Markov chain Monte Carlo

Art B. Owen

Stanford University

Adapted from “Monte Carlo theory, methods and examples”

http://statweb.stanford.edu/~owen/mc/

Outline

1) You can't always sample what you want
2) Sometimes a Markov chain does what you need
3) Review of Markov chains
4) Detailed balance, Metropolis-Hastings and Gibbs
5) Statistical analysis
6) Survey of: ABC, Hamiltonian, Variational

This sets the stage for Jeffrey Rosenthal's MCMC talks.

And one of Gareth Roberts

Apologies

We are skipping:

- 10s of authors
- 100s of ideas
- 1000s of references

Or maybe it's 1000s of ideas and 100s of references

Texts with more info

Gilks, Richardson, Spiegelhalter (1995), Liu (2001), Casella & Robert (2004)

BDA by Gelman, et al., and the Handbook of MCMC

Going deeper

Papers of

- Roberts & Rosenthal
- Diaconis, Wong, Meng, Doucet, Geyer, Andrieu $\cdots$

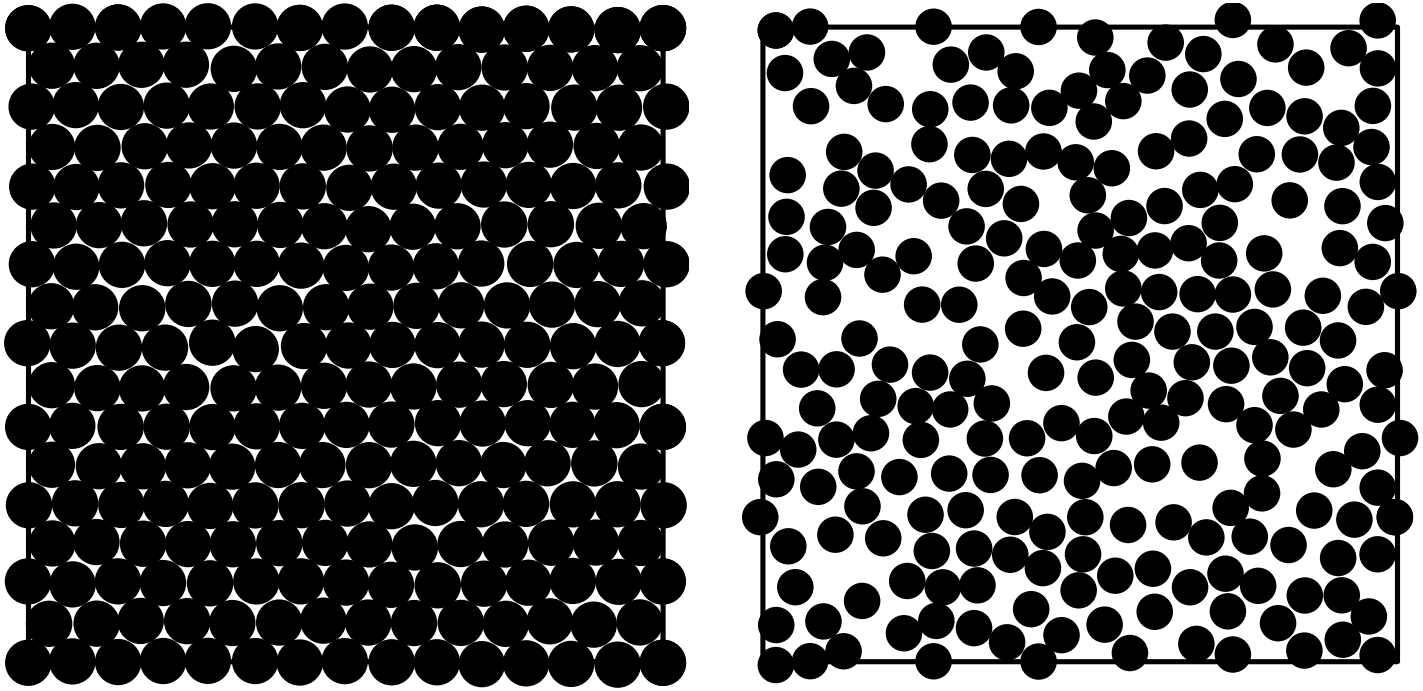
too many to name.

Also, I just ask James Johndrow

Hard shell model

Place N circles in a square.

Uniform conditionally on no overlap.



Hard to sample that IID.

Generally negative dependence is hard.

Bayes

Parameter $\theta \in \Theta$ with prior $\pi(\theta)$. Data $\boldsymbol{y} \sim p(\boldsymbol{y} \mid \theta)$

Posterior

$$\pi(\theta \mid \boldsymbol{y}) = \frac{\pi(\theta) p(\boldsymbol{y} \mid \theta)}{\int \pi(\theta) p(\boldsymbol{y} \mid \theta) \, \mathrm{d}\theta}$$

We want facts about $\theta \mid \boldsymbol{y}$.

We could get those facts by sampling $\theta \sim \pi(\theta \mid \boldsymbol{y})$.

(If we could sample.) Also π is usually unnormalized.

Astronomy example

Ruth Angus at Bayes Comp 2018

$$\pi(\theta \mid \boldsymbol{y}) \propto \pi(\theta) \times p(\boldsymbol{y} \mid \theta)$$

$\boldsymbol{y}$ is data on stars and exoplanets.

$p(\boldsymbol{y} \mid \theta)$ includes:

- propeties about their telescopes
- how many sunspots the stars have
- how fast they spin
- how that speed relates to age of stars
- angles that stars' axes make to us

θ is about whether stars gain or lose exoplanets over time.

$\pi(\theta \mid \boldsymbol{y})$ is not a named distribution. Not in Devroye (1986).

What we will do

First we change θ to $\boldsymbol{x}$. [Revert back later]

We want to sample $\boldsymbol{x} \sim \pi$ (which has everything we know).

We now want $\int f(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for various choices of f.

Then we will

- Devise a Markov chain $\boldsymbol{x}_i$ with stationary distribution π
- and sample $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ from P
- and use $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)$.

What could go wrong?

Markov chains

Refresher. See Norris (1998) or Levin, Peres, Wilmer (2009)

$$\mathbb{P}(X_i \in A \mid X_0 = x_0, \ldots, X_{i-1} = x_{i-1}) = \mathbb{P}(X_i \in A \mid X_{i-1} = x_{i-1})$$

We will work mostly with $X_i \in \Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$.

[Think discrete; act continuous.]

Homogeneous chain

$$\mathbb{P}(X_{i+1} = y \mid X_i = x) = \mathbb{P}(X_1 = y \mid X_0 = x) \equiv \mathbb{P}(x \to y)$$

Distribution of this chain is now determined by

$\mathbb{P}(x \to y)$, and

$p_0(x) = \mathbb{P}(X_0 = x)$

Where does this chain go?

$$p_1(\omega_k) \equiv \mathbb{P}(X_i = \omega_k) = \sum_{j=1}^{M} p_0(\omega_j) P(\omega_j \to \omega_k), \quad \text{i.e.,}$$

$$\boldsymbol{p}_1 = \boldsymbol{p}_0 P$$

$$\boldsymbol{p}_n = \boldsymbol{p}_0 P^n$$

Note

$\boldsymbol{p}_i = (\mathbb{P}(X_i = \omega_1), \ldots, \mathbb{P}(X_i = \omega_M)) \in \mathbb{R}^{1 \times M}$

a row vector.

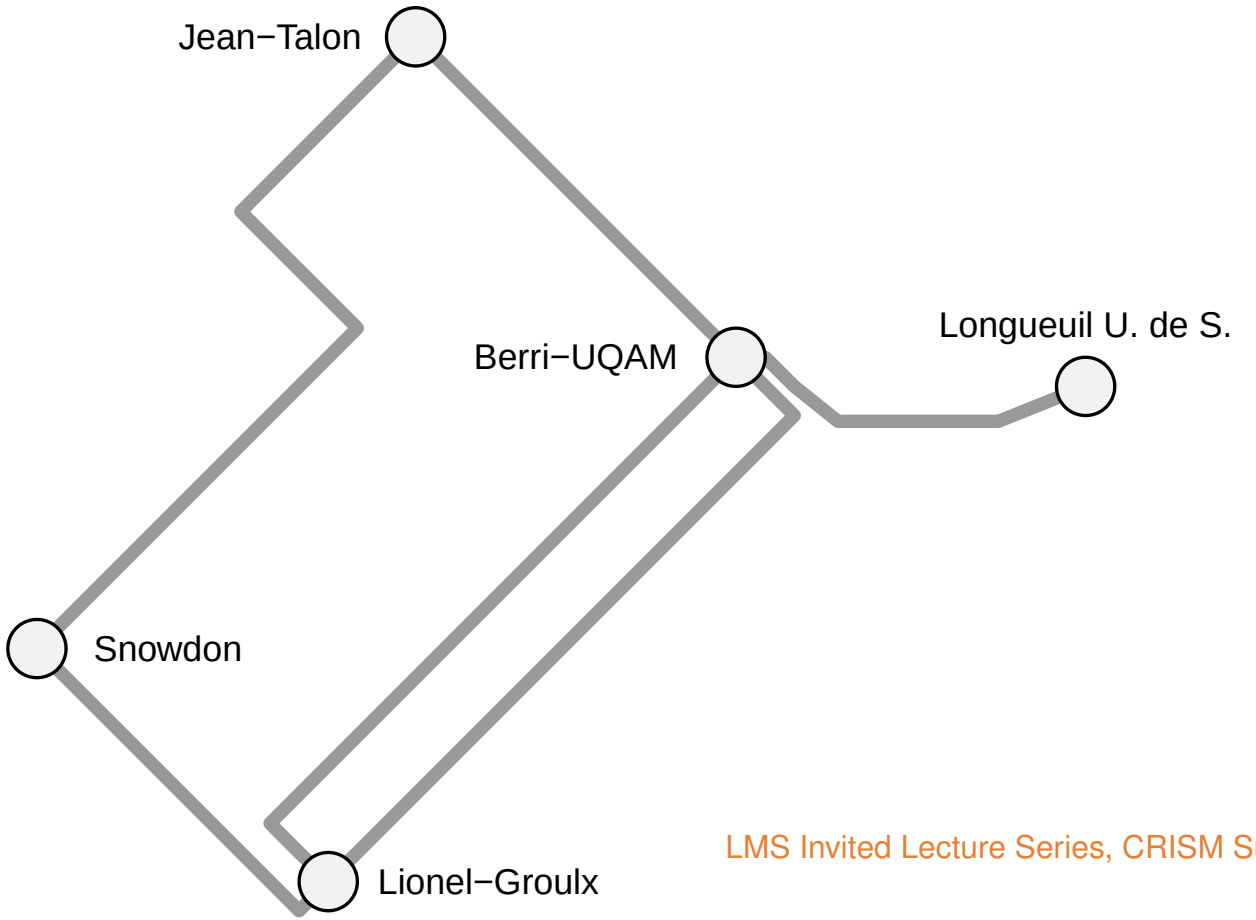
$\boldsymbol{p}_0 P^n \boldsymbol{f}$ is $\mathbb{E}(f(\boldsymbol{x}_n))$, for $f : \Omega \to \mathbb{R}$

A random walk

Follow a uniform random link, but linger with $p = 1/2$ at Longeuil.

(Van Houtte Coffee, 4.7 stars)

A portion of the Montréal métro



Transition matrix

	JT	S	LG	B	L
Jean-Talon	0	$1/2$	0	$1/2$	0
Snowdon	$1/2$	0	$1/2$	0	0
Lionel-Groulx	0	$1/3$	0	$2/3$	0
Berri-UQAM	$1/4$	0	$1/2$	0	$1/4$
Longueuil	0	0	0	$1/2$	$1/2$

For example

$$P\big(\text{Lionel-Groulx} \to \text{Berri-UQAM}\big) = \frac{2}{3}$$

After 100 steps

$$P^{100} \doteq \begin{array}{c} \\ \text{JT} \\ \text{S} \\ \text{LG} \\ \text{B} \\ \text{L} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{JT} & \text{S} & \text{LG} & \text{B} & \text{L} \end{array} \\ \begin{pmatrix} 0.1546 & 0.1530 & 0.2319 & 0.3063 & 0.1541 \\ 0.1530 & 0.1547 & 0.2296 & 0.3091 & 0.1536 \\ 0.1546 & 0.1530 & 0.2319 & 0.3064 & 0.1541 \\ 0.1532 & 0.1546 & 0.2298 & 0.3089 & 0.1536 \\ 0.1541 & 0.1536 & 0.2311 & 0.3073 & 0.1539 \end{pmatrix} \end{array}$$

No matter where you start $p_{100}(\text{Berri}) \doteq 0.31$.

Same for p_{200}, p_{300} et cetera.

These are almost IID from the stationary distribution.

NB

We will not throw out 99% of the data!

Stationary distribution

$\pi = \pi P$ so $\pi^{\mathsf{T}} = P^{\mathsf{T}}\pi^{\mathsf{T}}$

π^{T} is an eigenvector of P^{T} with eigenvalue 1

Perron-Frobenius does it.

Avoid bad chains

$$P_1 = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix} \quad \text{and} \quad P_2 = \begin{pmatrix} 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix}.$$

These both have $\mathbf{U}\{1, 2, 3, 4\}$ as stationary distributions.

P_1 is separable, has multiple stationary distributions, and will get stuck.

P_2 alternates so distribution of X_n does not converge to π.

That is less serious.

Law of large numbers

Let X_i be a time-homogenous Markov chain on a finite set Ω with transition matrix P and stationary distribution π.

If P is irreducible, then

$$\mathbb{P}_{\omega_0}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \sum_{\omega\in\Omega}\pi(\omega)f(\omega)\right) = 1$$

Levin, Peres, Wilmer (2009)

But it could be slow

$$P = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

If $\epsilon \ll 1/n$ then the LLN isn't helping.

What we will do

Given π we will find a transition matrix P with $\pi P = \pi$

Then sample $x_1, x_2, \ldots, x_n$ via P.

Here is what could go wrong

1) It might take a long time before $\boldsymbol{p}_n \approx \pi$ (Slow convergence.)

2) The x_n might get stuck for a long time (Slow mixing)

What helps

For any π there are lots of P to try!

Maybe one will work.

Detailed balance

Stationarity balances flow into y with flow out of y

$$\sum_{x \in \Omega} \pi(x) P(x \to y) = \pi(y) = \sum_{x \in \Omega} \pi(y) P(y \to x)$$

Detailed balance is stronger. Flow $x \to y =$ flow $y \to x$

$$\pi(x) P(x \to y) = \pi(y) P(y \to x) \quad \forall x, y \in \Omega$$

I.e., detailed balance $\implies$ balance.

Reversibility

Suppose that $X_0 \sim \pi$. Then (exercise)

$$\mathbb{P}(x_1 \to x_2 \to \cdots \to x_n) = \mathbb{P}(x_n \to x_{n-1} \to \cdots \to x_1)$$

Notes

- The running chain looks the same forwards or backwards
- Some physical laws/models obey detailed balance
- We needed $X_0 \sim \pi$ for this

Reversibility uses

- P has only real eigenvalues
- Simpler CLT under reversibility Kipnis & Varadhan (1986)
- For uniform walk on graph $\pi(v) \propto \text{degree}(v)$

Metropolis-Hastings

We are going to build a transition matrix P with detailed balance for π.

Propose $x_i \to y_{i+1}$. Accepting means $x_{i+1} = y_{i+1}$.

Rejecting means $x_{i+1} = x_i$. (Stay put.)

Propose and accept

$$Q(x \to y) = \mathbb{P}(\text{propose } y \mid \text{at } x)$$

$$A(x \to y) = \mathbb{P}(\text{accept } y \mid \text{it was proposed from } x)$$

For $x \neq y$

$$P(x \to y) = Q(x \to y) \times A(x \to y)$$

Detailed balance

$$\pi(x)P(x \to y) = \pi(y)P(y \to x)$$

$$\pi(x)Q(x \to y)A(x \to y) = \pi(y)Q(y \to x)A(y \to x)$$

Solve the D.B. equation

$$\pi(x)Q(x \to y)A(x \to y) = \pi(y)Q(y \to x)A(y \to x)$$

We can assume $\pi(x) > 0$ or else the chain would not be there.

So don't start with $\pi(x_0) = 0$!

We can assume $Q(x \to y) > 0$ or we would not have proposed it.

Therefore

$$A(x \to y) = \frac{\pi(y)}{\pi(x)} \frac{Q(y \to x)}{Q(x \to y)} A(y \to x)$$

Scale

If $A(x \to y)$ works so does $\frac{1}{2}A(x \to y)$.

We would prefer $2A(x \to y)$.

But we need $A(x \to y) \leqslant 1$.

Metropolis-Hastings

Maximize probabilities $A(x \to y)$ and $A(y \to x)$ subject to

$$A(x \to y) = \frac{\pi(y)}{\pi(x)} \frac{Q(y \to x)}{Q(x \to y)} A(y \to x)$$

Result

$$A(x \to y) = \min\left(\frac{\pi(y)}{\pi(x)} \frac{Q(y \to x)}{Q(x \to y)}, 1\right)$$

Mnemonics

$$\frac{\pi(y)}{\pi(x)} \implies \text{Moving up hill is good}$$

$$\frac{Q(y \to x)}{Q(x \to y)} \implies \text{But getting stuck there is bad}$$

Unnormalized π

Let $\pi(x) = \pi_u(x)/Z$ for unknown Z.

$$\frac{\pi(y)}{\pi(x)}\frac{Q(y \to x)}{Q(x \to y)} = \frac{\pi_u(y)/Z}{\pi_u(x)/Z}\frac{Q(y \to x)}{Q(x \to y)} = \frac{\pi_u(y)}{\pi_u(x)}\frac{Q(y \to x)}{Q(x \to y)}$$

So we can use π_u instead of π.

Metropolis

Original algorithm Metropolis et al. (1953) had $Q(x \to y) = Q(y \to x)$

$$A(x \to y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right)$$

For hard disk: propose to move one disk.

Accept it unless there is overlap.

Hastings (1970) generalized it.

Peskun's theorem

Let P and $\widetilde{P}$ be irreducible $M \times M$ transition matrices, that both satisfy detailed balance for the same stationary distribution π. Suppose that $\widetilde{P}(x \to y) \leqslant P(x \to y)$ holds for all $x \neq y$. For $i \geqslant 1$, let X_i be sampled from the transition matrix P starting at x_0. Similarly, for $i \geqslant 1$, let $\widetilde{X}_i$ be sampled from the transition matrix $\widetilde{P}$ starting at $\widetilde{x}_0$. Then

$$\lim_{n\to\infty} n\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\Big) \leqslant \lim_{n\to\infty} n\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} f(\widetilde{X}_i)\Big).$$

Peskun (1973)

Upshot

So Hastings was right to maximize $A(x \to y)$.

When we want detailed balance, be creative about $Q(x \to y)$, but use Metropolis-Hastings for $A(x \to y)$.

Estimating μ

The law of large numbers supports:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i)$$

If $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i$ because $\boldsymbol{y}_i$ was rejected **be sure** to count it again.

Metropolis et al. warn repeatedly about this.

Counting the new points only will not sample π.

Burn-in $\equiv$ warmup

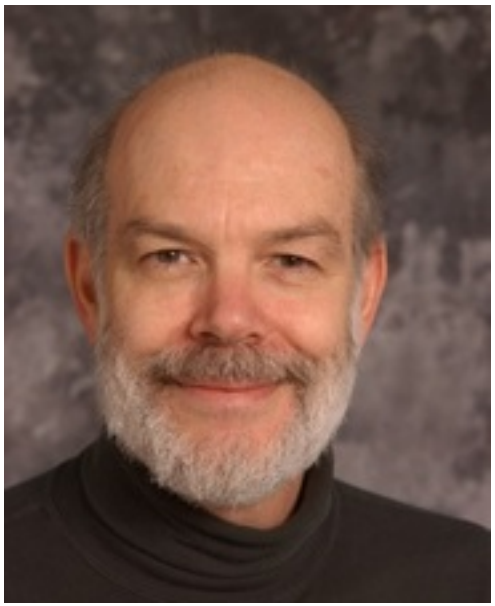
$$\hat{\mu} = \frac{1}{n-b} \sum_{i=b+1}^{n} f(\boldsymbol{x}_i)$$

Skip a few observations. Maybe they're not so close to π.

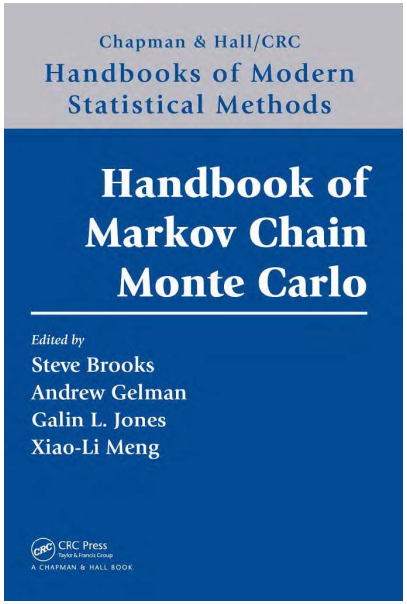
Should we? Yes and no.

About burn-in

Charlie Geyer



Won't throw out any data.



← In this book. →

Andrew Gelman

Likes to use $b = n/2$.

Geyer image from `www.stat.umn.edu`

Book image from `bookspics.com`

Gelman image from `rationallyspeakingpodcast.org`

Variance

Assume $\boldsymbol{x}_i \dot{\sim} \pi$ (e.g., burn-in) then for $Y_i = f(\boldsymbol{x}_i) \in \mathbb{R}$

$$\begin{aligned}\mathrm{Var}(\hat{\mu}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(Y_i, Y_j) \\ &= \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{|i-j|} \\ &\doteq \frac{\sigma^2}{n}\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)\end{aligned}$$

assuming that $|\rho_k|$ decrease. Typically they do, like $O(\rho^k)$ for some $\rho < 1$.

Variance estimation

Can be done by fussy time series analysis.

Or 'batching' Geyer (1992), Fishman

Treat B batch averages of n/B obs as independent.

Thinning

Just use every k'th observation, $k > 1$.

$$\hat{\mu}_k = \frac{1}{n/k} \sum_{i=1}^{n/k} f(\boldsymbol{x}_{k\times i})$$

- Points are usually less dependent than before
- Yet more variance: $\lim_{n\to\infty} \mathrm{Var}(\hat{\mu}_k)/\mathrm{Var}(\hat{\mu}_1) > 1$
- It does save storage.

Efficiency

If it costs 1 unit to advance $\boldsymbol{x}_i \to \boldsymbol{x}_{i+1}$ and $\theta > 0$ units to get $f(\boldsymbol{x}_i)$
Let $R = \sum_{\ell=1}^{\infty} \rho_\ell$ and $R_k = \sum_{\ell=1}^{\infty} \rho_{k\ell}$ and $R_{-k} = R - R_k$.
Thinning pays when

$$\frac{R_{-k}}{k-1} > \frac{1}{\theta+1}\left(R_k + \frac{1}{2}\right)$$

Random walk Metropolis

E.g., $\boldsymbol{y}_{i+1} = \boldsymbol{x}_i + \mathcal{N}(0, \sigma^2 I)$

Or $\boldsymbol{y}_{i+1} = \boldsymbol{x}_i + \mathbf{U}[-\Delta, \Delta]^d$

$$A(\boldsymbol{x} \to \boldsymbol{y}) = \min\left(\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}, 1\right)$$

How large a step?

Tiny step $\implies$ large $\pi(\boldsymbol{y}_{i+1})/\pi(\boldsymbol{x}_i)$ $\implies$ high acceptance

Large step $\implies$ small $\pi(\boldsymbol{y}_{i+1})/\pi(\boldsymbol{x}_i)$ $\implies$ low acceptance

We might have wanted high acceptance **and** large moves.

But there's a tradeoff.

0.234

Default advice:

try step sizes until about 23.4% of proposals are accepted. (Wide range ok)

Why?

Gelman, Roberts, Gilks (1996)

Consider exploring a high dimensional unimodal density,
such as $\pi = \mathcal{N}(0, I_d)$ with $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}, \sigma_d^2 I_d)$,
or $\pi = \mathcal{N}(\mu, \Sigma)$ with $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}, \sigma_d^2 \Sigma)$.

They find the asymptotically optimal σ_d is $2.38/\sqrt{d}$.

It is hard to scale the problem to make $\Sigma = I$. Easy to monitor acceptance rate.

However the optimal σ_d yields 23.4% acceptance as $d \to \infty$
And close to that for $d \geqslant 5$.

Multimodal problems

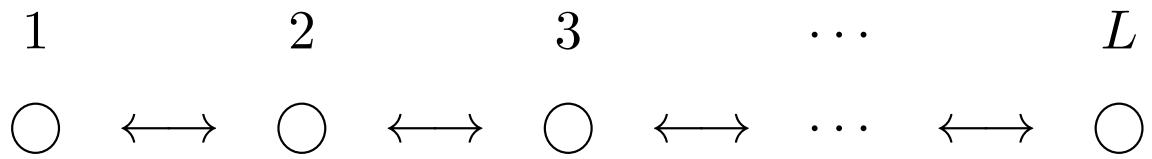
Requires larger steps and lower acceptance.

About that rate

If $\sigma_d = 2.34/\sqrt{d}$ then to cross the support of π takes $L = O(\sqrt{d})$ steps.

The chain is about as likely to go forward as backward.

A random walk on L steps takes $O(L^2)$ time for a round trip.



So it takes about $O(L^2) = O(d)$ steps to go across support of π.

Efficiency

It is about $\dfrac{0.331}{d}$ times as efficient as plain Monte Carlo sampling.

Which we would do if we could.

Improvements on RWM

For $\pi \doteq \mathcal{N}(\mu, \Sigma)$ use sample $\boldsymbol{x}_i$ to estimate Σ.

Then proposals are $\boldsymbol{y}_{i+1} \sim \mathcal{N}(\boldsymbol{x}_i, \lambda\widehat{\Sigma})$. (tune λ too)

Haario, Saksman & Tamminen (2001)

See also Roberts & Rosenthal (2009), Andrieu & Atchadé (2006)

Momentum

It is kind of a pity to endure that random walk behavior.

Hamiltonian MCMC builds in some momentum.

Surveys by Neal (2011), Betancourt (2017)

New Bouncy Particles and Zig-Zag Bierkens, Fernhead, Roberts++

Hamiltonian MCMC is the basis of the **Stan** probabilistic programming language.

Carpenter, Gelman, Hoffman, Daniel, Goodrich, Betancourt, Brubaker, Guo, Li, Riddel (2017)

Stan makes it easy to implement MCMC for Bayes

The Gibbs sampler

For $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$

maybe we can sample one x_j at a time, with the others fixed

"the full conditional of x_j given x_k for $k \neq j$" i.e., x_j given $x_{\neg j}$

Random scan Gibbs

for $i = 1$ to n **do**

$j \sim \mathbf{U}\{1, \ldots, d\}$

$z \sim \pi_{j|\neg j}(\cdot \mid \boldsymbol{x}_{i-1,\neg j})$

$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_{i-1}$

$x_{ij} \leftarrow z$

Deterministic scan

j cycles through $1, \ldots, d$ (repeatedly)

$j = 1 + (i - 1) \bmod d$

Comparison is subtle

Roberts & Rosenthal (2016) 'surprises' paper.

One step of Gibbs

Is a Metropolis-Hastings that always accepts.

Proposal Q just changes component x_j to z

$$\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})} \times \frac{Q(\boldsymbol{y} \to \boldsymbol{x})}{Q(\boldsymbol{x} \to \boldsymbol{y})} = \frac{\pi(\boldsymbol{x}_{\neg j})\pi(z \mid \boldsymbol{x}_{\neg j})}{\pi(\boldsymbol{x}_{\neg j})\pi(x_j \mid \boldsymbol{x}_{\neg j})} \times \frac{\pi(x_j \mid \boldsymbol{x}_{\neg j})}{\pi(z \mid \boldsymbol{x}_{\neg j})} = 1$$

Grouping

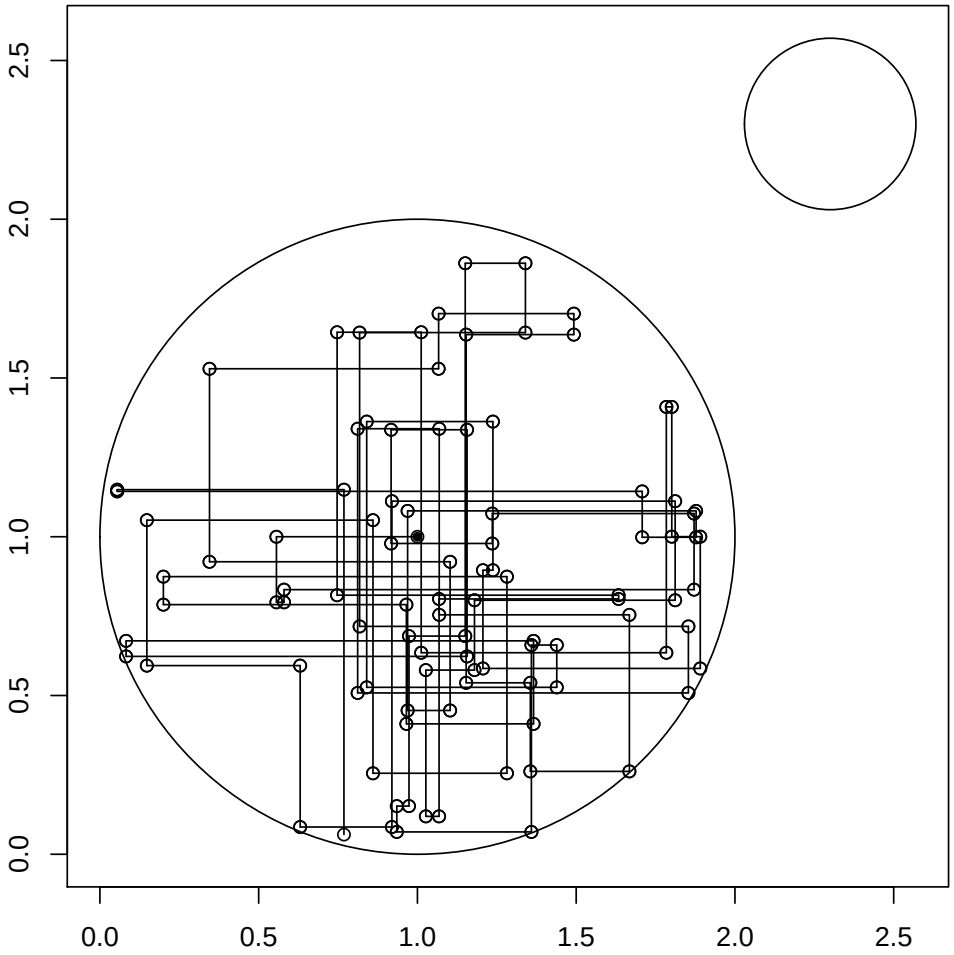
If possible sample x_j in groups.

Metropolis within Gibbs

Use for 'unnamed' $x_j \mid x_{\neg j}$

Used on the original Metropolis et al. paper

Reducible Gibbs



- Uniform in two circles
- Update horizontal then vertical etc.
- We get stuck on Earth
- Never sample the Moon

Gibbs for Bayes

Hierarchical model, mixture of binomials

$$\mathbb{P}(X = x) = \sum_{j=1}^{J} \eta_j \binom{m}{x} p_j^x (1-p_j)^{m-x}, \qquad \eta_j \geqslant 0, \sum_j \eta_j = 1$$

$$\pi(\eta) \propto \prod_{j=1}^{J} \eta_j^{\alpha_j - 1}, \qquad \text{(Dirichlet)}$$

$$p_j \overset{\text{iid}}{\sim} \mathbf{U}(0,1), \qquad \big(\text{Beta}(1,1)\big)$$

Likelihood

$$L = \prod_{i=1}^{n} \left[\binom{m_i}{x_i} \sum_{j=1}^{J} \eta_j p_j^{x_i} (1-p_j)^{m_i - x_i} \right]$$

Posterior

$$\pi(\eta, \boldsymbol{p}) \propto \prod_{j=1}^{J} \eta_j^{\alpha_j - 1} \times \prod_{i=1}^{n} \left[\binom{m_i}{x_i} \sum_{j=1}^{J} \eta_j p_j^{x_i} (1-p_j)^{m_i - x_i} \right]$$

Add (lots of) latent variables

We introduce $Z_{ij} = \begin{cases} 1 & \text{obs } i \text{ from group } j \\ 0 & \text{else} \end{cases}$

Now we want "$\int \pi(\eta, \boldsymbol{p}, \boldsymbol{Z})\, \mathrm{d}\boldsymbol{Z}$" where

$$\pi(\eta, \boldsymbol{p}, \boldsymbol{Z}) \propto \prod_{j=1}^{J} \eta_j^{\alpha_j - 1} \times \prod_{i=1}^{n} \binom{m_i}{x_i} \prod_{j=1}^{J} \left[\eta_j p_j^{x_i} (1 - p_j)^{m_i - x_i} \right]^{Z_{ij}}$$

If we freeze $\boldsymbol{Z}$ and $\boldsymbol{p}$

$$\pi(\eta \mid \boldsymbol{Z}, \boldsymbol{p}) \propto \prod_{j=1}^{J} \eta_j^{\alpha_j - 1} \times \prod_{i=1}^{n} \prod_{j=1}^{J} \eta_j^{Z_{ij}} = \prod_{j=1}^{J} \eta_j^{Z_{\bullet j} + \alpha_j - 1} \qquad \text{(Dirichlet)}$$

The other full conditionals are also named distributions:

- $\boldsymbol{Z}$ has independent multinomials
- $\boldsymbol{p}$ has independent beta

We update $\boldsymbol{Z} \in \{0, 1\}^{n \times J}$, $\boldsymbol{p} \in (0, 1)^J$ and $\eta \in \Delta^{J-1}$ in turn.

Survey

There's (way) more to MCMC than we can do in one hour.

Here are some sketches of interesting ideas.

You may have to make guesses, or read up later.

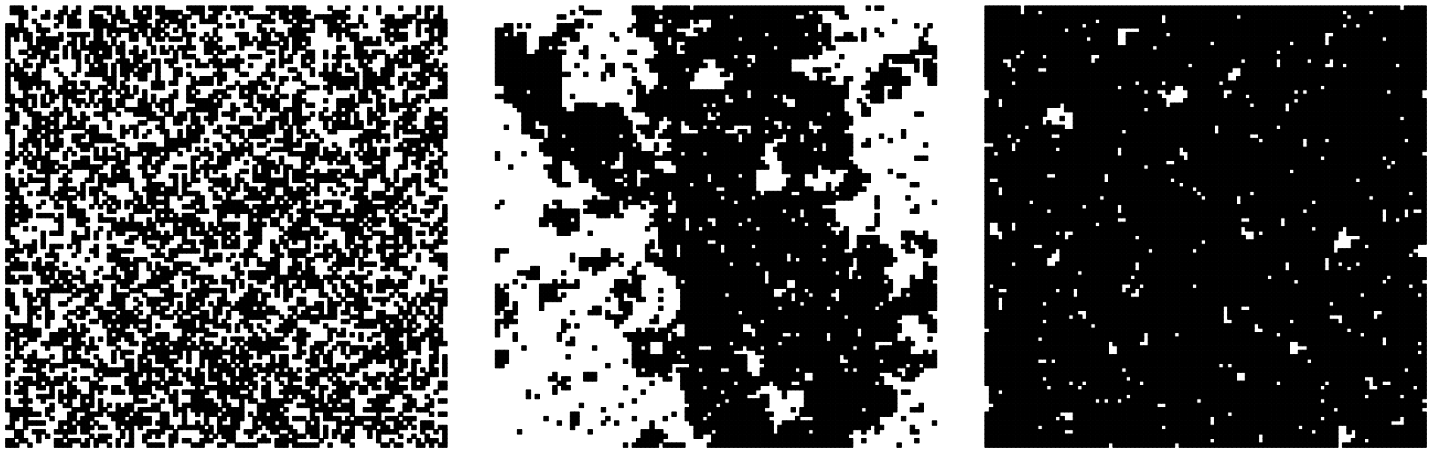
Think of it as vocabulary.

Ising model

Image $\boldsymbol{x} \in \{-1, 1\}^{R \times C}$ with $\pi(\boldsymbol{x}) = \exp(-H(\boldsymbol{x})/T)$ temperature $T > 0$

$$H(\boldsymbol{x}) = -\sum_{j \sim k} x_j x_k$$

Ising model



T = 8.0 T = 2.269 T = 2.0

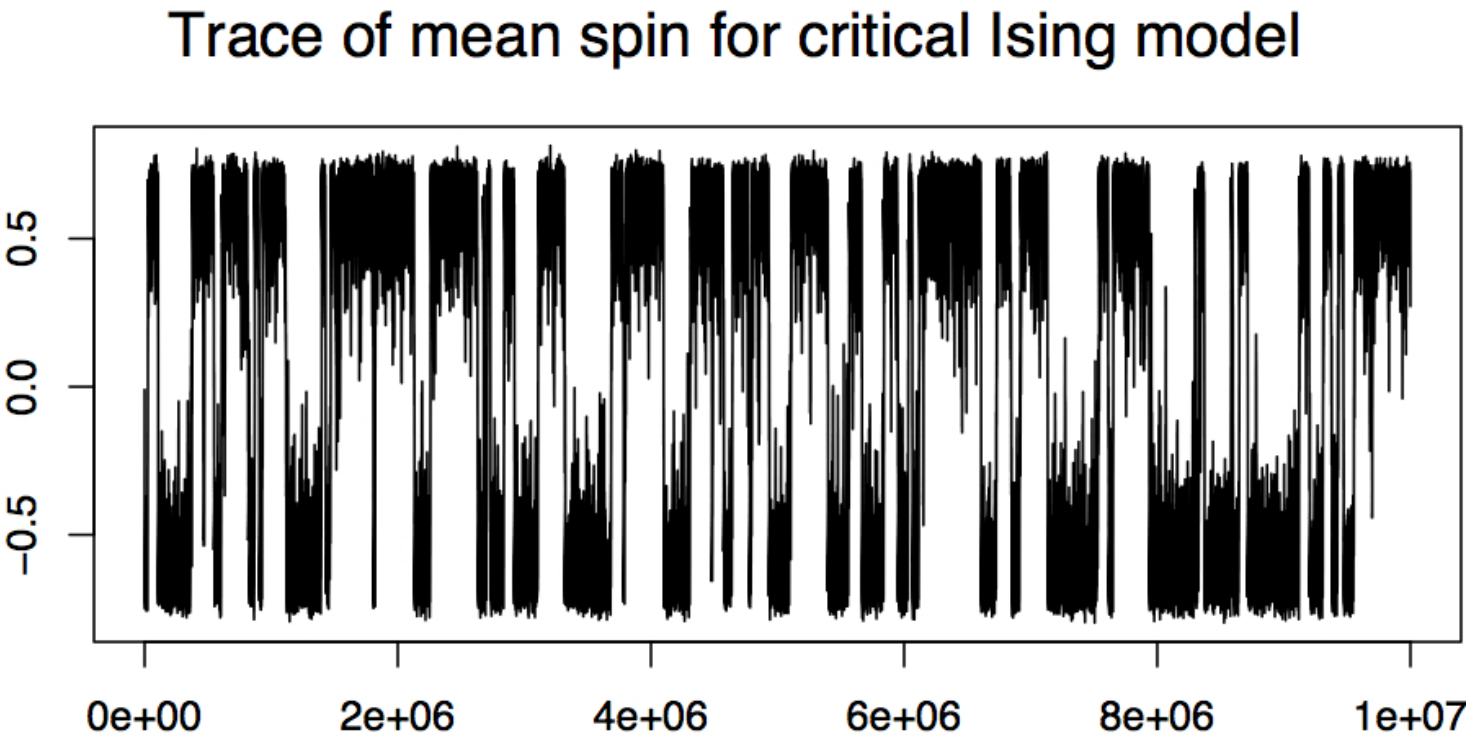
Used in physics (eg magnetism). Besag introduced it to image processing.

Ising model

There are very clever ways to sample the Ising model.

Or we can just flip bits conditionally on their 4 neighbours.

Let's trace mean spin $\frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij}$

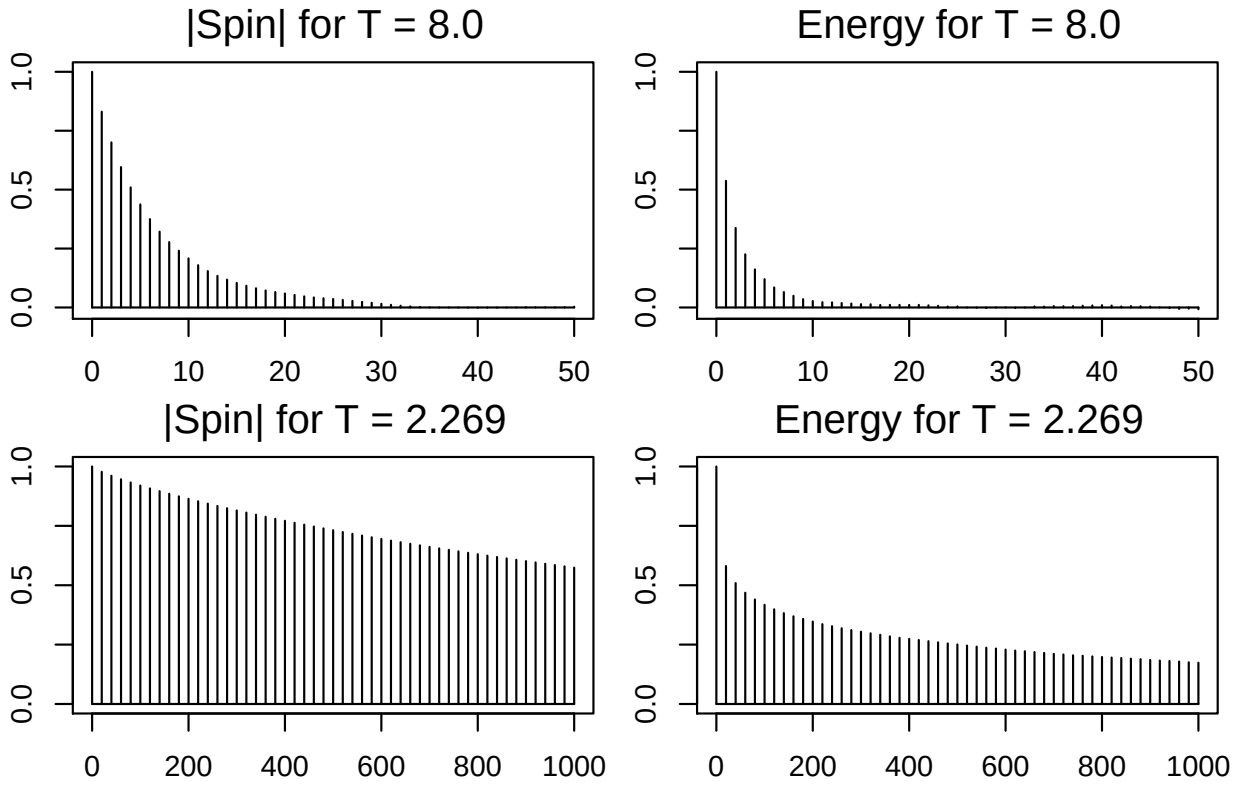


We see it makes a smallish number of round trips.

Autocorrelations

$$\hat{\rho}_k = \frac{1}{n}\sum_{i=1}^{n-k}(Y_i - \hat{\mu})(Y_{i+k} - \hat{\mu})$$

Autocorrelations for the Ising model



Did the chain mix well?

We can use the ACF or a trace.

Also: independent chains and Gelman, Rubin F statistic.

One way diagnostics

Bad ACF	$\Longrightarrow$	No
Good ACF	$\Longrightarrow$	Maybe
Bad trace	$\Longrightarrow$	No
Good trace	$\Longrightarrow$	Maybe
Bad F	$\Longrightarrow$	No
Good F	$\Longrightarrow$	Maybe

We could have missed a chunk of space.

Recent promising work by Gorham & Mackey using Stein discrepancy can provide a “Yes” (but it’s expensive).

Hamiltonian Monte Carlo

(In one slide!) Surveys by Neal (2011), Betancourt (2017)

Whatever $\pi(\theta \mid \boldsymbol{y})$ is we can write it as $\exp(-H(\theta))$.

Introduce momentum ϕ

e.g. $\pi(\theta \mid \boldsymbol{y}) \times \pi(\phi) \propto e^{-H(\theta) - \frac{1}{2}\phi^{\mathsf{T}}\phi} \equiv e^{-H(\theta,\phi)}$

In continuous time

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial \phi} H(\theta, \phi)$$
$$\frac{\partial \phi}{\partial t} = -\frac{\partial}{\partial \theta} H(\theta, \phi)$$

Then comes clever sampling within and between trajectories.

Hamiltonian MCMC is the basis of the **Stan** probabilistic programming language.

Named for Stanislaw Ulam

It automates MCMC for Bayes, especially hierarchical models.

Variational Bayes

Survey by Blei, Kucukelbir, McAuliffe (2017)

Sometimes we cannot even do MCMC on our posterior distribution.

VB approximates $\pi(\boldsymbol{x})$ often using $\pi(\boldsymbol{x}) \approx \prod_j g_j(x_j)$ Independent!

Hopefully it gets **something** right when x_j are dependent

Aimed at enormous problems.

VB underlies the **Edward** probabilistic programming language.

Named for George Edward P. Box

Approx. Bayesian computation

$$\pi(\theta \mid \boldsymbol{x}) \propto \pi(\theta) \times p(\boldsymbol{x} \mid \theta)$$

Sometimes we cannot compute the likelihood $p(\boldsymbol{x} \mid \theta)$.

E.g., θ describes how a colony of bacteria evolves over time, and

$\boldsymbol{x}$ is how it looks right now

A taste of ABC

Loop over i

- Sample $\theta_i \sim \pi(\theta)$.
- Sample $\boldsymbol{x}_i \mid \theta_i$
- Keep $\theta_i \iff \|\boldsymbol{x}_i - \boldsymbol{x}\| \leqslant \epsilon$

Use the retained θ_i

Many variants. Now a whole handbook.

Also ask Christian Robert

Tempering

Replace $\pi(\boldsymbol{x})$ by $\pi(\boldsymbol{x})^{1/T}$ for "temperature" $T > 1$

Large $T \implies$ more uniform $\implies$ better mixing.

Several tactics to connect hot fast mixing chains to target cold ones.

Parallel tempering keeps a vector of $\boldsymbol{x}$'s for temps

$$1 < T_2 < T_3 < \cdots < T_K$$

and uses clever swapping. Geyer (1991), Marinari and Parisi (1992)

Batching

Work on subsamples of IID data.

E.g., Li and Wong

Thanks

- Lecturers: Nicolas Chopin, Mark Huber, Jeffrey Rosenthal
- Guest speakers: Michael Giles, Gareth Roberts
- The London Mathematical Society: Elizabeth Fisher, Iain Stewart
- CRISM & The University of Warwick, Statistics
- Sponsors: Amazon, Google
- Partners: ISBA, MCQMC, BAYSM
- Poster: Talissa Gasser, Hidamari Design
- NSF: DMS-1407397 & DMS-1521145
- Planners: Murray Pollock, Christian Robert, Gareth Roberts
- Support: Paula Matthews, Murray Pollock, Shahin Tavakoli