem)



Gibbs Sampler - Some Theoretical Results

- If f satisfies the following positivity condition then the resulting Gibbs sampler is f-irreducible.

$$f^{(i)}(x_i) > 0 \forall i \implies f(x_1, \dots, x_p) > 0$$

  ($f^{(1)}, \dots, f^{(p)}$ denote the marginal distributions)
- If a Gibbs sampler is
  - f-irreducible with stationary distribution f and
  - for every x the transition probability $K(x, \cdot)$ is absolutely continuous with respect to f

  then the Gibbs sampler is Harris recurrent. (Tierney, 1994, Corollary 1)
- (Recall: Harris recurrence implies the usual ergodicity results)

BUGS software

- Bayesian inference Using Gibbs Sampling
- "flexible software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods"
- Allows specification of Bayesian models in the BUGS language. MCMC chain is constructed automatically.
- Original version: WinBUGS
- Open source version: OpenBUGS
- Similar: JAGS (based on C, hopefully more portable)

Stan

Outline

Introduction

- A variable dimension model is a “model where one of the things you do not know is the number of things you do now know” (Peter Green)
- in other words: the dimension of the parameter space is not fixed.
- can occur in model selection, checking, improvement, ...

Bayesian variable dimension model

- A Bayesian variable dimension model is defined as a collection of models $(k = 1, \dots, K)$,

$$\mathcal{M}_k = \{f(\cdot|\theta_k); \theta_k \in \Theta_k\},$$

  with a collection of priors on the parameters of these models,

$$\pi_k(\theta_k),$$

  and a prior distribution $\rho_k, k = 1, \dots, K$ on the indices of these models.
- Note: Θ_k may have different dimensions
- In this setting one can compute the posterior probability of models, i.e.

$$p(\mathcal{M}_k|\mathbf{y}) = \frac{\rho_k \int f_k(\mathbf{y}|\theta_k)\pi_k(\theta_k)d\theta_k}{\sum_j \rho_j \int f_j(\mathbf{y}|\theta_j)\pi_j(\theta_j)d\theta_j}$$

Reversible Jump Algorithm

- Want: proper framework for designing moves between models $\mathcal{M}_k$
- Construction of a reversible kernel K on $\Theta = \bigcup_k \{k\} \times \Theta_k$
- Main ideas of Green (1995):
  only consider moves between pairs of models.
  construct "dimension matching" moves.
  accept a move with probability similiar to the
  Metropolis-Hastings algorithm

Toy Example

(from a tutorial written by Peter Green, see `http://www.maths.bris.ac.uk/~mapjg/slides/tdtut4.pdf`)

- $x \in \mathbb{R} \cup \mathbb{R}^2$
- $\pi(x)$ is a mixture:
  - x is $U(0,1)$ with probability p_1
  - x is uniform on the triangle $0 < x_2 < x_1 < 1$ with probability $1 - p_1$.
- Three moves:
  1. within $\mathbb{R}$: $x \to U(\max(0, x - \epsilon), \min(1, x + \epsilon))$
  2. within $\mathbb{R}^2$: $(x_1, x_2) \to (1 - x_2, 1 - x_1)$
  3. between $\mathbb{R}$ and $\mathbb{R}^2$

  If $x \in \mathbb{R}$: choose moves (1), (3) with probability $1 - r_1$, r_1
  If $x \in \mathbb{R}^2$: choose moves (2), (3) with probability $1 - r_2$, r_2

Toy Example (cont)

- Trans-dimensional move [(3)]:
  - From $x \in \mathbb{R}$ to $(x_1, x_2) \in \mathbb{R}^2$: draw u from $U(0,1)$, propose (x, u)
    Accept with probability

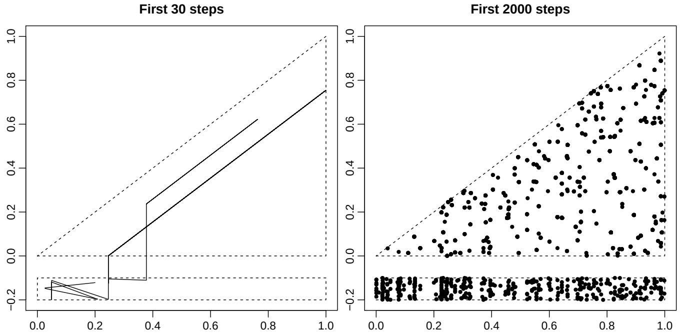
$$\alpha = \min(1, \frac{2(1-p_1)r_2}{p_1 r_1})\mathbb{I}(u < x)$$

  - From $(x_1, x_2) \in \mathbb{R}^2$ to $x \in \mathbb{R}$: propose $x = x_1$

$$\alpha = \min(1, \frac{p_1 r_1}{2(1-p_1)r_2})$$

Toy Example - Results

$p_1 = 0.2,\ r_1 = 0.7,\ r_2 = 0.4,\ \epsilon = 0.3$



Outline

Diagnosing Convergence

- To diagnose convergence to the stationary distribution: plot the parameter ("trace plots").
- Start multiple chains and compare the "within chain variance" to the variance when all chains are thrown together.
- Fundamental problem is mixing - you will never see if you have not explored the entire parameter space.
- No "magic" solution
- Even if you have (somehow) established that the chain is exploring the entire parameter space, there is still the issue of convergence - how long should you run the chain(s)?

Confidence intervals for standard Monte Carlo simulations

- Standard CLT: Suppose $X, X_1, X_2, \dots$ iid with $0 < \mathrm{Var}(X) < \infty$. Then
$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}(X)\right) \xrightarrow{d} N(0, \mathrm{Var}(X)) \quad (n \to \infty)$$
- $\mathrm{Var}(X)$ can be reasonably well estimated by the sample variance
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2$$
- Thus an asymptotic $1-\alpha$ confidence interval for $\mathrm{E}(X)$ is
$$\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{\sqrt{n}} cS, \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{\sqrt{n}} cS\right]$$
where c is such that $\Phi(1-c) = \frac{\alpha}{2}$.

CLT for Markov chains

- Suppose $X_1, X_2, \dots$ is a stationary Markov chain. Then, under suitable conditions,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}(X)\right) \xrightarrow{d} N(0,\sigma^2) \quad (n \to \infty)$$

where

$$\sigma^2 = \mathrm{Var}(X_i) + 2\sum_{k=1}^{\infty} \mathrm{Cov}(X_i, X_{i+k}). \tag{1}$$

Limiting variance is more complicated.

Batch Means

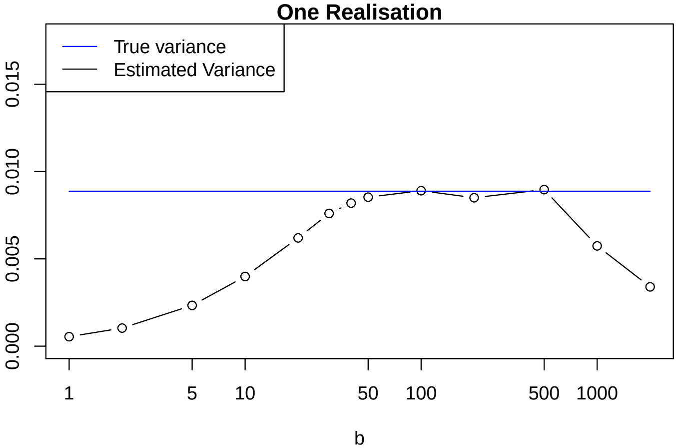
- Markov chain $X_1, X_2, \ldots$. Interested in $\mu = \mathrm{E}(g(X))$. Assume we want to use the estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$.

$$\underbrace{g(X_1) \cdots g(X_\ell)}_{\hat{\mu}_1} \quad \underbrace{g(X_{\ell+1}) \cdots g(X_{2\ell})}_{\hat{\mu}_2} \quad \cdots, \quad \underbrace{g(X_{n-\ell+1}) \cdots g(X_n)}_{\hat{\mu}_{n/\ell}}$$

- Assuming b divides n, let $\hat{\mu}_k = \frac{1}{b}\sum_{i=(k-1)b+1}^{kb} X_i$.
  Then $\hat{\mu} = \frac{1}{n/b}\sum_{k=1}^{n/b} \hat{\mu}_k$.
- $\hat{\mu}_1$, $\hat{\mu}_2$, ... is again a Markov chain with a similar CLT.
- Pragmatic approach: hope that the autocovariance is much smaller, so that $\hat{\mu}_1, \hat{\mu}_2, \ldots$ can be treated as an iid sample.
- Then construct confidence intervals using $\frac{1}{n/b} S_b^2$ as estimate of the variance of $\hat{\mu}$, where S_b^2 is the sample variance of $\hat{\mu}_1, \ldots, \hat{\mu}_{n/b}$.
- Note: $\frac{1}{n/b} S_b^2$ tends to underestimate the variance of $\hat{\mu}$ (as we are ignoring terms in (1)).

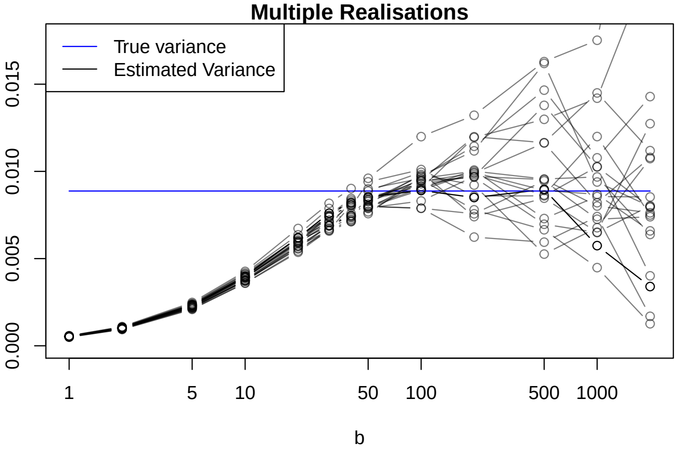
Batch Means - Example AR(1)

- $X_i = 0.9 \cdot X_{i-1} + \epsilon_i$, $\epsilon_i \sim N(0,1)$ independently, $i = 1, \dots, 10000$.



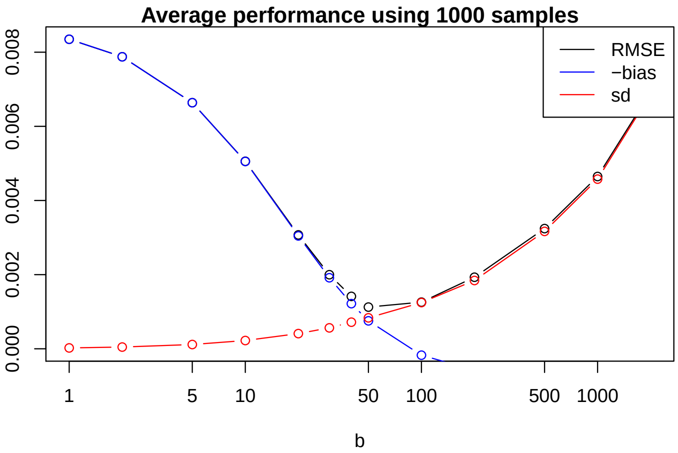
Batch Means - Example AR(1)

- $X_i = 0.9 \cdot X_{i-1} + \epsilon_i$, $\epsilon_i \sim N(0, 1)$ independently, $i = 1, \dots, 10000$.



Batch Means - Example AR(1)

- $X_i = 0.9 \cdot X_{i-1} + \epsilon_i$, $\epsilon_i \sim N(0,1)$ independently, $i = 1, \dots, 10000$.



Comments

- Bias-variance trade-off (small batch size: bias, underestimation of the variance, large batch size: variance).
- The batches can also be taken to be overlapping.
- Other approaches try to estimate the coefficients in (1) directly, see e.g. (Brooks et al., 2011, Section 1.10.2)

Outline

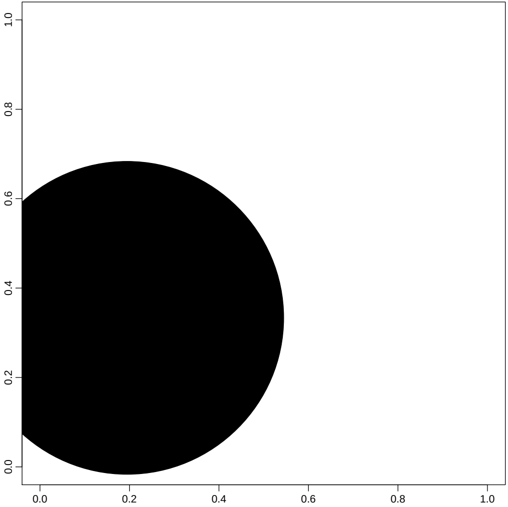
Perfect Sampling - Introduction

- So far: run Markov chain forward
- downside: converge to the stationary distribution only asymptotically
- Perfect Sampling: get a sample from precisely the stationary distribution.
- Methods in this section are not (yet?) in mainstream use

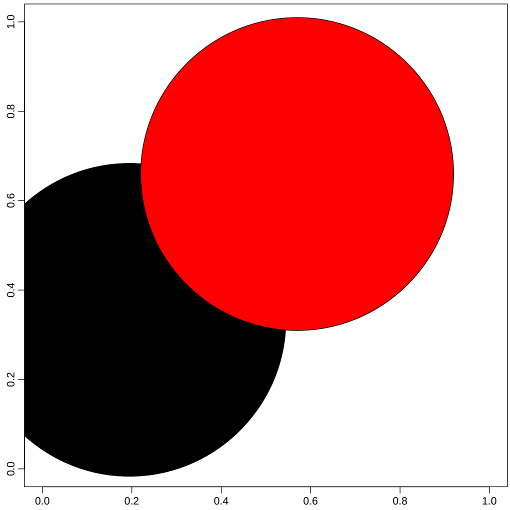
Example - Falling Leaves

- observe the square (0,1)x(0,1)
- leave = circle of radius r=0.35
- centre of falling leaves follows a Poisson distribution (will sample it on (-r,1+r)x(r,1+r))
- Markov chain with state space: leaves seen from the top
- Interested in obtaining a sample from the stationary distribution.

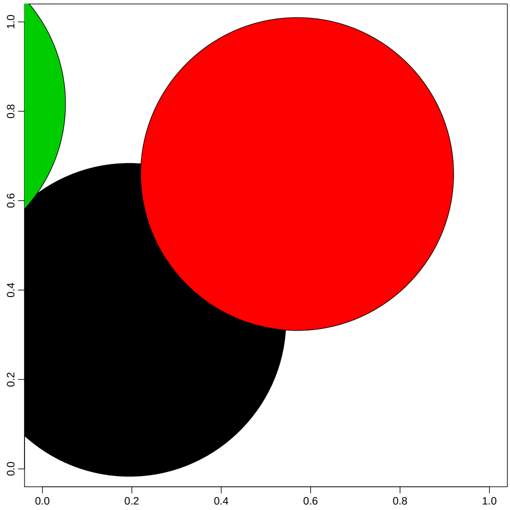
Time running forwards



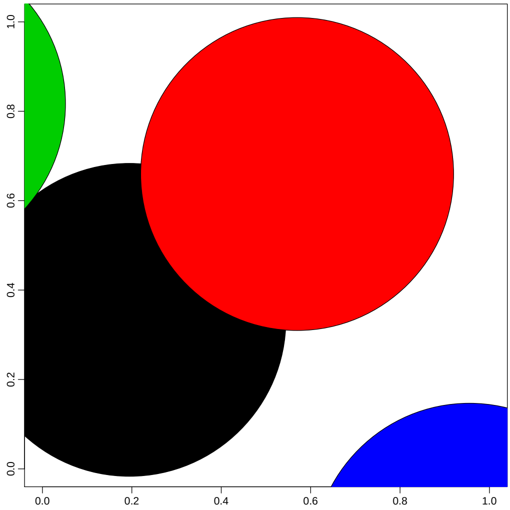
Time running forwards



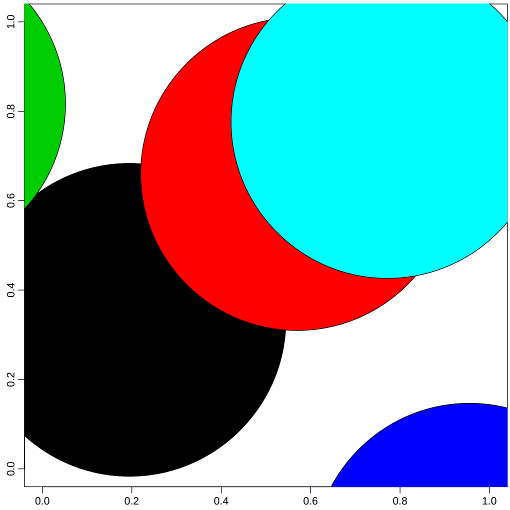
Time running forwards



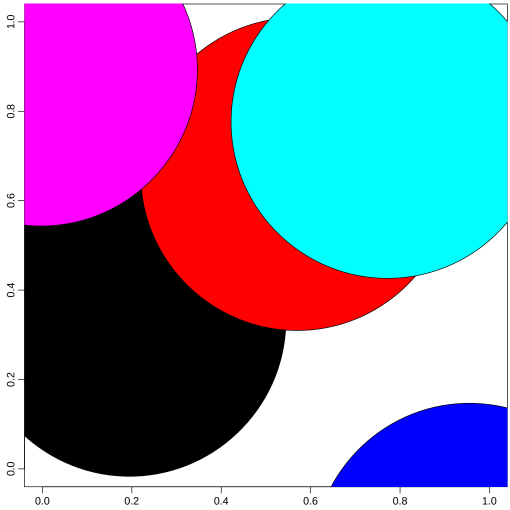
Time running forwards



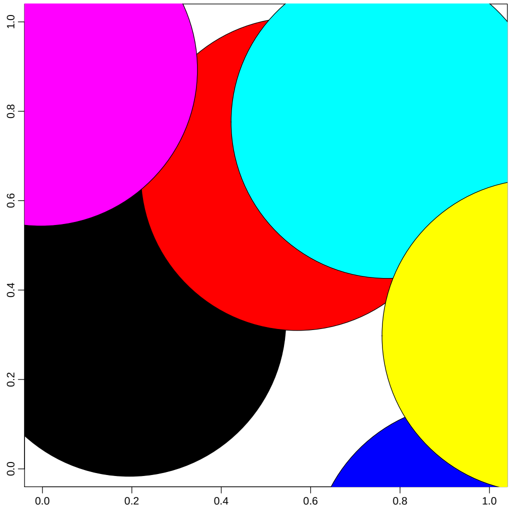
Time running forwards



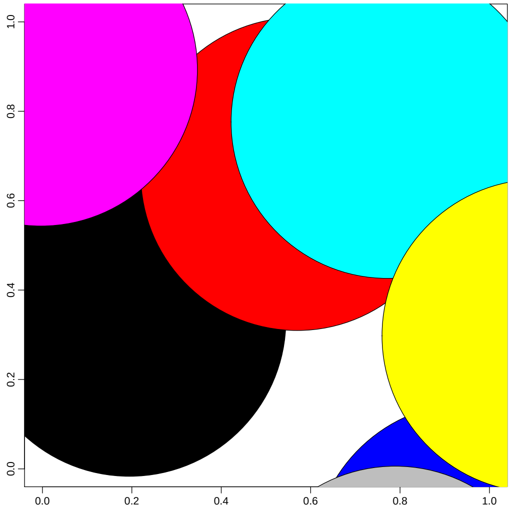
Time running forwards



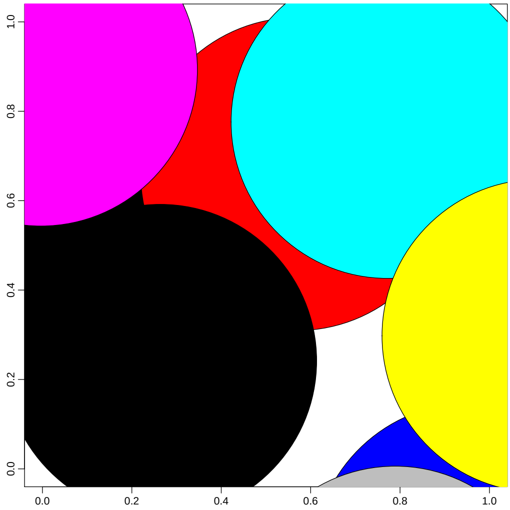
Time running forwards



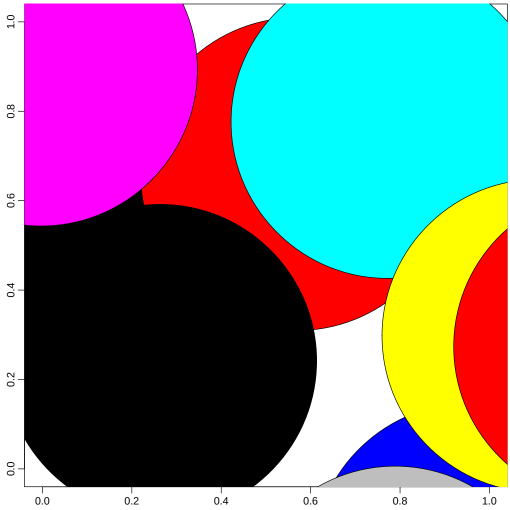
Time running forwards



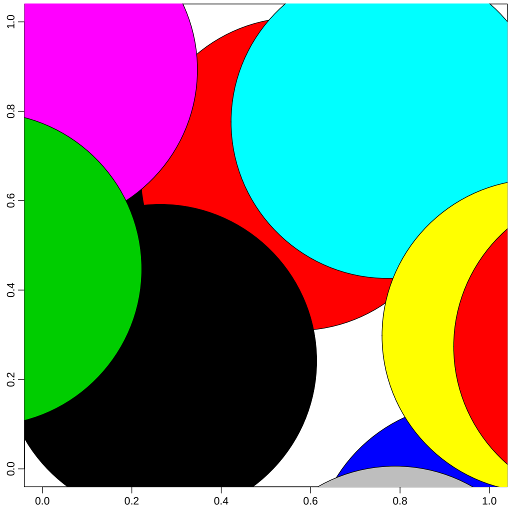
Time running forwards



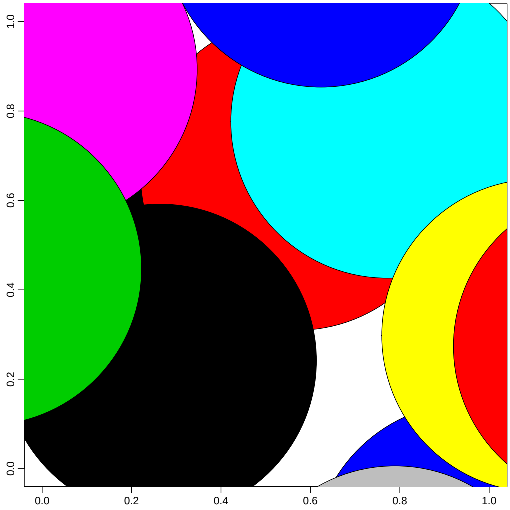
Time running forwards



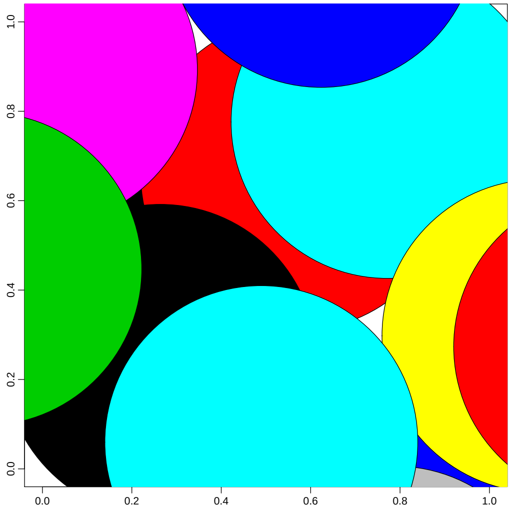
Time running forwards



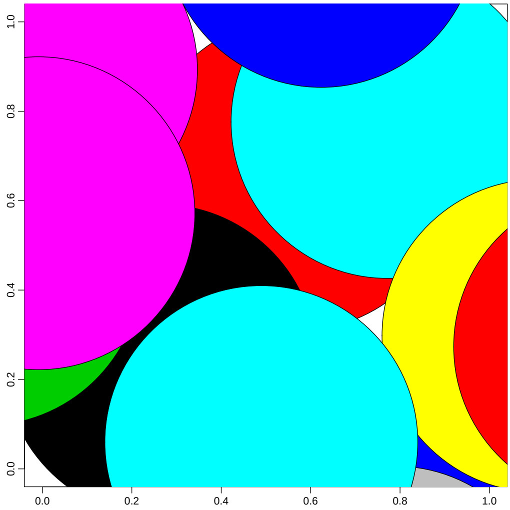
Time running forwards



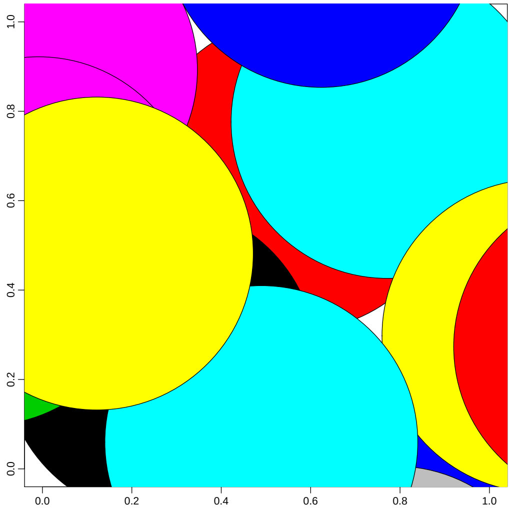
Time running forwards



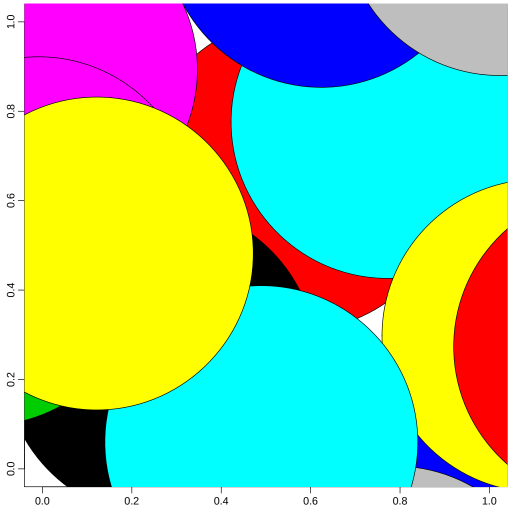
Time running forwards



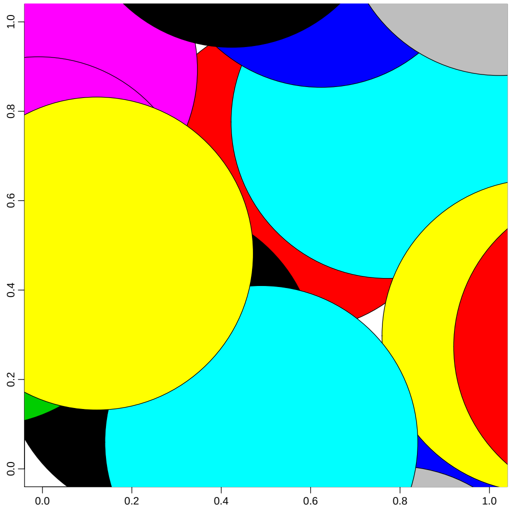
Time running forwards



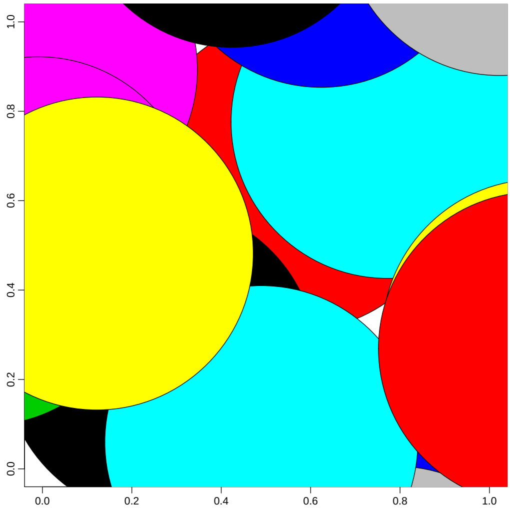
Time running forwards



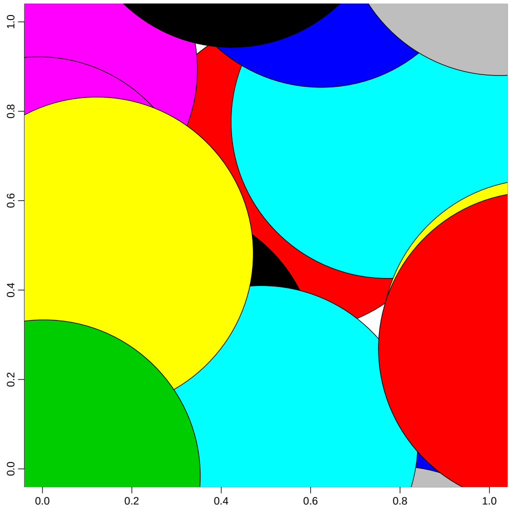
Time running forwards



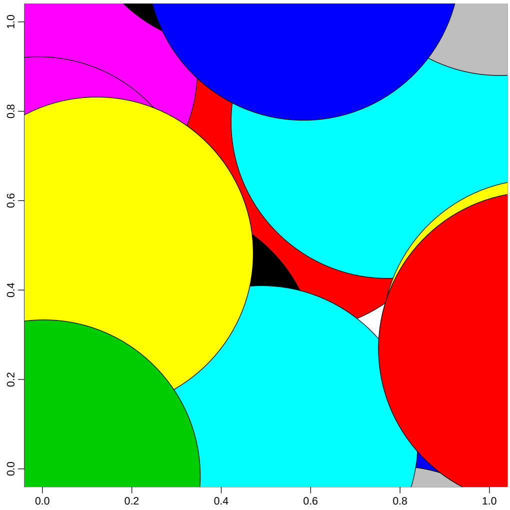
Time running forwards



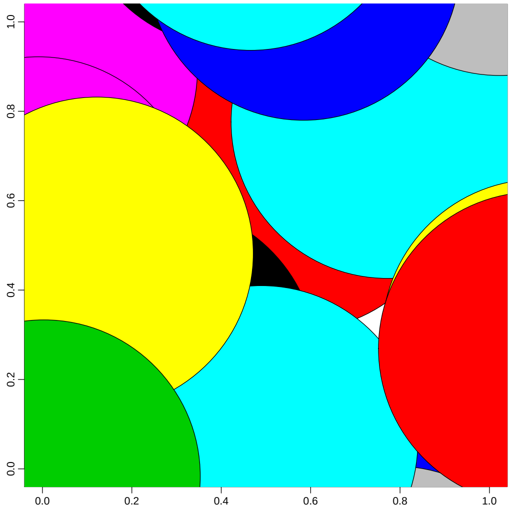
Time running forwards



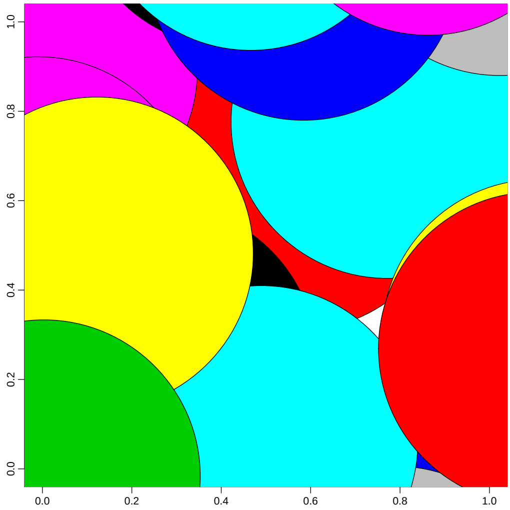
Time running forwards



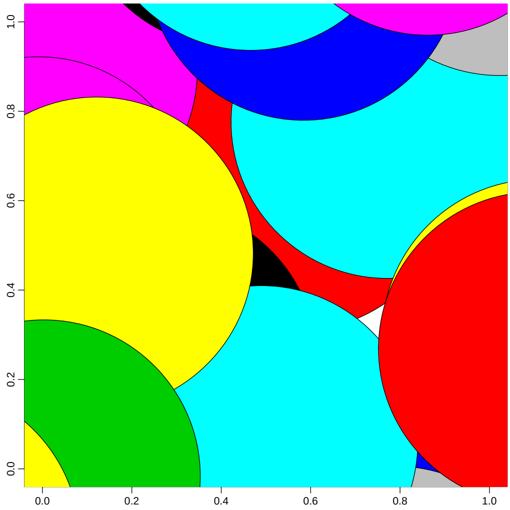
Time running forwards



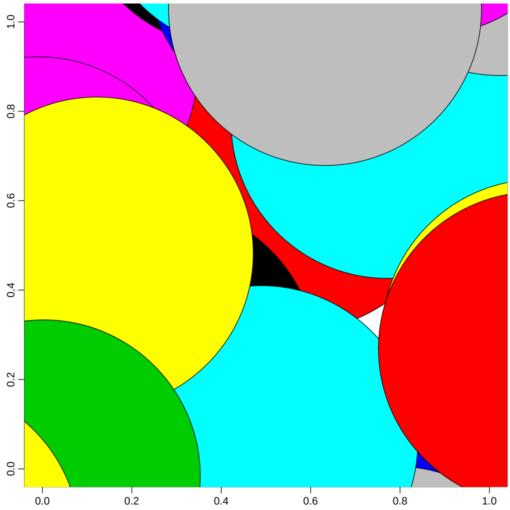
Time running forwards



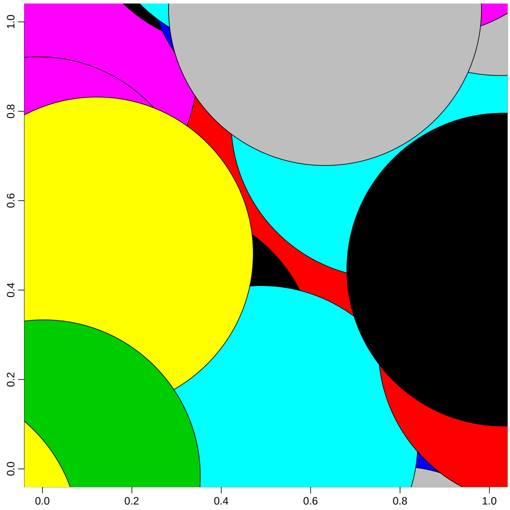
Time running forwards



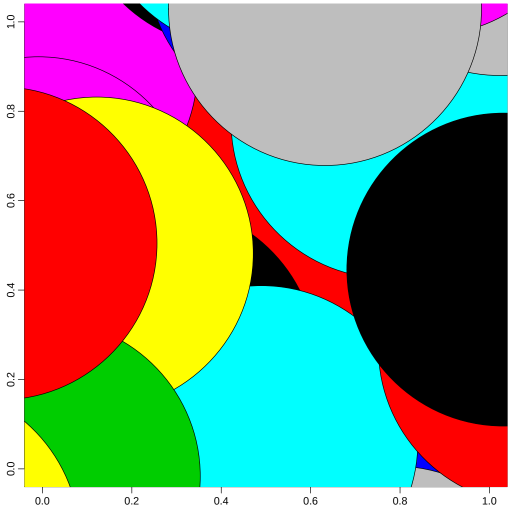
Time running forwards



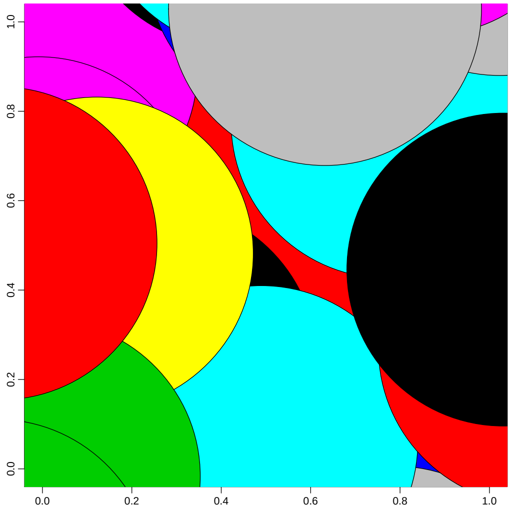
Time running forwards



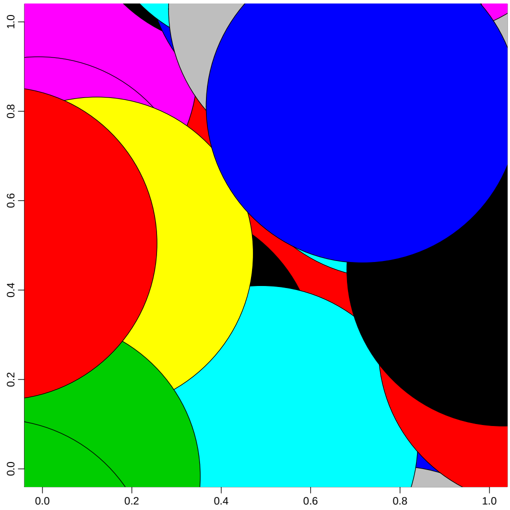
Time running forwards



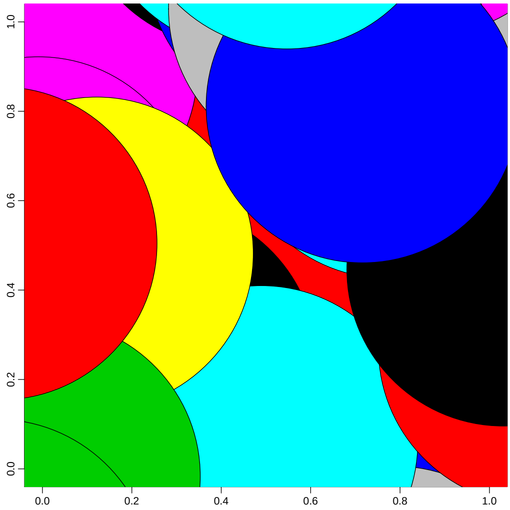
Time running forwards



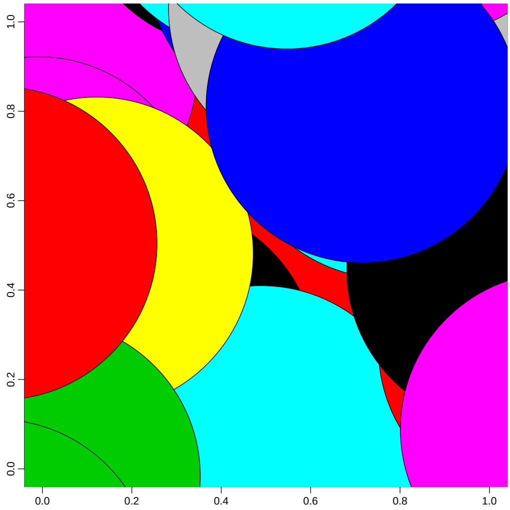
Time running forwards



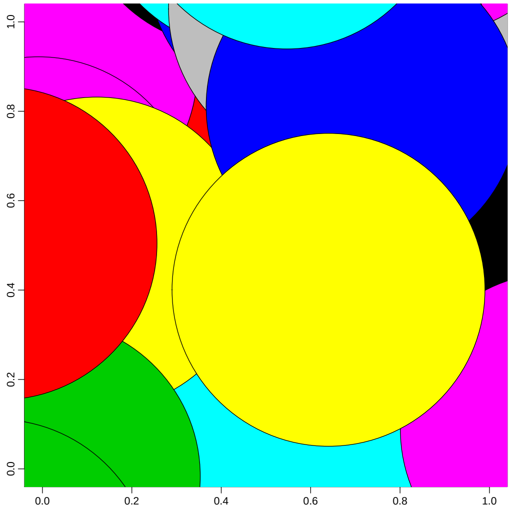
Time running forwards



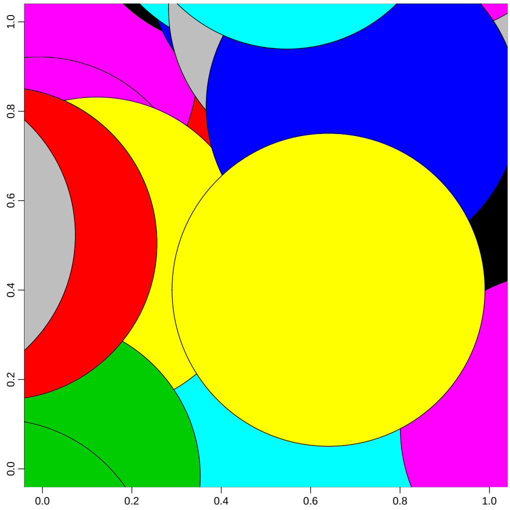
Time running forwards



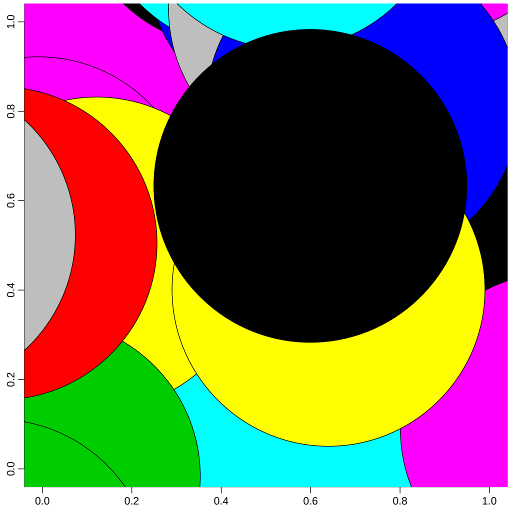
Time running forwards



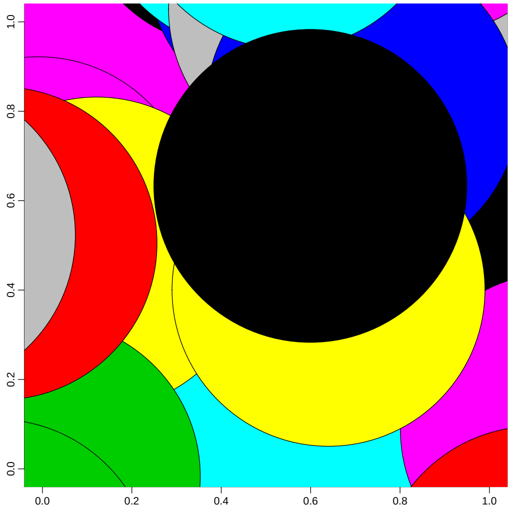
Time running forwards



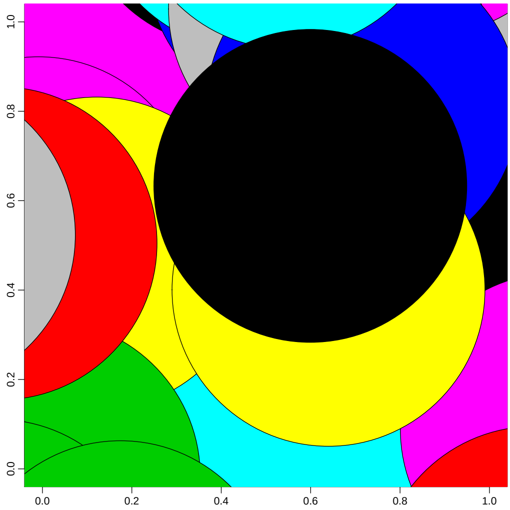
Time running forwards



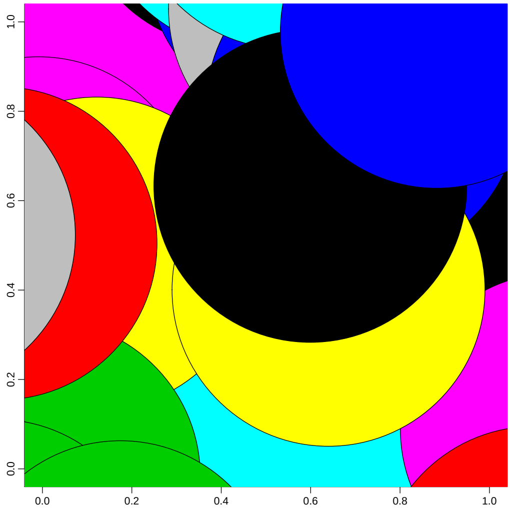
Time running forwards



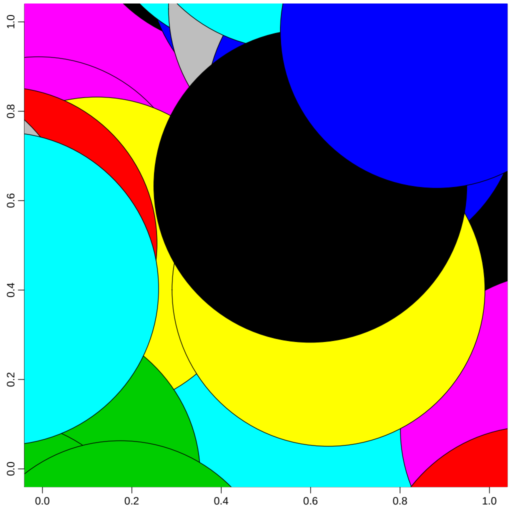
Time running forwards



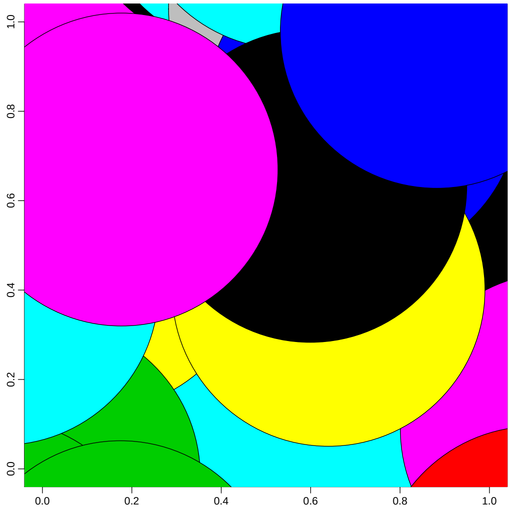
Time running forwards



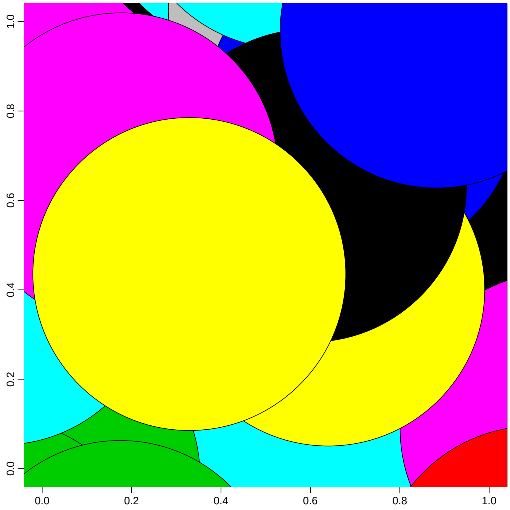
Time running forwards



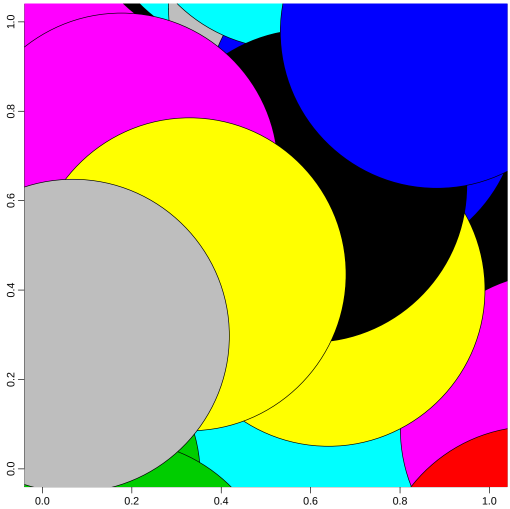
Time running forwards



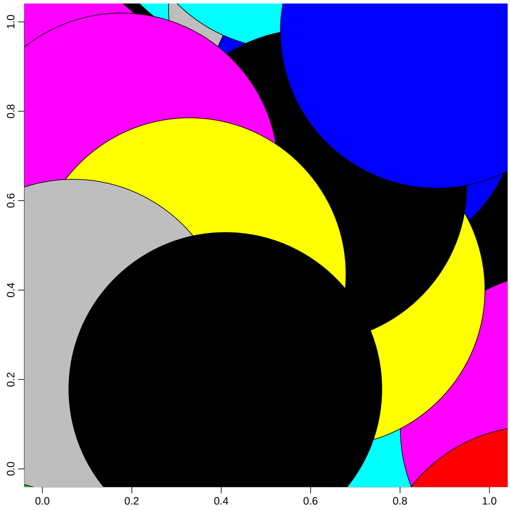
Time running forwards



Time running forwards



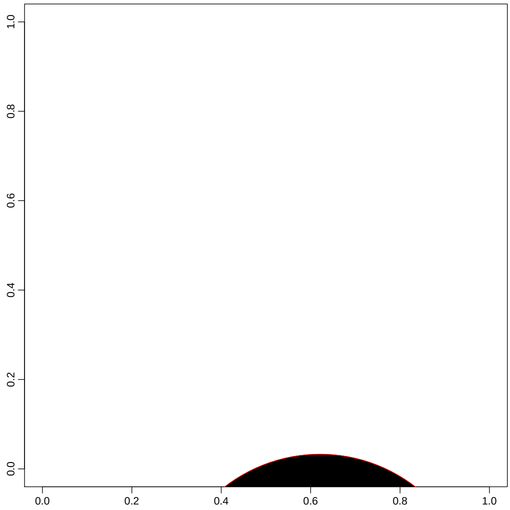
Time running forwards



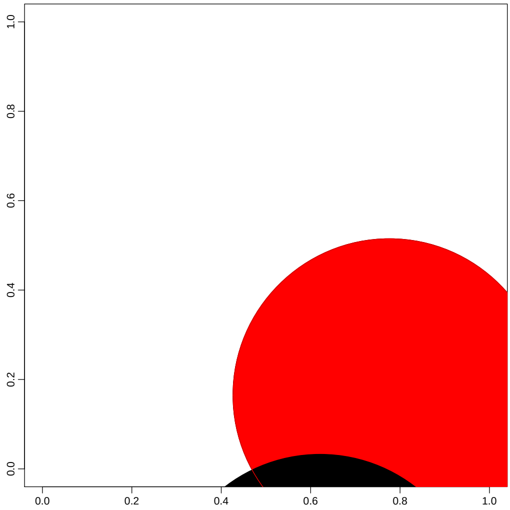
Comments

- Not clear how long to run the chain.
- At best, we can get a sample from an approximation to the distribution of interest.
- This is essentially a problem for all MCMC algorithms so far.

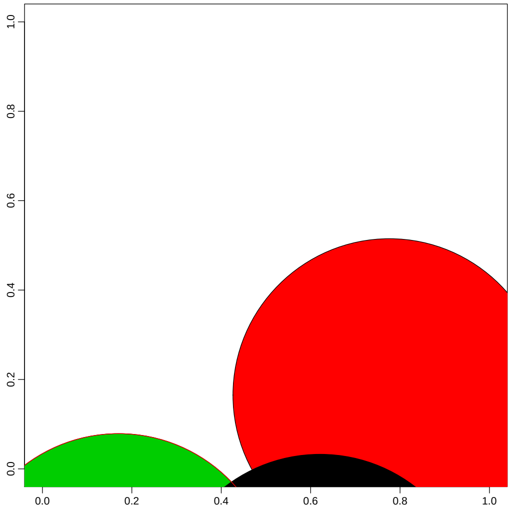
Time running backwards



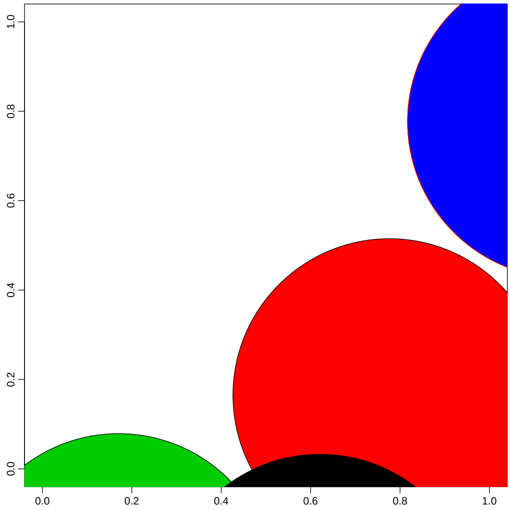
Time running backwards



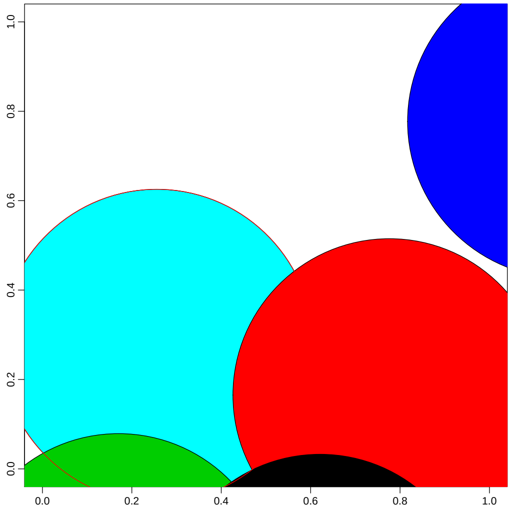
Time running backwards



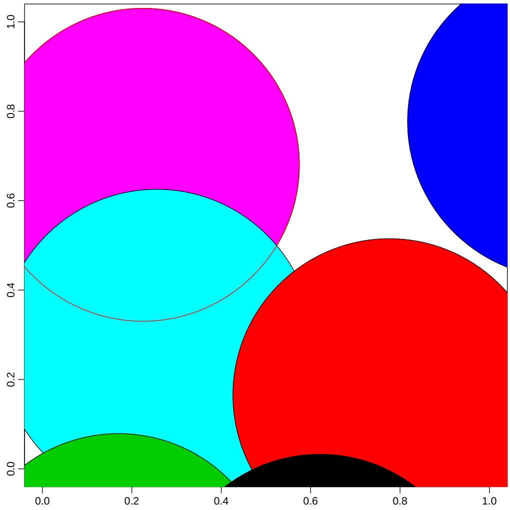
Time running backwards



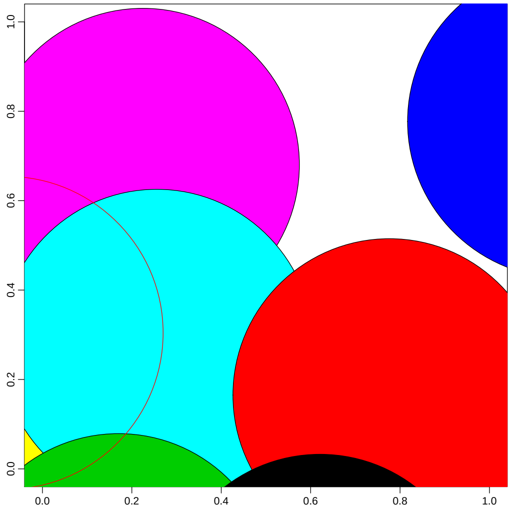
Time running backwards



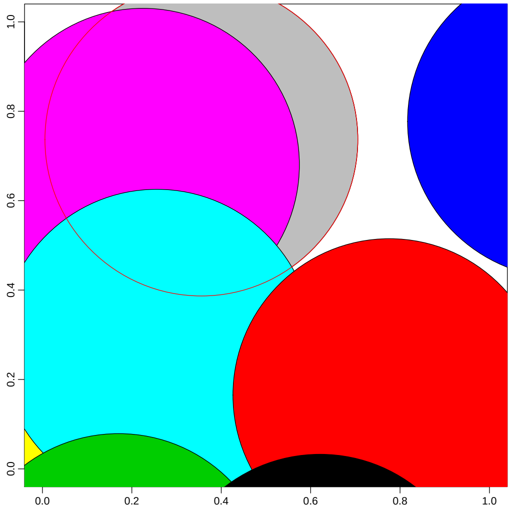
Time running backwards



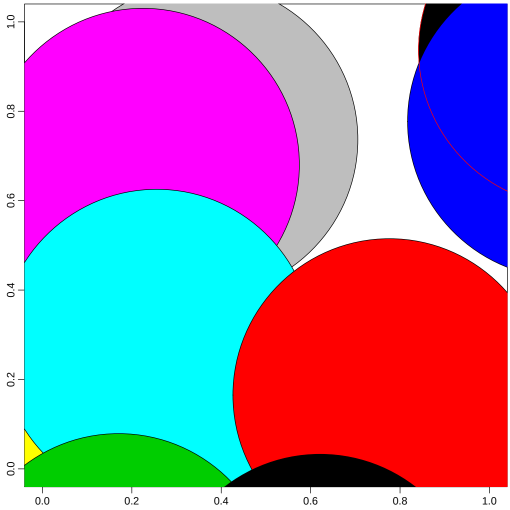
Time running backwards



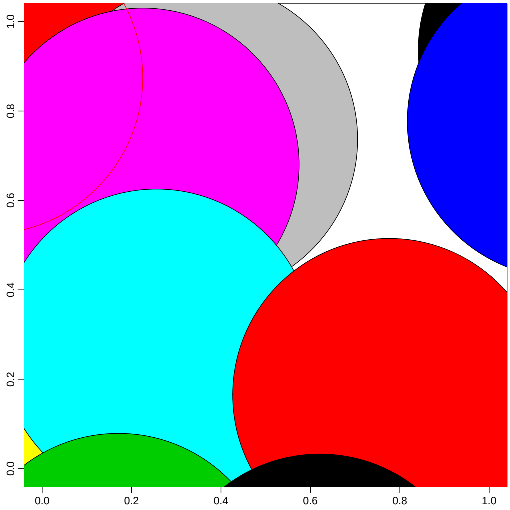
Time running backwards



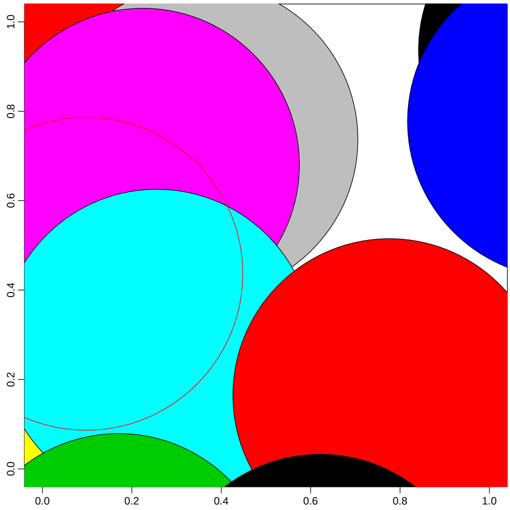
Time running backwards



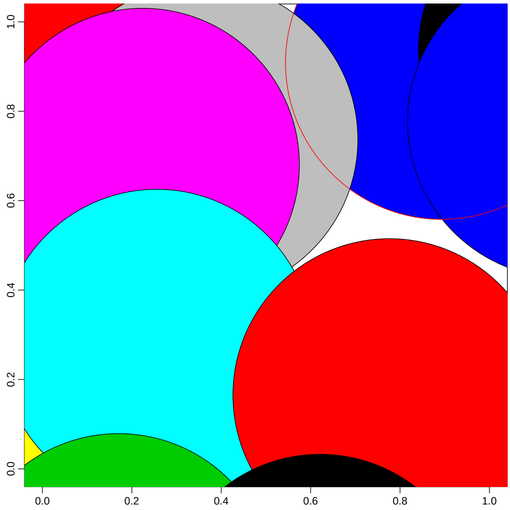
Time running backwards



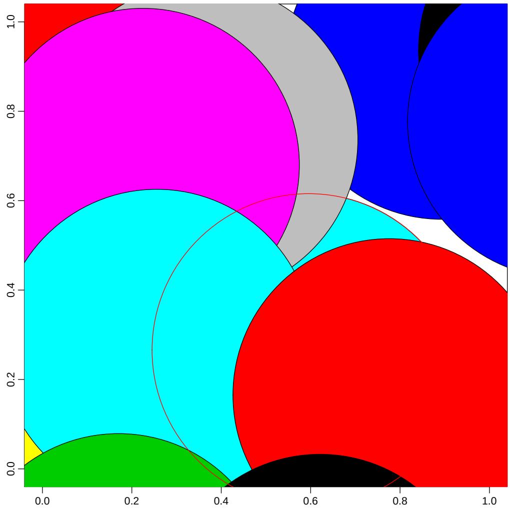
Time running backwards



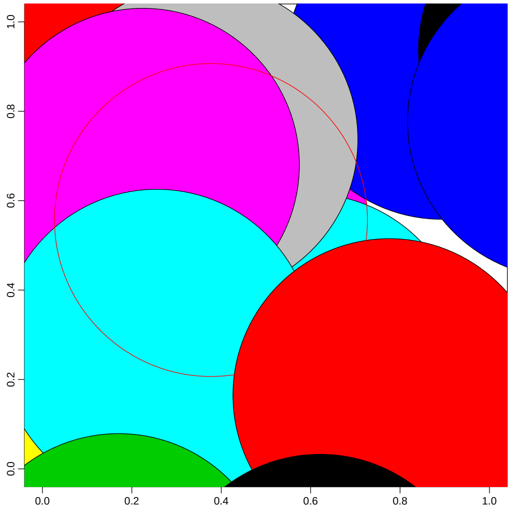
Time running backwards



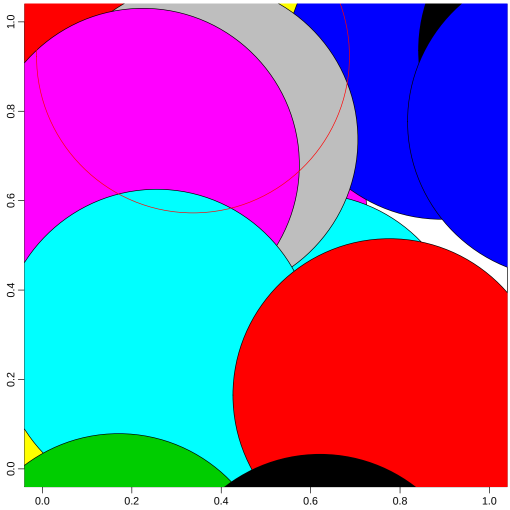
Time running backwards



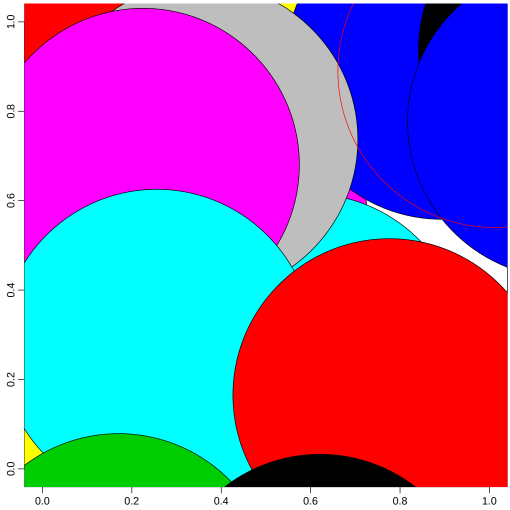
Time running backwards



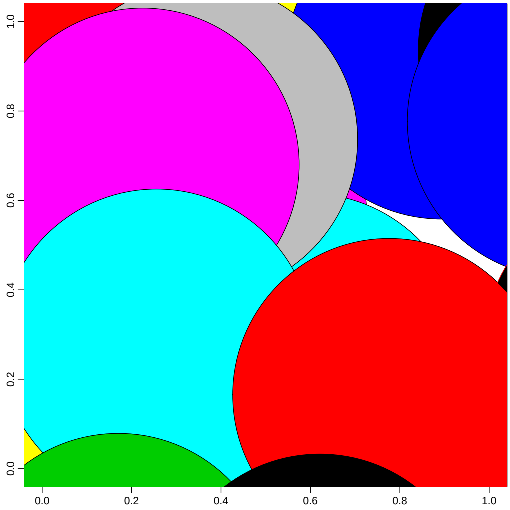
Time running backwards



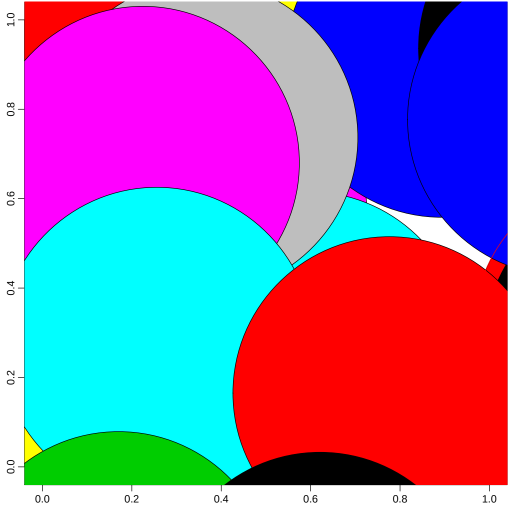
Time running backwards



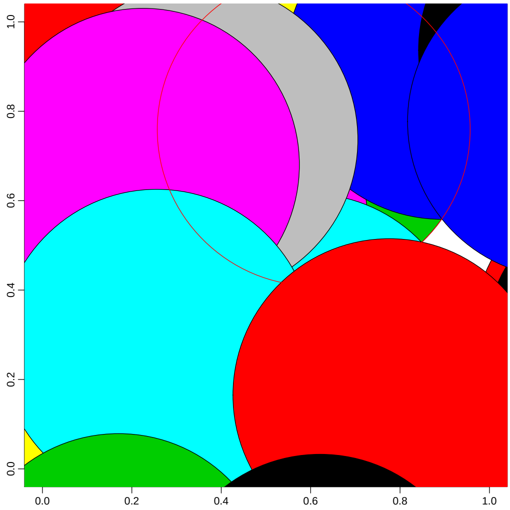
Time running backwards



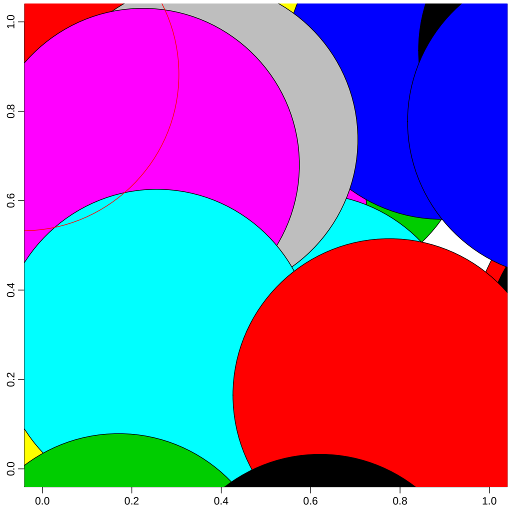
Time running backwards



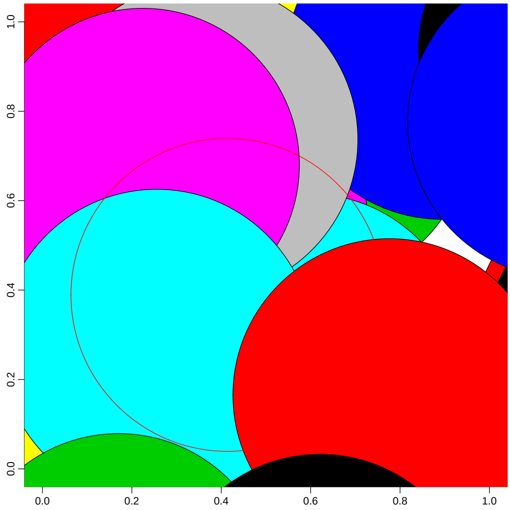
Time running backwards



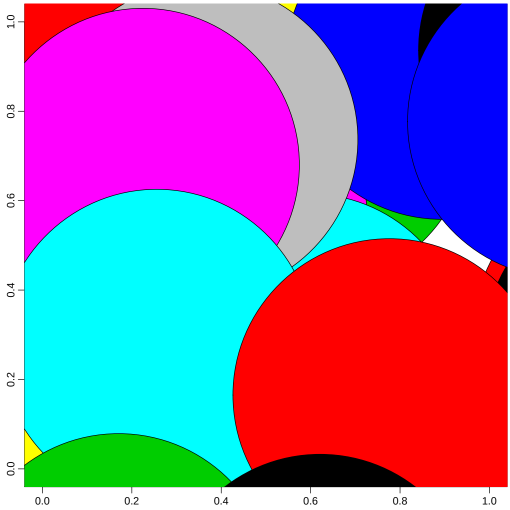
Time running backwards



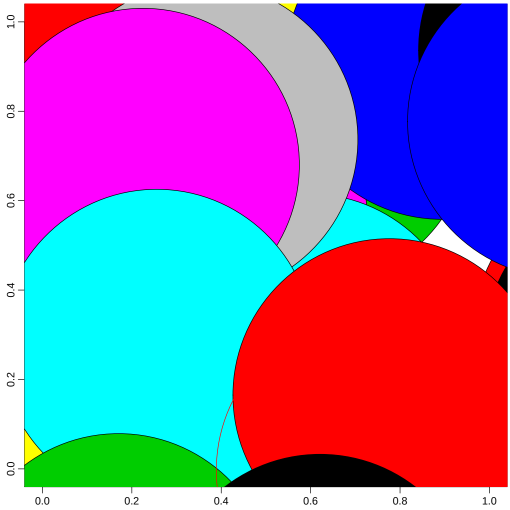
Time running backwards



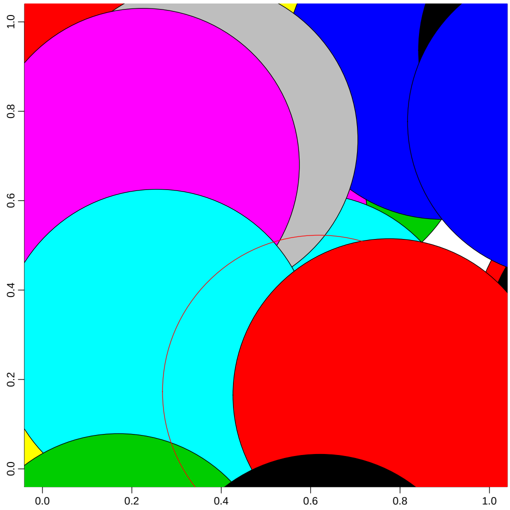
Time running backwards



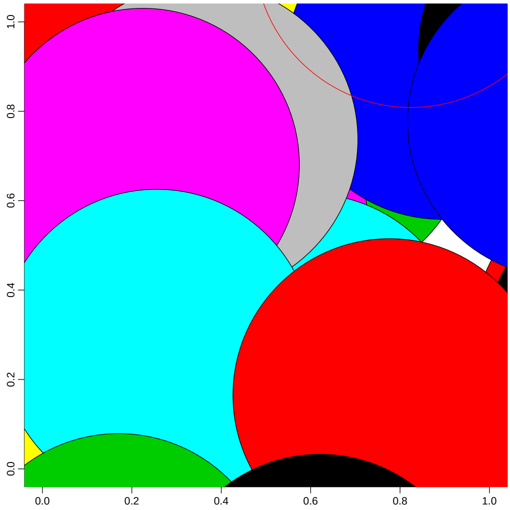
Time running backwards



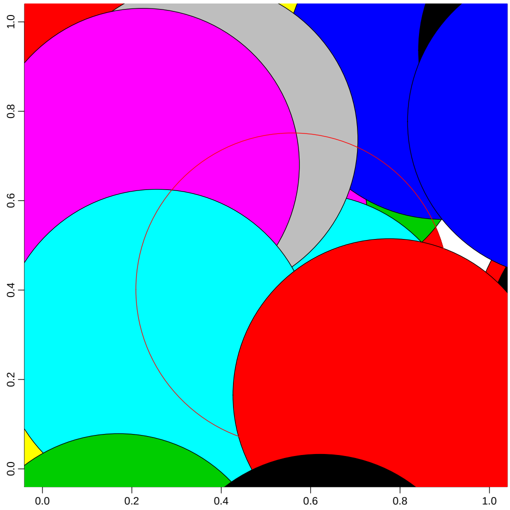
Time running backwards



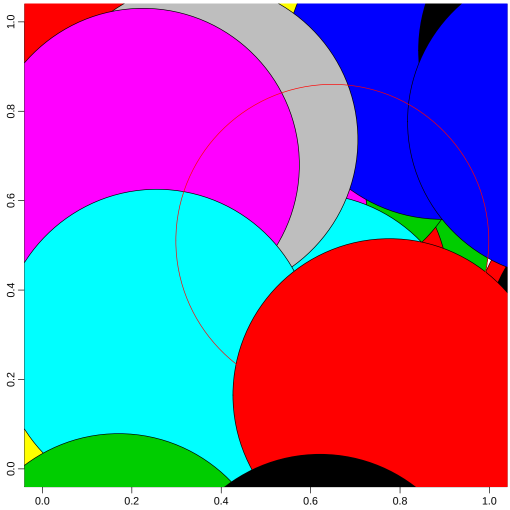
Time running backwards



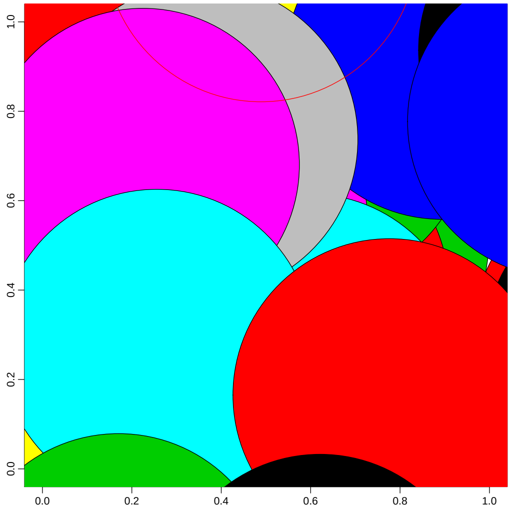
Time running backwards



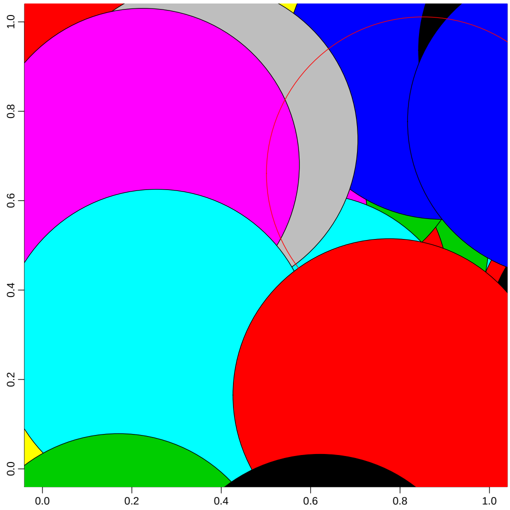
Time running backwards



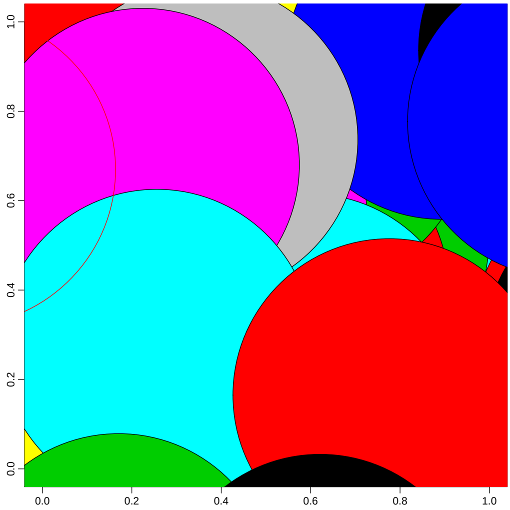
Time running backwards



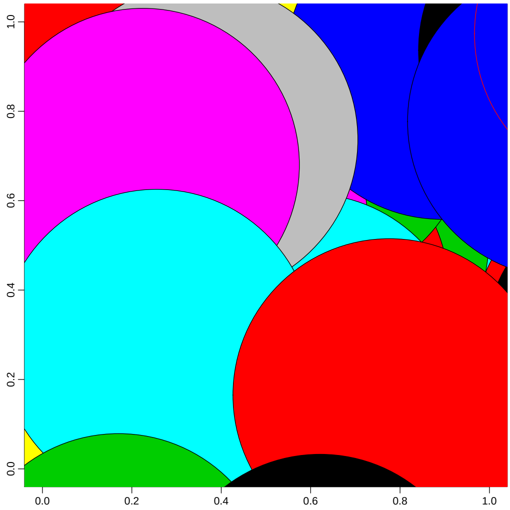
Time running backwards



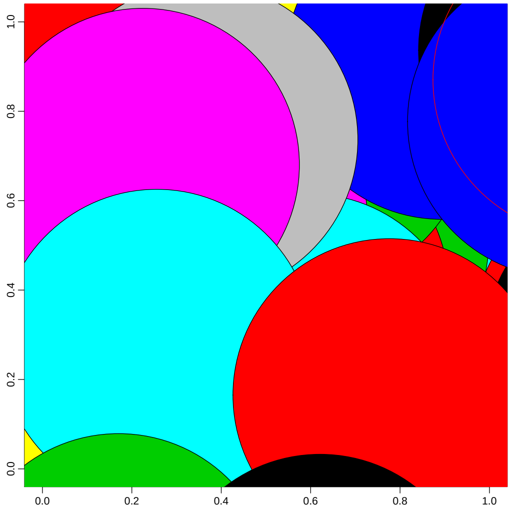
Time running backwards



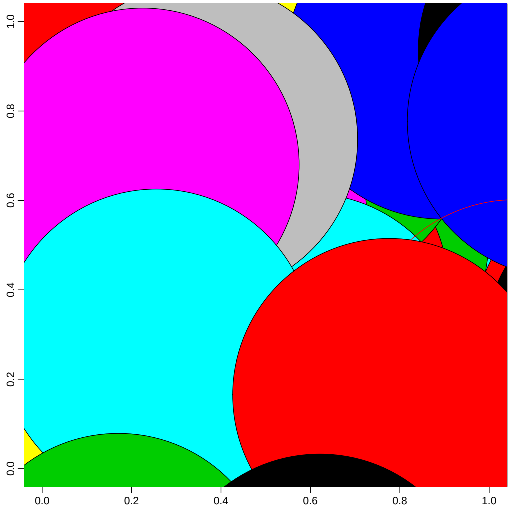
Time running backwards



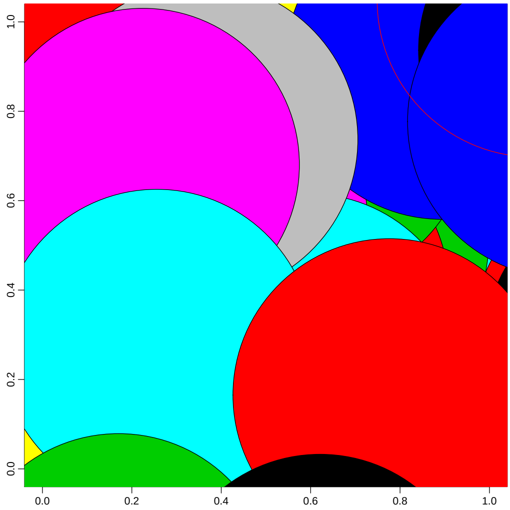
Time running backwards



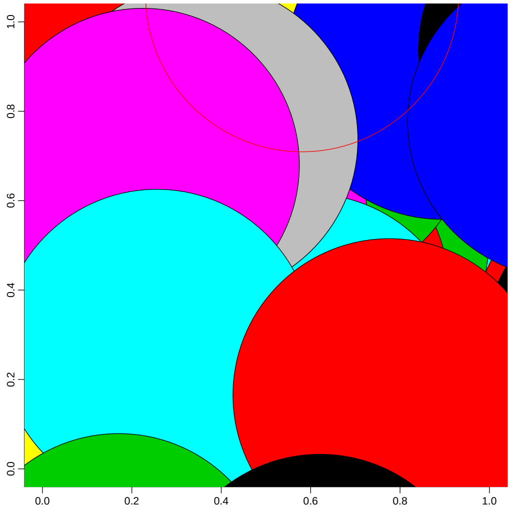
Time running backwards



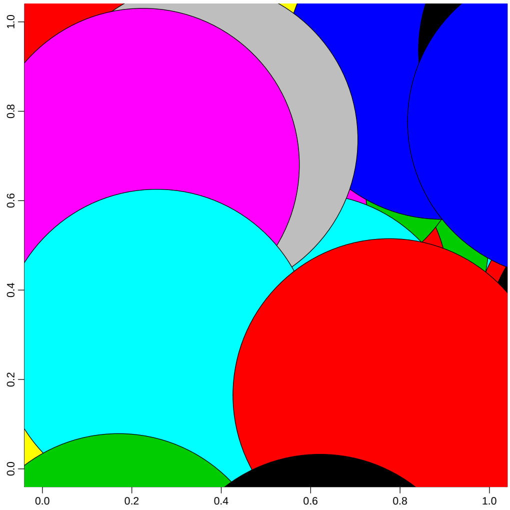
Time running backwards



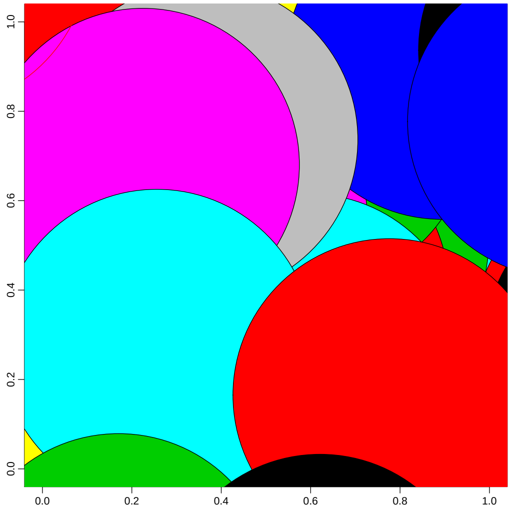
Time running backwards



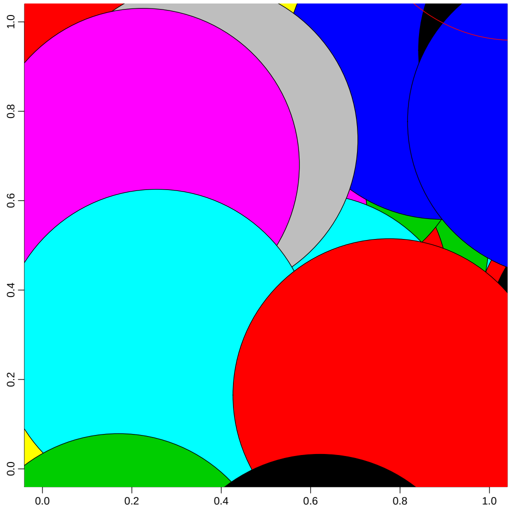
Time running backwards



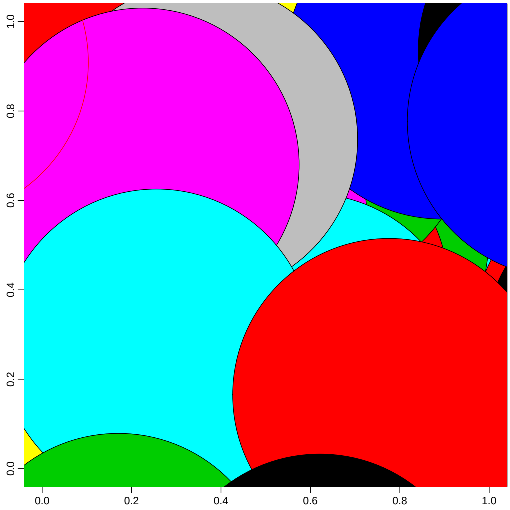
Time running backwards



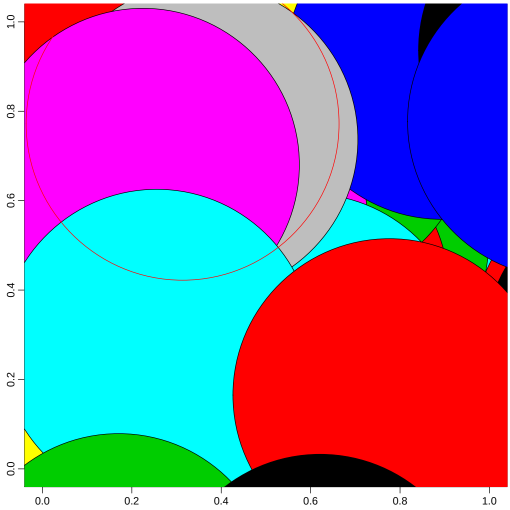
Time running backwards



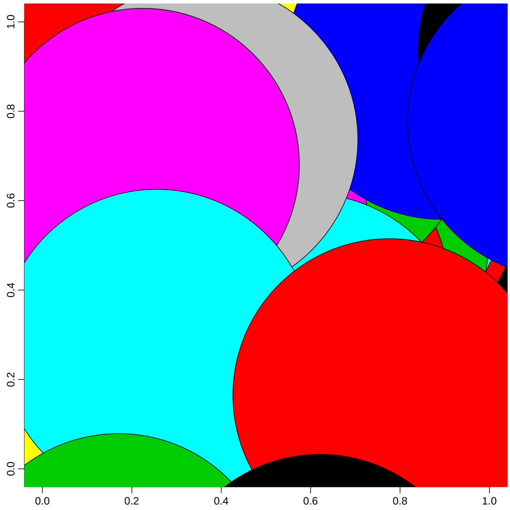
Time running backwards



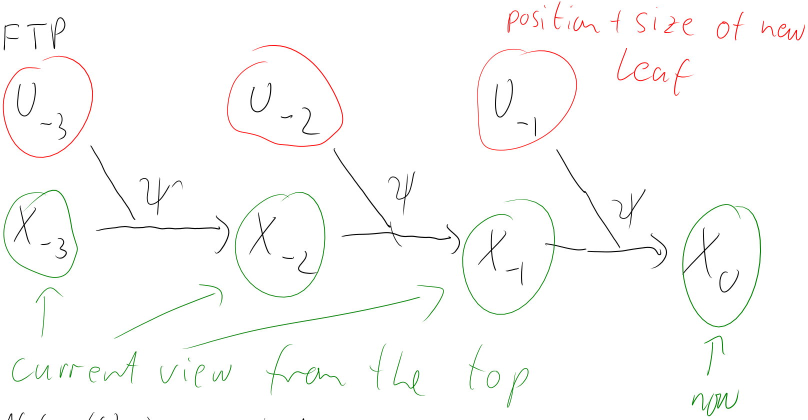
Time running backwards



Time running backwards



CFTP



Note: $(U_t)_{t\in\mathbb{Z}}$ iid

If ψ is suitably constructed, can assume $U_i \sim U[0,1]$

Coupling From the Past (CFTP)

- Propp & Wilson (1996), generates realisations from the stationary distribution of a Markov chain
- The transition of Markov chains can be represented as

$$X_{t+1} = \psi(X_t, U_t)$$

  where U_t are iid.
- Suppose (X_t) is a Markov Chain with stationary distribution f.
- Algorithm:
  - Generate $U_{-1}, U_{-2}, \dots$.
  - Let $\psi_t(\cdot) = \psi(\cdot, u_t)$ and

$$\phi_t(x) = \psi_{-t}(\psi_{-t+1}(\dots\psi_{-1}(x)\dots)$$

  - Determine T such that ϕ_T is constant by looking at $\phi_1, \phi_2, \phi_4, \phi_8, \dots$
  - Take $\phi_T(x)$ (for any x) as a realisation from f.

Using Monotonicity Structure

- Computationally intensive to verify if ϕ_T is constant
- Suppose $\psi(x, u)$ is monotonic in x, i.e.:
  - there exists an ordering $\preceq$ on the state space $\mathcal{X}$ such that $x \leq y \implies \psi(x, u) \leq \psi(y, u)$.
  - there exists a largest element $\overline{x}$ (and a smallest element $\underline{x}$) of $\mathcal{X}$ wrt $\preceq$
- Then it suffices to check if the chains started at $\underline{x}$ and $\overline{x}$ at time $-T$ have coupled before time t, i.e. if $\phi_T(\underline{x}) = \phi_T(\overline{x})$.

Forward Coupling

- Problem with Coupling from the Past:
  Algorithm cannot be interrupted
- Fill (1998) Forward-backward coupling algorithm:
- Main idea: Chain is run backward from a fixed time horizon T (in the future) and an arbitrary starting value X^T
  to time $0 \to X^0$.
- If coupling has occurred between 0 and T then X_0 is the sample otherwise increase T and begin again

Outline

Some Remarks

- MCMC not straightforward to parallelise - approaches
  - Could use parallel chains
  - Could use the conditional structure of the statistical model to parallelise the individual MCMC steps

  Can use parallel chains to facilitate jumps between different modes of the target density.
- Recent extensive treatment of MCMC methods: Brooks et al. (2011)
  (many examples and useful lists of references)
- Overview over R-packages for Bayesian computations:
  `http://cran.r-project.org/web/views/Bayesian.html`

Part I

Appendix

Topics in the coming lectures:

- Bootstrap
- Particle Filtering

References I

Andrieu, C. & Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.* **16**, 1462–1505.

Brooks, S., Gelman, A., Jones, G. L. & Meng, X. L. (eds.) (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Dalal, S. R., Fowlkes, E. B. & Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association* **84**, 945–957.

Fill, J. A. (1998). An interruptible algorithm for perfect sampling via markov chains. *The Annals of Applied Probability* **8**, 131–162.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

References II

Propp, J. G. & Wilson, D. B. (1996). Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.

Robert, C. & Casella, G. (2004). *Monte Carlo Statistical Methods*. Second ed., Springer.

Roberts, G. O., Gelman, A. & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability* **7**, 110–120.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics* **22**, 1701–1728.