# IEOR E4703: Monte-Carlo Simulation

## MCMC and Bayesian Modeling

**Martin Haugh**

Department of Industrial Engineering and Operations Research
Columbia University
Email: martin.b.haugh@gmail.com

Additional References: Ruppert and Matteson's *Statistics and Data Analysis for FE*, Christoper Bishop's *PRML*, David Barber's *BRML*, Gelman et al.'s *Bayesian Data Analysis*, Efron and Hastie's *Computer Age Statistical Inference*

Bayesian Modeling
- Conjugate Priors
- Computational Issues in Bayesian Modeling

The Sampling Problem
- One Approach: Acceptance-Rejection Algorithm

Another Approach: Markov Chain Monte-Carlo (MCMC)
- Markov Chains
- Metropolis-Hastings
- Examples

MCMC: Gibbs Sampling
- Examples
- Difficulties With Gibbs Sampling

MCMC Convergence Analysis and Output Analysis
- MCMC Output Analysis
- MCMC Convergence Diagnostics

Some Applications of Bayesian Modeling & MCMC

- Data Augmentation for Binary Response Regression
- Asset Allocation with Views
- A Novel Application of MCMC: Optimization and Code-Breaking
- Topic Modeling and LDA
- A Brief Detour on Graphical Models

Appendix

- Bayesian Model Checking
- Bayesian Model Selection
- Hamiltonian Monte-Carlo
- Empirical Bayes

# Bayes Theorem

Not surprisingly, Bayes's Theorem is the key result that drives Bayesian modeling and statistics.

Let $\mathcal{S}$ be a sample space and let $B_1, \ldots, B_K$ be a partition of $\mathcal{S}$ so that (i) $\bigcup_k B_k = \mathcal{S}$ and (ii) $B_i \bigcap B_j = \emptyset$ for all $i \neq j$.

**Bayes's Theorem:** Let $A$ be any event. Then for any $1 \leq k \leq K$ we have

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{P(A)} = \frac{P(A \mid B_k)P(B_k)}{\sum_{j=1}^{K} P(A \mid B_j)P(B_j)}.$$

Of course there is also a continuous version of Bayes's Theorem with sums replaced by integrals.

Bayes's Theorem provides us with a simple rule for updating probabilities when new information appears

- in Bayesian modeling and statistics this new information is the observed data
- and it allows us to update our prior beliefs about parameters of interest which are themselves assumed to be random variables.

# The Prior and Posterior Distributions

Let $\boldsymbol{\theta}$ be some unknown parameter vector of interest. We assume $\boldsymbol{\theta}$ is random with some distribution, $\pi(\boldsymbol{\theta})$

- this is our prior distribution which captures our prior uncertainty regarding $\boldsymbol{\theta}$.

There is also a random vector, $\mathbf{X}$, with PDF (or PMF) $p(\mathbf{x} \mid \boldsymbol{\theta})$

- this is the likelihood.

The joint distribution of $\boldsymbol{\theta}$ and $\mathbf{X}$ is then given by $p(\boldsymbol{\theta}, \mathbf{x}) = \pi(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta})$

- we can integrate to get the marginal distribution of $\mathbf{X}$

$$p(\mathbf{x}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

We can compute the posterior distribution via Bayes's Theorem:

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x})} \;=\; \frac{\pi(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta}) \, d\boldsymbol{\theta}} \tag{1}$$

# The Prior and Posterior Distributions

The mode of the posterior is called the maximum a posterior (MAP) estimator while the mean is of course $\mathrm{E}\left[\boldsymbol{\theta} \mid \mathbf{X} = \mathbf{x}\right] = \int \boldsymbol{\theta}\, \pi(\boldsymbol{\theta} \mid \mathbf{x})\, d\boldsymbol{\theta}$.

The posterior predictive distribution is the distribution of a new as yet unseen data-point, $\mathbf{X}_{new}$:

$$p(\mathbf{x}_{new}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta} \mid \mathbf{x}) p(\mathbf{x}_{new} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}$$

which is obtained using the posterior distribution of $\boldsymbol{\theta}$ given the observed data $\mathbf{X}$.

Much of Bayesian analysis is concerned with “understanding” the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{x})$. Note that

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

which is what we often work with:

1. Sometimes can recognize the form of the posterior by simply inspecting $\pi(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta})$
2. But often cannot recognize the posterior and can’t compute the denominator in (1) either
    - so approximate inference techniques such as MCMC must be used.

# E.G: A Beta Prior and Binomial Likelihood

Let $\theta \in (0,1)$ represent some unknown probability. We assume a Beta$(\alpha, \beta)$ prior so that

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < \theta < 1.$$

We also assume that $X \mid \theta \sim$ Bin$(n, \theta)$ so that

$$p(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, \ldots, n.$$

The posterior then satisfies

$$\begin{aligned} p(\theta \mid x) &\propto \pi(\theta) p(x \mid \theta) \\ &= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &\propto \theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1} \end{aligned}$$

which we recognize as the Beta$(\alpha + x, \beta + n - x)$ distribution!

**Question:** How can we interpret the prior distribution in this example?

# E.G: A Beta Prior and Binomial Likelihood

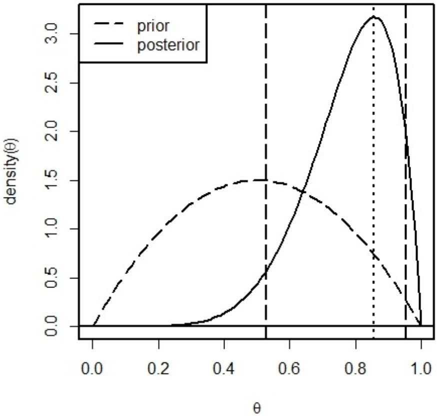


**Figure 20.1 from Ruppert's Statistics and Data Analysis for FE**: Prior and posterior densities for $\alpha = \beta = 2$ and $n = x = 5$. The dashed vertical lines are at the lower and upper $0.05$-quantiles of the posterior, so they mark off a $90\%$ equal-tailed posterior interval. The dotted vertical line shows the location of the posterior mode at $\theta = 6/7 = 0.857$.

# Conjugate Priors

Consider the following probabilistic model:

- parameter $\boldsymbol{\theta} \sim \pi(\,\cdot\,; \boldsymbol{\alpha}_0)$
- data $\mathbf{X} = (X_1, \ldots, X_N) \sim p(\mathbf{x} \mid \boldsymbol{\theta})$

As we saw before posterior distribution satisfies

$$p(\boldsymbol{\theta} \mid \mathbf{x}) \propto\; p(\boldsymbol{\theta}, \mathbf{x}) = p(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}; \boldsymbol{\alpha}_0)$$

Say the prior $\pi(\boldsymbol{\theta} \mid \alpha)$ is a conjugate prior for the likelihood $p(\mathbf{x} \mid \boldsymbol{\theta})$ if the posterior satisfies

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \pi(\boldsymbol{\theta}; \boldsymbol{\alpha}(\mathbf{x}))$$

so observations influence the posterior only via a parameter change $\boldsymbol{\alpha}_0 \rightarrow \boldsymbol{\alpha}(\mathbf{x})$

– the form or type of the distribution is unchanged.

**e.g.** In the previous example we saw the beta distribution is conjugate for the binomial likelihood.

# Conjugate Prior for Mean of a Normal Distribution

Suppose $\theta \sim \mathcal{N}(\mu_0, \gamma_0^2)$ and $p(X_i \mid \theta) = N(\theta, \sigma^2)$ for $i = 1, \ldots, N$

- so $\boldsymbol{\alpha}_0 = (\mu_0, \gamma_0^2)$ and $\sigma^2$ is assumed known.

If $\mathbf{X} = (X_1, \ldots, X_N)$ we then have

$$\begin{aligned} p(\theta \mid \mathbf{x}) &\propto p(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\theta; \boldsymbol{\alpha}_0) \\ &\propto e^{-\frac{(\theta-\mu_0)^2}{2\gamma_0^2}} \prod_{i=1}^{N} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \\ &\propto \exp\left(-\frac{(\theta-\mu_1)^2}{2\gamma_1^2}\right) \end{aligned}$$

where

$$\begin{aligned} \gamma_1^{-2} &:= \gamma_0^{-2} + N\sigma^{-2} \\ \text{and } \mu_1 &:= \gamma_1^2 \big(\mu_0 \gamma_0^{-2} + \sum_{i=1}^{n} x_i \sigma^{-2}\big). \end{aligned}$$

Of course we recognize $p(\theta \mid \mathbf{x})$ as the $\mathrm{N}(\mu_1, \gamma_1^2)$ distribution.

# Conjugate Prior for Mean and Variance of a Normal Dist.

Suppose that $p(X_i \mid \theta) = N(\mu, \sigma^2)$ for $i = 1, \ldots, N$ and let $\mathbf{X} := (X_1, \ldots, X_N)$.

Now assume $\mu$ and $\sigma^2$ are unknown so $\boldsymbol{\theta} = (\mu, \sigma^2)$.

We assume a joint prior of the form

$$\begin{aligned} \pi(\mu, \sigma^2) &= \pi(\mu \mid \sigma^2)\pi(\sigma^2) \\ &= \mathrm{N}\left(\mu_0, \sigma^2/\kappa_0\right) \times \text{Inv-}\chi^2\left(\nu_0, \sigma_0^2\right) \\ &\propto \sigma^{-1}\left(\sigma^2\right)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2\right]\right) \end{aligned}$$

– the N-Inv-$\chi^2(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$ density.

Note that $\mu$ and $\sigma^2$ are not independent under this joint prior.

**Exercise:** Show that multiplying this prior by the normal likelihood yields a N-Inv-$\chi^2$ distribution.

# The Exponential Family of Distributions

Canonical form of the exponential family distribution is

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = h(\mathbf{x}) e^{\boldsymbol{\theta}^\top \mathbf{u}(\mathbf{x}) - \psi(\boldsymbol{\theta})}$$

- $\boldsymbol{\theta} \in \mathbb{R}^m$ is a parameter vector
- and $\mathbf{u}(\mathbf{x}) = (u_1(\mathbf{x}), \ldots, u_m(\mathbf{x}))$ is the vector of sufficient statistics.

The exponential family includes Normal, Gamma, Beta, Poisson, Dirichlet, Wishart and Multinomial distributions as special cases.

The exponential family is essentially the only distribution with a non-trivial conjugate prior.

The conjugate prior takes the form

$$\pi(\boldsymbol{\theta}; \boldsymbol{\alpha}, \gamma) \propto e^{\boldsymbol{\theta}^\top \boldsymbol{\alpha} - \gamma \psi(\boldsymbol{\theta})}$$

since

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\alpha}, \gamma) \quad &\propto \quad e^{\boldsymbol{\theta}^\top \mathbf{u}(\mathbf{x}) - \psi(\boldsymbol{\theta})} e^{\boldsymbol{\theta}^\top \boldsymbol{\alpha} - \gamma \psi(\boldsymbol{\theta})} = e^{\boldsymbol{\theta}^\top (\alpha + \mathbf{u}(\mathbf{x})) - (\gamma + 1) \psi(\boldsymbol{\theta})} \\ &= \quad \pi(\boldsymbol{\theta} \mid \boldsymbol{\alpha} + \mathbf{u}(\mathbf{x}), \gamma + 1) \end{aligned}$$

# Selecting a Prior

Selecting an appropriate prior is a key component of Bayesian modeling.

With only a finite amount of data, the prior can have a very large influence on the posterior

- important to be aware of this and understand sensitivity of posterior inference to the choice of prior
- often try to use non-informative priors to limit this influence
- when possible conjugate priors often chosen for tractability reasons

A common misconception that the only advantage of the Bayesian approach over the frequentist approach is that the choice of prior allows us to express our prior beliefs on quantities of interest

- in fact there are many other more important advantages including modeling flexibility via MCMC, exact inference rather than asymptotic inference, ability to estimate functions of any parameters without "plugging" in MLE estimates, more accurate estimates of parameter uncertainty, etc.
- of course there are disadvantages as well including subjectivity induced by choice of prior and high computational costs.

# Inference in Bayesian Modeling

Despite differences in Bayesian and frequentist approaches we do have:

**Bernstein-von Mises Theorem:** Under suitable assumptions and for sufficiently large sample sizes,the posterior distribution of $\boldsymbol{\theta}$ is approximately normal with mean equal to the true value of $\boldsymbol{\theta}$ and variance equal to the inverse of the Fisher information matrix.

This theorem implies that Bayesian and MLE estimators have the same large sample properties

- not really surprising since influence of the prior should diminish with increasing sample sizes.

But this is a theoretical result and often don't have "large" sample sizes so quite possible for posterior to be (very) non-normal and even multi-modal.

Most of Bayesian inference is concerned with (which often means simulating from) the posterior

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) \;\propto\; \pi(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta}) \tag{2}$$

without knowing the constant of proportionality in (2).

This leads to the following sampling problem:

# The Basic Sampling Problem

Suppose we are given a distribution function

$$p(\mathbf{z}) = \frac{1}{Z_p}\tilde{p}(\mathbf{z})$$

where $\tilde{p}(\mathbf{z}) \geq 0$ is easy to compute but $Z_p$ is (too) hard to compute.

This very important situation arises in several contexts:

1. In Bayesian models where $\tilde{p}(\boldsymbol{\theta}) := p(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is easy to compute but $Z_p := p(\mathbf{x}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta})p(\mathbf{x} \mid \boldsymbol{\theta})d\boldsymbol{\theta}$ can be very difficult or impossible to compute.
2. In models from statistical physics, e.g. the Ising model, we only know $\tilde{p}(\mathbf{z}) = e^{-\mathcal{E}(\mathbf{z})}$, where $\mathcal{E}(\mathbf{z})$ is an "energy" function
   - the Ising model is an example of a Markov network or an undirected graphical model.
3. Dealing with evidence in directed graphical models such as belief networks aka directed acyclic graphs.

# How to Generate Samples from $p(\mathbf{z})$?

An important method is the acceptance-rejection algorithm:

- Choose a proposal density $q(\mathbf{z})$ from which it is easy to simulate.
- The support of $q(\cdot)$ must contain the support of $\tilde{p}(\mathbf{z})$
    - can therefore choose $k > 0$ so that $k \cdot q(\mathbf{z}) \geq \tilde{p}(\mathbf{z})$ for all $\mathbf{z}$.
- Generate $\mathbf{Z} \sim q(\cdot)$ and $U \sim U(0,1)$.
- Accept $\mathbf{Z}$ if $U \leq \frac{\tilde{p}(\mathbf{Z})}{k \cdot q(\mathbf{Z})}$. Otherwise re-sample $(\mathbf{Z}, U)$.

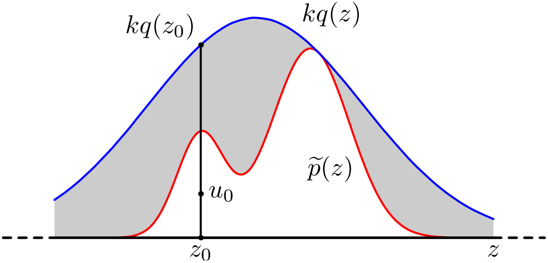


**Figure 11.4 from Bishop**

Alternative representation:

- Sample a point uniformly from the region under the curve $kq(\mathbf{z})$
- Accept the point if it lies in the white region.

# Efficiency of Acceptance-Rejection Algorithm

**Question:** How many iterations, $I$, does it take on average to generate a sample?

- The probability of success on each iteration is (why?) $Z_p/k$.
- Clear that $I$ has a geometric distribution.
- Therefore have $\mathbb{E}[I] = k/Z_p$.

Some implications . . .

- Would like to take $k$ as small as possible.
- When $q(\mathbf{z}) = p(\mathbf{z})$, we get $\mathbb{E}[I] = 1$.
- $\mathbb{E}[I]$ can be large if $q(\mathbf{z})$ is very different from $p(\mathbf{z})$
  - so would like have the proposal distribution close to the true one!

Acceptance-rejection can work very well in low-dimensions

- but can be extremely inefficient (why?) in high dimensions.

Nonetheless, can be a useful technique (even in high dimensions) when combined with MCMC methods.

# Another Approach: Markov Chain Monte-Carlo (MCMC)

- MCMC algorithms were originally developed in the 1940's by physicists at Los Alamos
  - Ulam (playing solitaire!), Von Neumann (acceptance-rejection!) and others.
- They were interested in modeling the probabilistic behavior of collections of atomic particles
  - could not be done analytically but maybe they could use simulation?
- Simulation was difficult – the normalization constant $Z_p$ was not known
  - and simulation hadn't (why?) been "discovered" yet
  - although simulation ideas had been around for some time

    e.g. Buffon's needle (1700's), Lord Kelvin (1901), Fermi (1930's).
  - in fact the term "Monte-Carlo" was coined at Los Alamos.
- Ulam and Metropolis overcame this problem by constructing a Markov chain for which the desired distribution was the stationary distribution
  - then only needed to simulate the Markov chain until stationarity achieved
  - they introduced the Metropolis algorithm and its impact was enormous.
- Introduced to statistics and generalized with the Metropolis-Hastings algorithm (1970) and the Gibbs sampler of Geman and Geman (1984).

# But First ... Some Markov Chain Theory

**Definition:** A sequence of random variables $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t\}$ on a discrete state space $\Omega$ is called a (first-order) Markov Chain if

$$p(\mathbf{X}_t = \mathbf{x}_t \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1}, \ldots, \mathbf{X}_1 = \mathbf{x}_1) = p(\mathbf{X}_t = \mathbf{x}_t \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1}).$$

We will restrict ourselves to time-homogeneous Markov chains:

$$p(\mathbf{X}_t = \mathbf{x}_t \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1}) = \mathbf{P}(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \in \mathbb{R}^{\Omega \times \Omega}$$

Easy to check that $[p(\mathbf{X}_{t+1} = \mathbf{x}_t \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1})]_{(\mathbf{x}_t, \mathbf{x}_{t-1}) \in \Omega} = \mathbf{P}^2$

**Definition:** A Markov chain is called ergodic if there exists $r$ such that $\mathbf{P}^r > 0$ – this is equivalent to the Markov chain being:

1. Irreducible: For all $\mathbf{x}, \mathbf{y} \in \Omega$, there exists $r(\mathbf{x}, \mathbf{y})$ s.t. $\mathbf{P}^{r(\mathbf{x},\mathbf{y})}(\mathbf{x}, \mathbf{y}) > 0$
2. Aperiodic: For all $\mathbf{x} \in \Omega$, GCD$\big\{r : P^r(\mathbf{x}, \mathbf{x}) > 0\big\} = 1$.

# Stationary Distributions of Markov Chains

**Definition:** A stationary distribution of a Markov chain is a distribution $\pi$ on $\Omega$ such that

$$\pi(\mathbf{y}) = \sum_{\mathbf{x}\in\Omega} P(\mathbf{y} \mid \mathbf{x})\pi(\mathbf{x}).$$

**Theorem**: A finite ergodic Markov Chain has a unique stationary distribution.

**Definition:** The total variation distance, $d_{TV}(\mu, \nu)$, between two probability measures $\mu, \nu$ on $\Omega$ is defined as

$$\|\mu - \nu\|_{TV} := \max_{S\subset\Omega}\{\mu(S) - \nu(S)\} = \frac{1}{2}\sum_{\mathbf{z}\in\Omega} |\mu(\mathbf{z}) - \nu(\mathbf{z})|$$

The mixing time of $\tau_{\mathrm{mix}}(\epsilon)$ defined as the time until the total variation distance to $\pi$ is below $\epsilon$

$$\tau_{\mathrm{mix}}(\epsilon) = \max_{\mathbf{x}_0\in\Omega} \min\left\{t : \left\|P^t(\cdot, \mathbf{x}_0) - \pi(\cdot)\right\|_{TV} \leq \epsilon\right\} \sim \ln\left(\frac{1}{\epsilon}\right)$$

Would like to have similar properties for continuous sample spaces!

# Reversible Markov Chains

**Definition:** A Markov chain is said to be reversible if there exists a probability measure $\pi$ on $\Omega$ such that

$$P(\mathbf{x} \mid \mathbf{y})\pi(\mathbf{y}) = P(\mathbf{y} \mid \mathbf{x})\pi(\mathbf{x}) \tag{3}$$

Easy to check that if $\pi$ satisfies (3) then it is the stationary distribution of the Markov chain since then

$$\sum_{\mathbf{x}} P(\mathbf{y} \mid \mathbf{x})\pi(\mathbf{x}) = \sum_{\mathbf{x}} P(\mathbf{x} \mid \mathbf{y})\pi(\mathbf{y}) = \pi(\mathbf{y})$$

- so (3) then implies the chain moves from **x** to **y** at the same rate it moves from **y** to **x** (when in equilibrium)
- for this reason (3) is often called the detailed balance equation

Satisfying the detailed balance equation is a sufficient (but not necessary) condition for $\pi$ to be a stationary distribution

- will also want to have ergodicity to guarantee that $\pi$ is the stationary distribution

There are analogous definitions and results for continuous Markov chains.

# The Metropolis-Hastings Algorithm

Suppose we want to sample from a distribution $p(\mathbf{x}) := \tilde{p}(\mathbf{x})/Z_p$.

We can construct a (reversible) Markov chain as follows. Let $\mathbf{X}_t = \mathbf{x}$ be the current state:

- Generate $\mathbf{Y} \sim Q(\cdot \mid \mathbf{x})$ for some Markov transition matrix $Q$.

  Let $\mathbf{y}$ be the generated value.
- Set $\mathbf{X}_{t+1} = \mathbf{y}$ with probability $\alpha(\mathbf{y} \mid \mathbf{x}) := \min\left\{\frac{\tilde{p}(\mathbf{y})}{\tilde{p}(\mathbf{x})} \cdot \frac{Q(\mathbf{x}|\mathbf{y})}{Q(\mathbf{y}|\mathbf{x})}, 1\right\}$.

  Otherwise set $\mathbf{X}_{t+1} = \mathbf{x}$.

**Claim**: The resulting Markov chain is reversible with stationary distribution $p(\mathbf{x}) = \tilde{p}(\mathbf{x})/Z_p$.

Note that $Z_p$ is not required for the algorithm!

Note also that if $\mathbf{Y} = \mathbf{y}$ is rejected then the current state $\mathbf{x}$ becomes the next state so that $\mathbf{X}_t = \mathbf{X}_{t+1} = \mathbf{x}$.

Can therefore sample from $p(\mathbf{x})$ by running the algorithm until stationarity is achieved and then using generated points as our samples.

# The Metropolis-Hastings Algorithm

**Proof of Claim**: We just check that $p(\mathbf{x})$ satisfies the detailed balance equations:

$$\begin{aligned}
\underbrace{\alpha(\mathbf{y} \mid \mathbf{x})Q(\mathbf{y} \mid \mathbf{x})}_{P(\mathbf{y}|\mathbf{x})}p(\mathbf{x}) &= \min\left\{\frac{p(\mathbf{y})}{p(\mathbf{x})} \cdot \frac{Q(\mathbf{x} \mid \mathbf{y})}{Q(\mathbf{y} \mid \mathbf{x})}, 1\right\} Q(\mathbf{y} \mid \mathbf{x})p(\mathbf{x}) \\
&= \min\left\{Q(\mathbf{x} \mid \mathbf{y})p(\mathbf{y}), Q(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})\right\} \\
&= \min\left\{1, \frac{p(\mathbf{x})}{p(\mathbf{y})} \cdot \frac{Q(\mathbf{y} \mid \mathbf{x})}{Q(\mathbf{x} \mid \mathbf{y})}\right\} Q(\mathbf{x} \mid \mathbf{y})p(\mathbf{y}) \\
&= \underbrace{\alpha(\mathbf{x} \mid \mathbf{y})Q(\mathbf{x} \mid \mathbf{y})}_{P(\mathbf{x}|\mathbf{y})}p(\mathbf{y}).
\end{aligned}$$

**Question:** How do we determine when stationarity is achieved?

- will use convergence diagnostics (to be discussed later) to do this.

**Question:** There are many possible choices of $Q(\cdot \mid \cdot)$. What should we use?

- an important question since $Q(\cdot \mid \cdot)$ influences how much time required to reach stationarity
- won't have time to say much on this question.

**Question:** Are the samples independent?

# Example: Sampling from a 2-D Gaussian

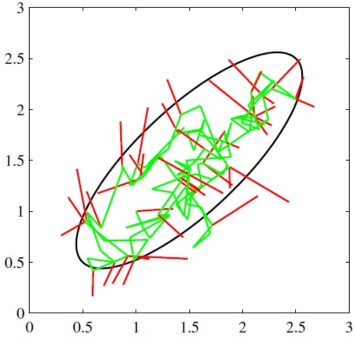


**Figure 11.9 from Bishop**: A simple illustration using Metropolis algorithm to sample from a Gaussian distribution whose one standard-deviation contour is shown by the ellipse. The proposal distribution is an isotropic Gaussian distribution whose standard deviation is $0.2$. Steps that are accepted are shown as green lines, and rejected steps are shown in red. A total of $150$ candidate samples are generated, of which $43$ are rejected.

# Example: Sampling from a Multi-Modal Distribution

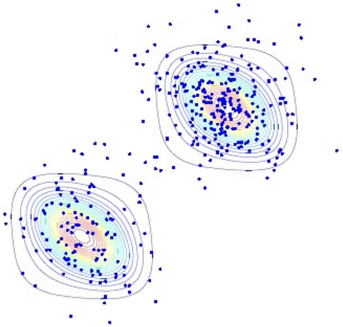


**Figure 27.8 from Barber**: Metropolis-Hastings samples from a bi-variate distribution $p(x_1, x_2)$ using a proposal $\tilde{q}(\mathbf{x}'|\mathbf{x}) = N(\mathbf{x}'|\mathbf{x}, \mathbf{I})$. We also plot the iso-probability contours of $p$. Although $p(\mathbf{x})$ is multi-modal, the dimensionality is low enough and the modes sufficiently close such that a simple Gaussian proposal distribution is able to bridge the two modes. In higher dimensions, such multi-modality is more problematic.

**Question:** Why do you think it might sometimes be difficult to sample from a multi-modal distribution?

# Gibbs Sampling

Gibbs sampling is an MCMC sampler introduced by Geman and Geman in 1984

- named after the physicist J. W. Gibbs who died 80 years earlier.

Let $\mathbf{x}^{(t)} \in \mathbb{R}^m$ denote the current sample. Then Gibbs sampling proceeds as follows:

1. Pick an index $k \in \{1, \ldots, m\}$ either via round-robin or uniformly at random
2. Set $\mathbf{x}_j^{(t+1)} = \mathbf{x}_j^{(t)}$, for $j \neq k$, i.e. $\mathbf{x}_{-k}^{(t+1)} = \mathbf{x}_{-k}^{(t)}$
3. Generate $\mathbf{x}_k^{(t+1)} \sim p(\mathbf{x}_k \mid \mathbf{x}_{-k}^{(t)})$

   – so only one component of $\mathbf{x}$ is updated at a time.

Common to simply order the $m$ components and update them sequentially. Can then let $\mathbf{x}_k^{(t+1)}$ be the value of the chain after all $m$ updates rather than each individual update.

A very popular method when the (true) conditional distributions, $p(\mathbf{x}_j \mid \mathbf{x}_{-k}^{(t)})$, are easy to simulate from

- which is the case for conditionally conjugate models and others.

# Gibbs Sampling

Easy to see that Gibbs sampling is a special case of Metropolis-Hastings sampling with

$$Q_k(\mathbf{y} \mid \mathbf{x}) = \begin{cases} p(\mathbf{y}_k \mid \mathbf{x}_{-k}) & \mathbf{y}_{-k} = \mathbf{x}_{-k} \\ 0 & \text{otherwise.} \end{cases}$$

and that each component update will be accepted with probability 1.

Have to be careful that the component-wise Markov Chain is ergodic

- see Barber's Figure 27.5 later in these slides.

Instead of updating just 1 component at a time can also split **x** into blocks and update 1 block at a time.

If not possible to simulate directly from one or more of the conditional distributions can use rejection-sampling or Metropolis-Hastings sampling for those updates

- sometimes called Metropolis-with-Gibbs.

# A Simple Example

Consider the distribution

$$p(x,y) = \frac{n!}{(n-x)!x!} y^{(x+\alpha-1)} (1-y)^{(n-x+\beta-1)}, \quad x \in \{0,\ldots,n\}, y \in [0,1].$$

Hard to simulate directly from $p(x,y)$ but the conditional distributions are easy to work with. We see that

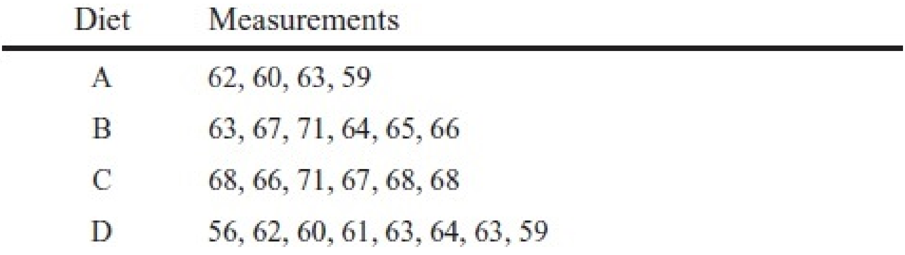
- $p(x \mid y) \equiv \text{Bin}(n,y)$
- $p(y \mid x) \equiv \text{Beta}(x+\alpha, n-x+\beta)$

Since it's easy to simulate from each conditional, it is easy to run a Gibbs sampler to simulate from the joint distribution.

**Question:** Given one of our earlier examples, can you identify a situation where this distribution might arise?

The marginal distribution of $x$ is the beta-binomial distribution.

# Hierarchical Models



| Diet | Measurements |
|---|---|
| A | 62, 60, 63, 59 |
| B | 63, 67, 71, 64, 65, 66 |
| C | 68, 66, 71, 67, 68, 68 |
| D | 56, 62, 60, 61, 63, 64, 63, 59 |

Table 11-2 taken from *Bayesian Data Analysis*, $2^{nd}$ edition by Gelman et al.

Gibbs sampling is particulary suited for hierarchical modeling

- we will consider an example from *Bayesian Data Analysis* by Gelman et al.
- the data is in Table 11-2 above.

# The Hierarchical Normal Model

Data $y_{ij}$, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$ are assumed to be independently normally distributed within each of $J$ groups with means $\theta_j$ and common variance $\sigma^2$. That is, $y_{ij} \mid \theta_j \sim N(\theta_j, \sigma^2)$.

Total number of observations is $n = \sum_{j=1}^{J} n_j$.

Group means are assumed to follow a normal distribution with unknown mean $\mu$ and variance $\tau^2$. That is $\theta_j \sim N(\mu, \tau^2)$.

A uniform prior is assumed for $(\mu, \log \sigma, \tau)$ so $p(\mu, \log \sigma, \log \tau) \propto \tau$

- if a uniform prior was assigned to $\log \tau$ then posterior would be improper as discussed in Gelman et al
    - this emphasizes the importance of understanding the **issues** associated with choosing priors.

The posterior then given by

$$p(\boldsymbol{\theta}, \mu, \log \sigma, \log \tau \mid \mathbf{y}) \propto \tau \prod_{j=1}^{J} \mathsf{N}\left(\theta_j \mid \mu, \tau^2\right) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathsf{N}\left(y_{ij} \mid \theta_j, \sigma^2\right).$$

# The Gibbs Sampler for the Hierarchical Normal Model

Will see that all conditional distributions required for Gibbs sampler have simple conjugate forms:

1. *Conditional Posterior Distribution of Each $\theta_j$*

   Just gather terms from posterior that only involve $\theta_j$ and then simplify to obtain

   $$\theta_j \mid (\boldsymbol{\theta}_{-j}, \mu, \sigma, \tau, \mathbf{y}) \sim \mathsf{N}\left(\hat{\theta}_j, V_{\theta_j}\right)$$

   where

   $$\begin{aligned} \hat{\theta}_j &:= \frac{\frac{1}{\tau^2}\mu + \frac{n_j}{\sigma^2}\bar{y}_{.j}}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}} \\ V_{\theta_j} &:= \frac{1}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}. \end{aligned}$$

   These conditional distributions are independent so generating the $\theta_j$'s one at a time is equivalent to drawing $\boldsymbol{\theta}$ all at once.

# The Gibbs Sampler for the Hierarchical Normal Model

2. *Conditional Posterior Distribution of $\mu$*

   Again, just gather terms from posterior that only involve $\mu$ and then simplify to obtain

$$\mu \mid (\boldsymbol{\theta}, \sigma, \tau, \mathbf{y}) \sim \mathsf{N}\left(\hat{\mu}, \frac{\tau^2}{J}\right)$$

   where

$$\hat{\mu} \quad := \quad \frac{1}{J}\sum_{j=1}^{J} \theta_j .$$

3. *Conditional Posterior Distribution of $\sigma^2$*

   Again, just gather terms from posterior that only involve $\sigma$ and then simplify to obtain

$$\sigma^2 \mid (\boldsymbol{\theta}, \mu, \tau, \mathbf{y}) \sim \mathsf{Inv}\text{-}\chi^2\left(n, \hat{\sigma}^2\right)$$

   where

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} \left(y_{ij} - \theta_j\right)^2 .$$

# The Gibbs Sampler for the Hierarchical Normal Model

4. *Conditional Posterior Distribution of $\tau^2$*

   Again, gather terms from posterior that only involve $\tau$ and then simplify to obtain

   $$\tau^2 \mid (\boldsymbol{\theta}, \mu, \sigma, \mathbf{y}) \sim \text{Inv-}\chi^2\left(J-1, \hat{\tau}^2\right)$$

   where

   $$\hat{\tau}^2 := \frac{1}{J-1} \sum_{j=1}^{J} \left(\theta_j - \mu\right)^2 .$$

To start the Gibbs sampler we need starting points for $\boldsymbol{\theta}$ and $\mu$

– but not (why?) for $\tau$ or $\sigma$.

# Difficulties With Gibbs Sampling

Gibbs sampling is a very popular MCMC technique that is widely used.

It does have some potential drawbacks, however:

1. Need to be able to show that the Gibbs sampler Markov chain is ergodic
   - obvious in many circumstances but sometimes an issue
   - for example Figure 27.5 from Barber shows a 2-dimensional example where the chain is not irreducible.
2. If the variables are strongly correlated (negatively or positively) then it may take too long to reach the stationary distribution
   - see Figure 27.7 from Barber and Figure 11.11 from Bishop.

**Question:** Suppose the random variables $x_1, \ldots, x_d$ are independent. How long do you think it will take the Gibbs sampler to reach stationarity in that case?

# An Example Where Gibbs Fails

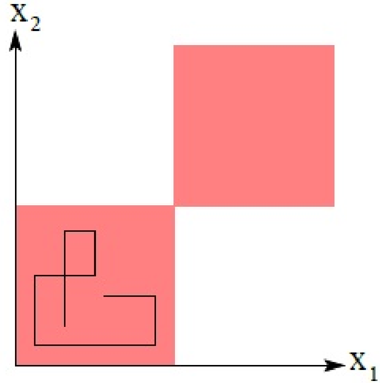
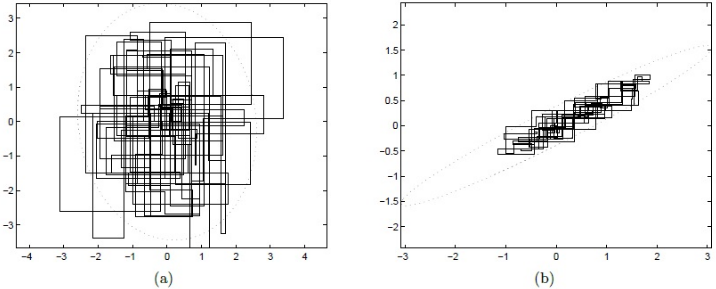


**Figure 27.5 from Barber**: A two dimensional distribution for which Gibbs sampling fails. The distribution has mass only in the shaded quadrants. Gibbs sampling proceeds from the $l^{th}$ sample state $(x_1^l, x_2^l)$ and then sampling from $p(x_2|x_1^l)$, which we write $(x_1^{l+1}, x_2^{l+1})$ where $x_1^{l+1} = x_1^l$. One then continues with a sample from $p(x_1|x_2 = x_2^{l+1})$, etc. If we start in the lower left quadrant and proceed this way, the upper right region is never explored.

# Gibbs is More Effective When Variables Are Less Correlated



(a) (b)

**Figure 27.7 from Barber**: Two hundred Gibbs samples for a two dimensional Gaussian. At each stage only a single component is updated. (a): For a Gaussian with low correlation, Gibbs sampling can move through the likely regions effectively. (b): For a strongly correlated Gaussian, Gibbs sampling is less effective and does not rapidly explore the likely regions.

When the variables are very correlated a common strategy is to seek variable transformations so that the transformed variables are approximately independent.

# Gibbs is More Effective When Variables Are Less Correlated

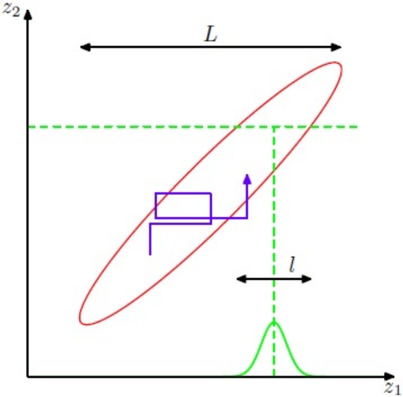


**Figure 11.11 from Bishop**: Illustration of Gibbs sampling by alternate updates of two variables whose distribution is a correlated Gaussian. The step size is governed by the standard deviation of the conditional distribution (green curve), and is $O(l)$, leading to slow progress in the direction of elongation of the joint distribution (red ellipse). The number of steps needed to obtain an independent sample from the distribution is $O((L/l)2)$.

# A Cautionary Example (From Casella and George, 1992)

Gibbs sampling implies that conditional distributions are sufficient to define the joint distribution.

But there is a subtle issue here: it is not the case that a set of proper well-defined conditional distributions will determine a proper marginal.

e.g. Consider the following $2$-dimensional example with

$$\begin{aligned} f(x \,|\, y) &= ye^{-yx}, \qquad 0 < x < \infty & (4) \\ f(y \,|\, x) &= xe^{-xy}, \qquad 0 < y < \infty & (5) \end{aligned}$$

so both conditionals are exponential distributions (and therefore well-defined).

If we apply a Gibbs sampler to (4) and (5), however, will not obtain a sample from any marginal or joint distribution!

This is because (4) and (5) do not correspond to any joint distribution on $(x, y)$.

# MCMC Output Analysis

We are usually interested in scalar-valued functions of the parameter vector $\boldsymbol{\theta}$.

Let $\psi(\boldsymbol{\theta})$ be one such function.

If we have $n$ MCMC samples from the stationary distribution then we have $n$ samples of $\psi(\boldsymbol{\theta})$:

$$\{\psi_1 := \psi(\boldsymbol{\theta}_1),\ \ldots\ ,\ \psi_n := \psi(\boldsymbol{\theta}_n)\}\,.$$

The sample mean is then given by $\bar{\psi} = n^{-1} \sum_{i=1}^{n} \psi_1$.

Posterior intervals for $\psi(\boldsymbol{\theta})$ can also be calculated:

1. Let $L(\alpha_1) := \alpha_1$ lower sample quantile and $U(\alpha_2) := \alpha_2$ upper sample quantile of $\psi_1, \ldots, \psi_n$. Then $(L(\alpha_1),\ U(\alpha_2))$ is a $1 - (\alpha_1 + \alpha_2)$ posterior interval.
2. If $\alpha_1 = \alpha_2 = \alpha/2$ then we obtain an equi-tailed $1 - \alpha$ posterior interval.
3. For a highest posterior density interval we solve (numerically) for $\alpha_1$ and $\alpha_2$ such that $\alpha = \alpha_1 + \alpha_2$ and $U(\alpha_2) - L(\alpha_1)$ is minimized
    - could be a union of intervals if posterior of $\psi(\boldsymbol{\theta})$ is not unimodal
    - kernel density estimates of the posterior density can be plotted to help determine number of modes.

# Convergence Diagnostics

In order to use the MCMC samples for inference we must:

1. Ensure the Markov chains have reached stationarity
2. Only use those samples that have been generated after stationarity has been reached.

But it's impossible to ensure when these two conditions are satisfied since the Markov chain does not begin with the stationary distribution. Instead we can use various methods to assess whether or not stationarity appears to have been reached :

1. Visual inspection where we plot variables (of interest) vs iteration #, plot running means of variables (of interest) etc.
   - can be very informative but they also require "manual" work.
2. Statistical summaries of MCMC output which are designed to diagnose convergence / non-convergence
   - they can be programmed and so "manual" labor not required

   We will consider the popular Gelman-Rubin methodology
   - will not justify everything here but details can be found in *Bayesian Data Analysis* by Gelman et.al. and also Chapter 20 of *SDFE* by Ruppert and Matteson. (The latter is available online.)

# Convergence Diagnostics Via Visual Inspection

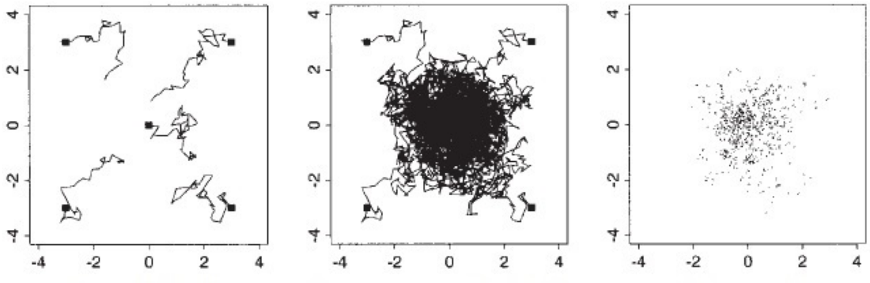


**Figure 11.2 from Gelman et al. (2nd Edition)**: Five independent sequences of a Markov chain simulation for the bivariate unit normal distribution, with over-dispersed starting points indicated by solid squares. (a) After 50 iterations, the sequences are still far from convergence. (b) After 1000 iterations, the sequences are nearer to convergence. Figure (c) shows the iterates from the second halves of the sequences. The points in Figure (c) have been jittered so that steps in which the random walk stood still are not hidden.

# The Gelman & Rubin Approach

Gelman & Rubin approach runs $m/2$ chains for a total of $n_0 + 2n$ iterations each.

The chains are begun from over-dispersed starting points

- usually obtained by generating them from some over-dispersed distribution.

We discard the first $n_0$ samples from each chain

- these samples constitute the burn-in period where the chains are assumed to be in their transient phase
- common to take $n_0 = 2n$ so first half of each chain is discarded.

Remaining component of each chain is then split into two (sub-)chains, each containing $n$ samples

- chain splitting will allow process (described below) to determine if each chain has reached stationarity.

At this point we therefore have $m$ chains each containing $n$ samples

- we hope these $m \times n$ samples are from the stationarity distribution
- so we check that this appears to be the case by comparing the between-chain variance with the within-chain variance for all scalar quantities, $\psi$, of interest.

# The Gelman & Rubin Approach

Because the method is based on means and variances generally a good idea to transform the scalar estimands so they are approximately normal

- e.g. take logs of strictly positive quantities
- e.g. take logits of quantities that must lie in $(0,1)$.

Let $\psi_{ij}$ for $i = 1, \ldots n$ and $j = 1, \ldots, m$ be the MCMC samples

- computed after the burn-in period
- and then splitting the non-burn-in component of each chain in two.

The between- and within-sequence variances, $B$ and $W$, are computed as

$$B := \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\psi}_{.j} - \bar{\psi}_{..}\right)^2$$

$$W := \frac{1}{m} \sum_{j=1}^{m} s_j^2 \quad \text{where} \quad s_j^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(\psi_{ij} - \bar{\psi}_{.j}\right)^2$$

and where $\bar{\psi}_{.j} := \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}$ and $\bar{\psi}_{..} := \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j}$.

# The Gelman & Rubin Approach

$B$ contains a factor of $n$ because it is based on the variance of the within-sequence means, $\bar{\psi}_{.j}$, each of which is an average of $n$ values.

We can estimate $\mathsf{Var}(\psi \mid \mathbf{X})$ as a weighted average of $W$ and $B$ with

$$\widehat{\mathsf{Var}}^{+}(\psi \mid \mathbf{X}) = \frac{n-1}{n} W \; + \; \frac{1}{n} B$$

- overestimates the marginal posterior variance since starting distribution is over-dispersed
- but unbiased when sampling from the desired stationary distribution.

But also have for any finite $n$ that $W$ should be an underestimate of $\mathsf{Var}(\psi \mid \mathbf{X})$

- since each individual sequence may not have had time to explore all of the target, i.e. stationary, distribution
- but $W$ should approach $\mathsf{Var}(\psi \mid \mathbf{X})$ in limit as $n \to \infty$.

# The Gelman & Rubin Approach

We therefore monitor convergence through

$$\hat{R} := \sqrt{\frac{\widehat{\mathsf{Var}}^{+}(\psi \mid \mathbf{X})}{W}}$$

Note that should have $\hat{R} > 1$ for any finite $n$ by above argument.
But also have $\hat{R} \to 1$ as $n \to \infty$.

Rule of Thumb: Values of $\hat{R} < 1.1$ are acceptable but closer $\hat{R}$ is to $1$ the better.

We then monitor $\hat{R}$ for all quantities $\psi$ of interest.

# The Gelman & Rubin Approach

Note that $B/n$ is the sample variance of $m$ chain means so $B/mn$ therefore estimates Monte-Carlo variance of $\bar{\psi}_{..}$.

Suppose now that we could take an independent sample of size $n_{eff}$.
Variance of the mean of this sample would be estimated as $\widehat{\mathsf{Var}}^{+}(\psi \mid \mathbf{X})/n_{eff}$.

Equating the two estimates yields the effective sample size, $n_{eff}$, as

$$n_{eff} := mn \frac{\widehat{\mathsf{Var}}^{+}(\psi \mid \mathbf{X})}{B} \tag{6}$$

Generally $n_{eff} < mn$ since samples within each sequence will be auto-correlated

- $n_{eff}/mn$ is then a measure of the simulation efficiency.

If $m$ is small then $B$ will have high sampling variability in which case $n_{eff}$ will be a crude estimate

- might prefer to report $\min(n_{eff}, mn)$ in this case.

# Applications of Bayesian Modeling & MCMC

Inference in (complex) Bayesian models is typically done via one of:

1. Sampling from the posterior using MCMC algorithms such as Metropolis-Hastings, Gibbs sampling or auxiliary variable methods such as slice sampling or Hamiltonian Monte-Carlo (HMC)
2. Approximating the posterior with more tractable distributions – a process known as deterministic inference
    - methods include variational Bayes and expectation propagation.

Over the past couple of decades a lot of software such `WinBugs`, `OpenBugs` and `JAGS` have been made freely available

- they use Gibbs sampling to simulate from posterior and also perform various convergence diagnostics

More recently `STAN` has been developed (mainly by researchers at Columbia U)

- relies on HMC to overcome slow mixing / convergence of Gibbs for very complex models.

There are Bayesian versions of classification, regression etc. as well as many other applications including ...

# Data Augmentation for Binary Response Regression

Have binary response variables $\mathbf{y} := (y_1, \ldots, y_m)$ and corresponding covariate vectors $\mathbf{x}_i := (x_{i1}, \ldots, x_{ik})$.

The probit regression model is a GLM where

$$p_i := P(y_i = 1) = \Phi\left(x_{i1}\beta_1 + \cdots + x_{ik}\beta_k\right).$$

Goal is to estimate $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_k)$

- can be done using standard GLM software using the 'probit' link function.

But we will use a Bayesian approach!

If we assume a prior $\pi(\boldsymbol{\beta})$ on $\boldsymbol{\beta}$ then posterior given by

$$\begin{aligned} g(\boldsymbol{\beta} \mid \mathbf{y}) &\propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \\ &= \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)^{y_i} \left(1 - \Phi\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right)^{1-y_i}. \end{aligned} \tag{7}$$

Not clear how to generate samples of $\boldsymbol{\beta}$ from the posterior in (7) using Gibbs.

# Data Augmentation for Binary Response Regression

A clever way to resolve this problem is to define latent variables

$$z_i := x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \epsilon_i$$

where the $\epsilon_i$'s are IID $\mathrm{N}(0,1)$ for $i = 1, \ldots, n$.

Note that (why?)

$$p_i = P(z_i > 0) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

Can now regard the problem as a missing data problem where instead of observing the $z_i$'s we only observe the indicators $y_i := 1_{\{z_i > 0\}}$.

Posterior distribution is now over $(\boldsymbol{\beta}, \mathbf{z})$ and is given by

$$\begin{aligned} g(\boldsymbol{\beta}, \mathbf{z} \,|\, \mathbf{y}) &\propto g(\boldsymbol{\beta}, \mathbf{z}, y) \\ &= \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \left[ 1_{\{z_i > 0\}} 1_{\{y_i = 1\}} + 1_{\{z_i \leq 0\}} 1_{\{y_i = 0\}} \right] \phi(z_i \,;\, \mathbf{x}_i^\top \boldsymbol{\beta}, 1) \quad (8) \end{aligned}$$

where $\phi(\cdot\,;\, \mu, \sigma^2)$ denotes the PDF for a normal random variable with mean $\mu$ and variance $\sigma^2$.

# Data Augmentation for Binary Response Regression

Posterior in (8) is in a particularly convenient form for Gibbs sampling.

Suppose we assume $\pi(\boldsymbol{\beta}) \equiv 1$, i.e. a uniform prior on $\boldsymbol{\beta}$.

Can then use a block Gibbs sampler where we simulate successively from $g(\boldsymbol{\beta} \,|\, \mathbf{z}, \mathbf{y})$ and $g(\mathbf{z} \,|\, \boldsymbol{\beta}, \mathbf{y})$.

Relatively(!) easy then to see that

$$g(\boldsymbol{\beta} \,|\, \mathbf{z}, \mathbf{y}) \sim \mathrm{MVN}_k \left( (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{z},\, (\mathbf{X}^\top \mathbf{X})^{-1} \right). \tag{9}$$

**Question:** How would you justify (9)?

**Question:** How can we simulate from $g(\mathbf{z} \,|\, \boldsymbol{\beta}, \mathbf{y})$?

# An Application to the Donner Party Wagon Trail Dataset

Consider the data-set on the Donner party, a group of wagon trail emigrants who struggled to cross the Sierra Nevada mountains in California in 1846-47.

Interested in estimating the model

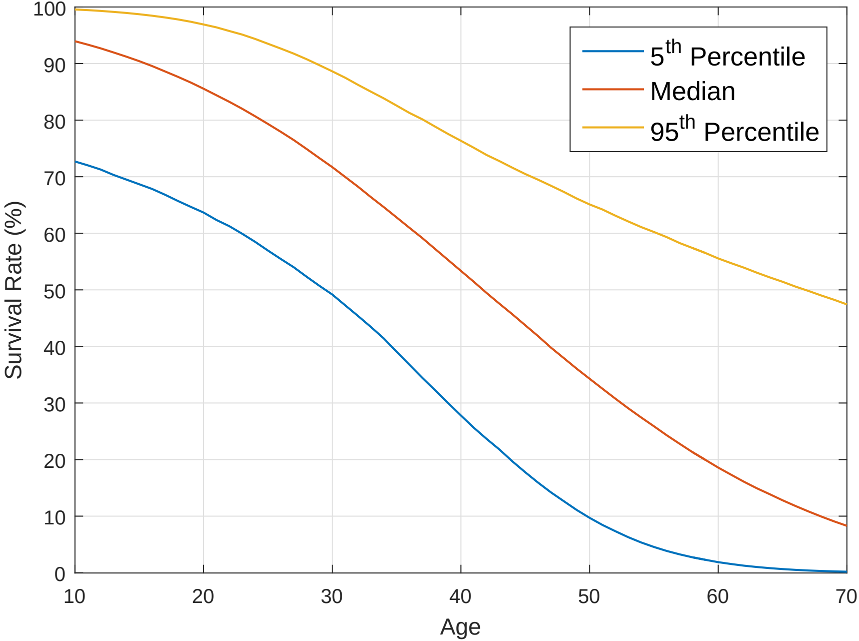
$$P(y_i = 1) = \Phi\left(\beta_0 + \beta_1 \mathsf{Male}_i + \beta_2 \mathsf{Age}_i\right) \tag{10}$$

where $y_i = 1$ denotes the death of the $i^{th}$ person in the party and $y_i = 0$ denotes their survival.

Have two covariates: Male (1 for males, 0 for females) and Age (in years).

Figure on next slide displays estimated percentile survival rates for men of various ages based in the Donner party

- computed by running the block Gibbs sampler and using the $\boldsymbol{\beta}$ samples (after convergence had been diagnosed) together with (10).


100
90
80
70
60
50
40
30
20
10
0
Survival Rate (%)
10
20
30
40
50
60
70
Age
5th Percentile
Median
95th Percentile

# Example: Asset Allocation with Views

In finance one can use sophisticated statistical / time series techniques to construct an objective model of security returns or risk factors.

Let $\mathbf{X}_{t+1}$ denote the change in risk factors between dates $t$ and $t+1$

- then all security returns from $t$ to $t+1$ depend on $\mathbf{X}_{t+1}$ only plus idiosyncratic noise.

Let $f(\cdot)$ denote the (objective) distribution of $\mathbf{X}_{t+1}$ based on all information available in the market place at date $t$.

The investor would like to construct an optimal portfolio based on the distribution $f(\cdot)$ as well as her own subjective views of what will happen in the market between dates $t$ and $t+1$.

**Question**: How can she do this?

# Example: Asset Allocation with Views

**Solution**: Let $\mathbf{V} = g(\mathbf{X}_{t+1}) + \boldsymbol{\epsilon}$ be a random vector where

- $g(\cdot)$ is a function representing how these views depend on $\mathbf{X}_{t+1}$
- and $\boldsymbol{\epsilon}$ is a noise vector reflecting how certain the investor is in her views.
  - $\boldsymbol{\epsilon}$ is assumed to be independent of $\mathbf{X}_{t+1}$ with distribution $\text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ say.

Suppose the investor believes that $g(\mathbf{X}_{t+1})$ will equal $\mathbf{v}$.

Then we construct the conditional distribution of $\mathbf{X}_{t+1}$ given $\mathbf{V} = \mathbf{v}$ and obtain

$$\begin{aligned} f(\mathbf{X}_{t+1} \,|\, \mathbf{V} = \mathbf{v}) &\propto f(\mathbf{X}_{t+1}, \mathbf{v}) \\ &= f(\mathbf{v} \,|\, \mathbf{X}_{t+1})\, f(\mathbf{X}_{t+1}). \end{aligned} \tag{11}$$

We can use MCMC to simulate many samples from (11) which can then be used to construct portfolios.

We obtain the famous Black-Litterman model when $\mathbf{X}_{t+1}$ is the vector of security returns, $g(\cdot)$ is linear, and all distributions are multivariate normal

- in this case the posterior can be calculated analytically.

# A Novel Application: Optimization and Code-Breaking

One day a psychologist from California's state prison system showed up at the consulting service of Stanford's Statistics department.

The problem was to decode a collection of coded messages – see example below.

Student in consulting service guessed it was a simple substitution cipher

- so each symbol represents a letter, number, punctuation mark or a space.

Goal then is to crack this cipher and find the function

$$f\ :\ \{\text{code space}\} \rightarrow \{\text{usual alphabet}\}. \qquad (12)$$

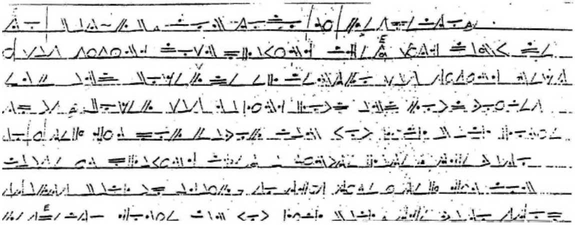


Figure taken from "The Markov Chain Monte Carlo Revolution", by Persi Diaconis in the *Bulletin of the American Mathematical Society* (2008).

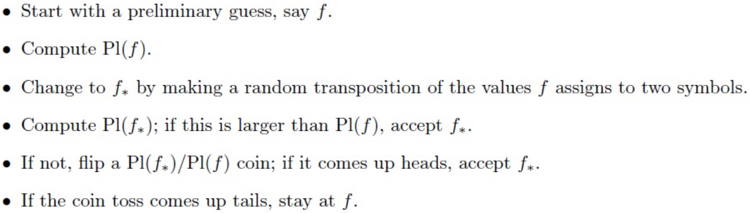
# A Novel Application: Optimization and Code-Breaking

Solution approach:

1. Find a text, e.g. *War and Peace*, and record the first-order transitions, i.e. the proportion of consecutive text symbols from $x$ to $y$
    - yields a matrix $M(x, y)$ of transitions
2. Can then define a plausibility to any function $f(\cdot)$ vis

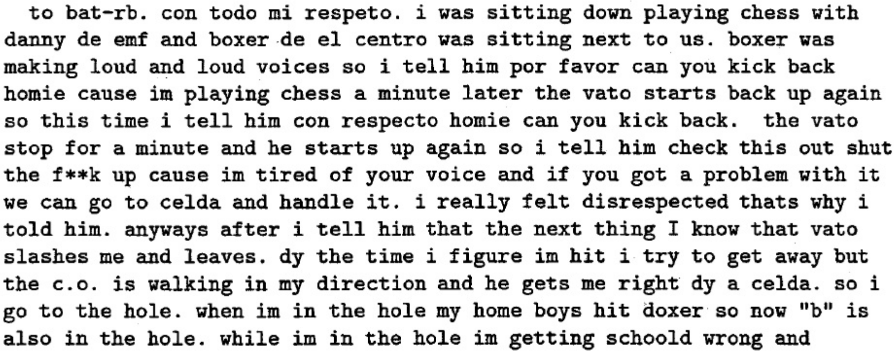
$$\text{Pl}(f) := \prod_i M\left(f(s_i), f(s_{i+1})\right)$$

   where $s_i$ runs over symbols in coded message.
3. Functions with high values of Pl$(f)$ are good candidates for decryption code in (12).
4. So search for maximal $f(\cdot)'s$ by running the following MCMC:



- Start with a preliminary guess, say $f$.
- Compute Pl$(f)$.
- Change to $f_*$ by making a random transposition of the values $f$ assigns to two symbols.
- Compute Pl$(f_*)$; if this is larger than Pl$(f)$, accept $f_*$.
- If not, flip a Pl$(f_*)$/Pl$(f)$ coin; if it comes up heads, accept $f_*$.
- If the coin toss comes up tails, stay at $f$.

# And the Solution ....



to bat-rb. con todo mi respeto. i was sitting down playing chess with danny de emf and boxer de el centro was sitting next to us. boxer was making loud and loud voices so i tell him por favor can you kick back homie cause im playing chess a minute later the vato starts back up again so this time i tell him con respecto homie can you kick back. the vato stop for a minute and he starts up again so i tell him check this out shut the f**k up cause im tired of your voice and if you got a problem with it we can go to celda and handle it. i really felt disrespected thats why i told him. anyways after i tell him that the next thing I know that vato slashes me and leaves. dy the time i figure im hit i try to get away but the c.o. is walking in my direction and he gets me right dy a celda. so i go to the hole. when im in the hole my home boys hit doxer so now "b" is also in the hole. while im in the hole im getting schoold wrong and

Figure taken from “The Markov Chain Monte Carlo Revolution”, by Persi Diaconis in the *Bulletin of the American Mathematical Society* (2008).

# Example: Topic Modeling

LDA is a hierarchical model used to model text documents:

- Each document is modeled as a mixture of topics.
- Each topic is then defined as a distribution over the words in the vocabulary.

We assume there are:

- A total of $K$ topics.
- A total of $D$ documents.
- A total of $M$ words in the vocabulary / dictionary
  - words are numbered from $1$ to $M$.

The latent Dirichlet allocation (LDA) topic model is obtained in the following generative fashion:

# Example: Topic Modeling

1. A topic mixture $\boldsymbol{\theta}_d$ for each document is drawn independently from a $\text{Dir}_K(\alpha\mathbf{1})$ distribution, where $\text{Dir}_K(\boldsymbol{\phi})$ is a Dirichlet distribution over the $K$-dimensional simplex with parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$.
2. Each of the $K$ topics $\{\boldsymbol{\beta}_k\}_{k=1}^K$ are drawn independently from a $\text{Dir}_M(\gamma\mathbf{1})$ distribution.
3. Then for each of the $i = 1 \ldots, N_d$ words in document $d$, an assignment variable $z_i^d$ is drawn from $\text{Mult}(\boldsymbol{\theta}_d)$.
4. Conditional on the assignment variable $z_i^d$, word $i$ in document $d$, denoted as $w_i^d$, is drawn independently from $\text{Mult}(\boldsymbol{\beta}_{z_i^d})$

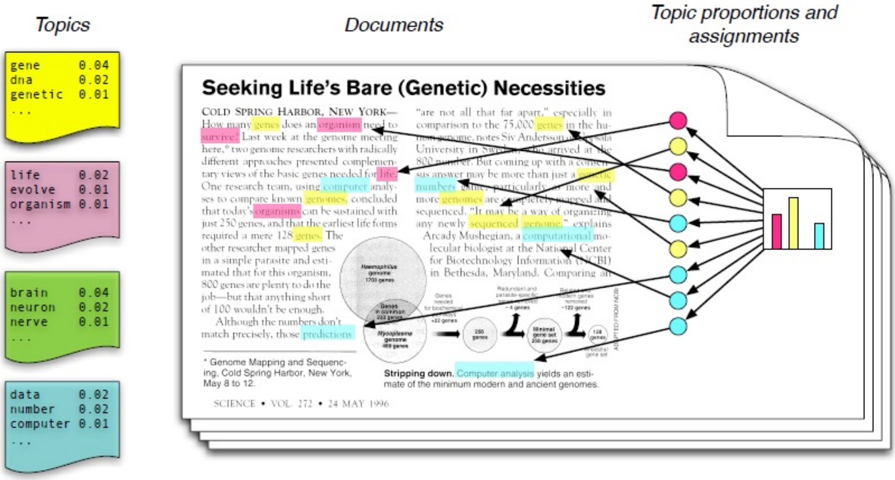
This is a hierarchical model and it is easy to write out the joint distribution of all the data.

Only the $w_i^d$'s are observed, however, so we need to use the conditional distribution to learn the topic mixtures for each document, the $K$ topic distributions and the latent variables $z_i^d$

- typically done via Gibbs sampling or variational Bayes.

**Question**: Is this a bag-of-words model?

# Example: Topic Modeling


Topics
Documents
Topic proportions and assignments
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 468 genes
Genes needed for biochemical pathways +22 genes
Redundant and parasite-specific genes removed −4 genes
Related and modern genes removed −122 genes
256 genes
Minimal gene set 250 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

| | |
|---|---|
| gene | 0.04 |
| dna | 0.02 |
| genetic | 0.01 |
| ... | |

| | |
|---|---|
| life | 0.02 |
| evolve | 0.01 |
| organism | 0.01 |
| ... | |

| | |
|---|---|
| brain | 0.04 |
| neuron | 0.02 |
| nerve | 0.01 |
| ... | |

| | |
|---|---|
| data | 0.02 |
| number | 0.02 |
| computer | 0.01 |
| ... | |

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Figure taken from "Introduction to Probabilistic Topic Models", by D.M. Blei. (2011).

# An Extremely Brief Detour on Graphical Models

Graphical models are used to describe dependence / independence relationships between variables.

Two main types of graphical models:

1. Undirected graphical models which are also known as Markov networks.
2. Directed graphical models which are also known as Bayesian networks
    - belief networks, i.e. directed acyclic graphs (DAG's), are an important subclass.

Each node in graph corresponds to a random variable.

The edge structure of the graph (and edge direction in case of directed graphs) help determine the conditional independence / dependence relationships between random variables

- these relationships often enable inference, e.g. computation of conditional distributions, to be performed very efficiently.

Graphical models now very popular in statistics and machine learning.

# Directed Acyclic Graphs (DAGs)

There are no directed cycles in a DAG

- implies there is a node numbering such that any link from any node always goes to a higher numbered node.

Many efficient algorithms exist for performing inference in belief networks

- inference is the problem of "understanding" the conditional distribution of the graph when some nodes are observed

# Directed Acyclic Graphs (DAGs)

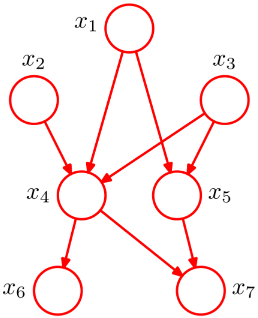


**Figure 8.2 from Bishop**

Note ordering of nodes in the DAG of Figure 8.2. This ordering can be used to write

$$\begin{aligned} p(x_1, x_2, \ldots, x_7) \quad &= \quad p(x_7 \mid x_4, x_5) \cdot p(x_6 \mid x_4) \cdot \\ & \quad\quad p(x_5 \mid x_1, x_3) \cdot p(x_4 \mid x_1, x_2, x_3 \\ & \quad\quad p(x_3) \cdot p(x_2) \cdot p(x_1). \end{aligned}$$

More generally for any DAG we have

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k \mid \mathsf{pa}(x_k)) \tag{13}$$

where pa$(x)$ denotes the "parents" of node $x_k$.

It's easy (why?) to simulate from a belief network using (13)

- simulating using representation in (13) is called ancestral sampling.

Not easy to simulate from conditional distribution when some nodes are observed

- but will see that Gibbs sampling easy to implement in that case.

# Dealing with Evidence in a Belief Network

Suppose now that $x_3$, $x_5$ and $x_6$ have been observed and we want to compute the conditional distribution of the unobserved variables.

Using (13) this conditional distribution satisfies

$$\begin{aligned} p(x_1, x_2, x_4, x_7 \,|\, x_3, x_5, x_6) &= \frac{p(x_1, x_2, x_3, x_4, x_5, x_6, x_7)}{p(x_3, x_5, x_6)} \\ &= \frac{p(x_1, x_2, x_3, x_4, x_5, x_6, x_7)}{\sum_{x_1, x_2, x_4, x_7} p(x_1, x_2, x_3, x_4, x_5, x_6, x_7)} \\ &= \frac{\prod_{k=1}^{7} p(x_k \mid \mathsf{pa}(x_k))}{\sum_{x_1, x_2, x_4, x_7} \prod_{k=1}^{7} p(x_k \mid \mathsf{pa}(x_k))} \qquad (14) \end{aligned}$$

where $x_3$, $x_5$ and $x_6$ are "clamped" at their observed values in (14).

Computing the normalizing factor, i.e. the denominator, in (14) can be computationally demanding — especially for very large DAGs.

Note also that the ordering of the original DAG (with no observed variables) is now lost. e.g. $x_1$ and $x_3$ are no longer independent once $x_5$ has been observed.

# Sampling Given Evidence in a Directed Acyclic Graphs

**Questions:** Can we use still ancestral sampling to simulate from $p(x_1, x_2, x_4, x_7 \,|\, x_3, x_5, x_6)$? If so, is it efficient?

**Question:** Can we simulate efficiently from from $p(x_1, x_2, x_4, x_7 \,|\, x_3, x_5, x_6)$?
**Solution:** Yes, using Gibbs sampling!

At each step of the Gibbs sampler we need to simulate from $p(x_i \mid \mathbf{x}_{-i})$ where any observed values in $\mathbf{x}_{-i}$ are clamped at these values throughout the simulation.

But it's easy to see (why?) that

$$p(x_i \mid \mathbf{x}_{-i}) = \frac{1}{Z} p(x_i \mid pa(x_i)) \prod_{j \in ch(i)} p(x_j \mid pa(x_j))$$

where $pa(x_i)$ and $ch(i)$ are the parent and children nodes, respectively, of $x_i$, and $Z$ is the (usually easy to compute) normalization constant

$$Z = \sum_{x_i} p(x_i \mid pa(x_i)) \prod_{j \in ch(i)} p(x_j \mid pa(x_j)).$$

The parents of $x_i$, the children of $x_i$ and the parents of the children of $x_i$ are known collectively as the Markov blanket of $x_i$.

# Oil Exploration and Inference Using a DAG

**e.g.** Consider the oil exploration example on the next slide:

A directed graphical model can be used to model the geology of a particular area below the seabed of the North Sea

- this geology is complex and locating oil requires both exploration and inference.

A decision has been made to drill at node A and the figures display the changes in probabilities of oil being present at every other node conditional on:

(i) oil being found at A (left-hand heat-map).

(ii) only partial oil being found at A (right-hand heat-map).

The probabilities, and therefore all of the changes in probabilities, can be estimated using Gibbs sampling as described above.

# Oil Exploration and Inference Using a DAG

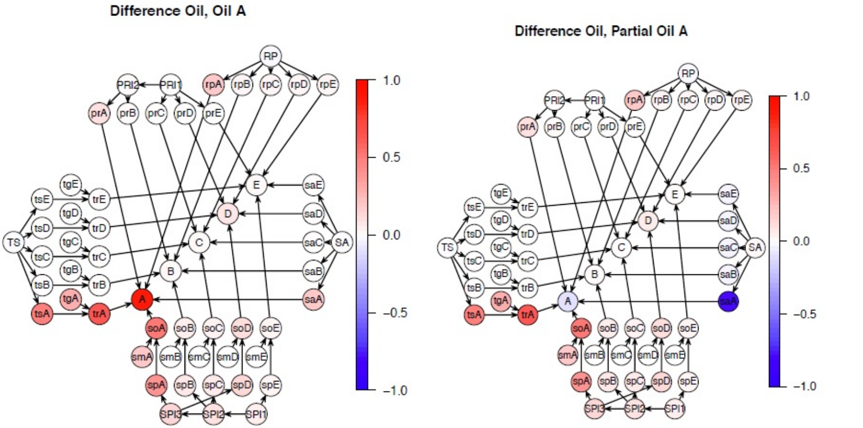


**Figure 4: Difference between conditional probabilities given evidence and prior probabilities. Evidence is observed in prospect A, and follows the explanations in section 3.4 . Figure shows the effect of Evidence 1 (left) and 2 (right).**

Figure taken from “Strategies for Petroleum Exploration Based on Bayesian Networks: a Case Study”, by Martinelli et al. (2012).

# Appendix: Bayesian Model Checking

After (successfully) confirming stationarity of Markov chains, can use samples to estimate quantities of interest.

But often this is just part of a bigger analysis. In particular often need to: (1) assess model performance and (2) choose among competing models.

There are many ways to assess model performance including:

1. Comparing posterior distributions of parameters to domain knowledge.
2. Simulating samples from the posterior predictive distribution and checking them for "reasonableness"
   - can do this by first simulating $\boldsymbol{\theta}$ from posterior distribution (already have these samples from the MCMC!) and then simulating $\mathbf{X}_{rep} \mid \boldsymbol{\theta}$
3. Posterior predictive checking: design test statistics of interest and compare their posterior predictive distributions (using simulated samples) to observed values of these test statistics
   - a form of internal model validation.

But see *Bayesian Data Analysis* (BDA) by Gelman et al. for discussion of model checking and examples.

# An Example of a Posterior Predictive Check (Gelman et al.)

Consider a sequence of binary outcomes $\mathbf{y} = [y_1, \ldots, y_n]$.

We model them as a specified number of IID Bernoulli trials with a uniform prior on the probability of success, $\theta$.

Let $s := \sum y_i$. Then posterior is

$$\begin{aligned} p(\theta \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \theta)p(\theta) \\ &= \theta^s(1-\theta)^{n-s} \end{aligned}$$

Recognize this as the Beta $\left(\sum y_i + 1, n - \sum y_i + 1\right)$ distribution.

The data is $y = [1\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$

- so $n = 20$ and $s = 7$.

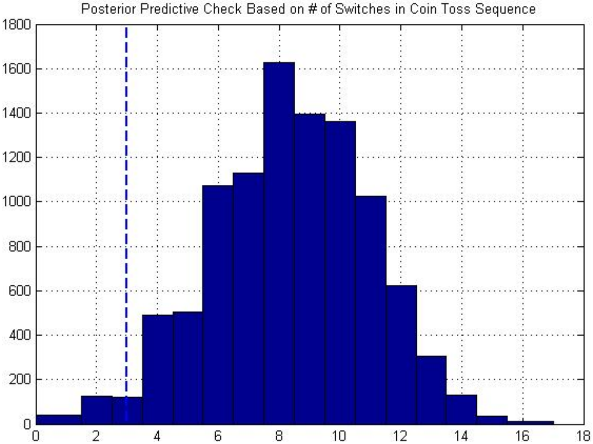
**Questions:** Is this a good model? Do the data look IID (given $\theta$)?

We note the sequence is strongly autocorrelated with $T(\mathbf{y}) = 3$ where $T(\cdot)$ counts the number of switches between $0$ and $1$.

So lets simulate $m$ samples $T(\mathbf{y}_1^{rep}), \ldots, T(\mathbf{y}_m^{rep})$ from posterior predictive dist.

- and compare them with $T(\mathbf{y})$.

# Appendix: An Example of a Posterior Predictive Check



We took $m = 10k$ and found only $2.8\%$ of the samples were $\leq T(\mathbf{y}) = 3$

- pretty strong evidence against the model!

Posterior predictive checks are a form of internal model validation and in this case suggests the model is inadequate and should be improved / expanded.

# Appendix: Bayesian Model Selection

Suppose now that we have several "good" models that have "passed" various posterior predictive checks etc.

**Question**: How do we pick the "best" model?

There are also several approaches to this mode selection problem:

1. Information criteria approaches that estimate an in-sample error and penalize the effective number of parameters, $p_D$.

   Two common criteria are:

   (i) The deviance information criterion (DIC)
      - only suitable for certain types of Bayesian models

   (ii) The Watanabe-Akaike information criterion (WAIC)
      - recently developed and more generally applicable than DIC
      - but not suitable for models where the data is dependent (given $\theta$) like time-series and spatial models.

Note that $p_D$ is a random variable that depends on the data
- is estimated differently for DIC and WAIC.

When comparing models, a smaller DIC or WAIC is better.

Both DIC and WAIC are easily estimated from the output of an MCMC
- important given the computational demands of Bayesian modeling.

# Appendix: Bayesian Model Selection

2. Bayesian cross-validation where the data is divided into $K$ folds

   Error on each fold computed by fitting model on remaining $K-1$ folds. Can be computed using:

   (i) mean-squared prediction error – requires predicted values of hold-out data
      - can use posterior predictive mean which can often be estimated from MCMC.

   (ii) the $\log$ posterior predictive distribution evaluated at the hold-out data.

   Cross-validation can clearly be computationally very demanding.

3. Bayes factors can also be useful when choosing among models.

   Given two models $H_1$ and $H_2$, the Bayes factor, $B(H_2; H_1)$, is

$$B(H_2; H_1) := \frac{p(\mathbf{X} \mid H_2)}{p(\mathbf{X} \mid H_1)} = \frac{\int_{\boldsymbol{\theta}_2} p(\mathbf{X} \mid \boldsymbol{\theta}_2, H_2) p(\boldsymbol{\theta}_2 \mid H_2)\, d\boldsymbol{\theta}_2}{\int_{\boldsymbol{\theta}_1} p(\mathbf{X} \mid \boldsymbol{\theta}_1, H_1) p(\boldsymbol{\theta}_1 \mid H_1)\, d\boldsymbol{\theta}_1}$$

   – not defined if priors $p(\boldsymbol{\theta}_i \mid H_i)$ not proper

   – in general need to estimate the two integrals.

Bayesian Model Averaging (BMA) is a related technique that performs inference using a weighted average of several "good" models
- weights are computed via Bayes factors.

# Appendix: Auxiliary Variable MCMC Methods

A real concern with MCMC methods is that the chain move through all areas of significant probability

- guaranteed in theory but in practice too many iterations may be required.

**e.g.** consider a Metropolis-Hastings algorithm with a local proposal distribution, i.e. a proposal unlikely to propose a candidate point $x_{t+1}$ that's far from $x_t$.

If the target distribution has many modes or "islands" of high density, then it will take a long time to move from one island to another.

But if we use a global proposal distribution, i.e. one with very large variance, then chance of landing on a high-density island is small.

Convergence diagnostics can help us determine if the MCMC chain for a specific application is converging too slowly.

But we could also use auxiliary variable MCMC methods such as Hamiltonian Monte-Carlo or the slice sampler

- they've become very popular in recent years and (with Gibbs sampling) have begun to render (basic) Metropolis-Hastings almost obsolete!

# Appendix: Hamiltonian Monte-Carlo

Hamiltonian Monte-Carlo is an MCMC method for continuous variables. It makes non-local jumps possible so we can jump from one mode to another.

To begin, we write the target distribution as

$$p(\mathbf{x}) = \frac{1}{Z_x} e^{H_x(\mathbf{x})}$$

where as usual $Z_x$ is unknown.

Now introduce a new auxiliary variable $\mathbf{y}$ with

$$p(\mathbf{y}) = \frac{1}{Z_y} e^{H_y(\mathbf{y})}$$

– almost always choose $\mathbf{y}$ to be Gaussian so that $H_y(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^\top \mathbf{y}$.

We also assume

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) = \frac{1}{Z_x Z_y} e^{H_x(\mathbf{x})+H_y(\mathbf{y})} = \frac{1}{Z} e^{H(\mathbf{x},\mathbf{y})}$$

where $Z := Z_x Z_y$ and $H(\mathbf{x}, \mathbf{y}) := H_x(\mathbf{x}) + H_y(\mathbf{y})$.

# Appendix: Hamiltonian Monte-Carlo

The goal is to define an MCMC algorithm for generating samples of $(\mathbf{x}, \mathbf{y})$ with the stationary distribution $p(\mathbf{x}, \mathbf{y})$

- once stationarity is reached we can simply discard the $\mathbf{y}$ samples.

The "trick" is to define the proposal distribution so that we can easily jump from one mode (of $p(\mathbf{x})$) to another.

Can do this as follows: given a current sample $(\mathbf{x}, \mathbf{y})$ we:

1. Simulate $\mathbf{y}'$ from $p(\mathbf{y})$
2. And then simulate $\mathbf{x}'$ from $p(\mathbf{x} \mid \mathbf{y}')$ using a Metropolis-Hastings sampler

We want the new sample $(\mathbf{x}', \mathbf{y}')$ to satisfy

$$H(\mathbf{x}', \mathbf{y}') \approx H(\mathbf{x}, \mathbf{y})$$

so that it will be accepted with high probability in the M-H algorithm.

We can achieve this by moving (approximately) along a contour of $H$ from $(\mathbf{x}, \mathbf{y})$ to $(\mathbf{x}', \mathbf{y}')$ where $(\mathbf{x}', \mathbf{y}') = (\mathbf{x} + \Delta\mathbf{x},\ \mathbf{y} + \Delta\mathbf{y})$.

# Appendix: Hamiltonian Monte-Carlo

A first-order Taylor approximation implies

$$\begin{aligned} H(\mathbf{x}', \mathbf{y}') &= H(\mathbf{x} + \Delta\mathbf{x},\ \mathbf{y} + \Delta\mathbf{y}) \\ &\approx H(\mathbf{x}, \mathbf{y}) + \boldsymbol{\nabla}_x H_x(\mathbf{x})^\top \Delta\mathbf{x} + \boldsymbol{\nabla}_y H_y(\mathbf{y})^\top \Delta\mathbf{y} \qquad (15) \end{aligned}$$

To move (approximately) along a contour of $H$ would like to set sum of last two terms in (15) to 0 – a $1$-dimensional constraint so many solutions possible.

Customary to use so-called Hamiltonian dynamics whereby

$$\Delta\mathbf{x} := \epsilon \boldsymbol{\nabla}_y H(\mathbf{y}) \qquad \text{and} \qquad \Delta\mathbf{y} := -\epsilon \boldsymbol{\nabla}_x H(\mathbf{x})$$

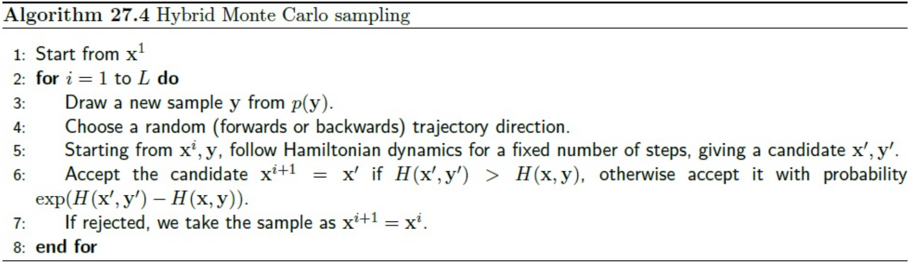
so that $H(\mathbf{x}', \mathbf{y}') \approx H(\mathbf{x}, \mathbf{y})$ as desired.

We take $L$ such Hamiltonian steps all with the same value of $\epsilon$ which is drawn randomly according to

$$\epsilon = \begin{cases} +\epsilon_0, & \text{with prob. } 0.5 \\ -\epsilon_0, & \text{with prob. } 0.5 \end{cases}$$

so that the proposal distribution, $Q(\cdot \mid \cdot)$, is symmetric.

# Appendix: Hamiltonian Monte-Carlo Algorithm



**Algorithm 27.4** Hybrid Monte Carlo sampling

```
Start from $\mathrm{x}^1$
for $i = 1$ to $L$ do
    Draw a new sample $\mathbf{y}$ from $p(\mathbf{y})$.
    Choose a random (forwards or backwards) trajectory direction.
    Starting from $\mathrm{x}^i, \mathrm{y}$, follow Hamiltonian dynamics for a fixed number of steps, giving a candidate $\mathrm{x}', \mathrm{y}'$.
    Accept the candidate $\mathrm{x}^{i+1} = \mathrm{x}'$ if $H(\mathrm{x}', \mathrm{y}') > H(\mathrm{x}, \mathrm{y})$, otherwise accept it with probability $\exp(H(\mathrm{x}', \mathrm{y}') - H(\mathrm{x}, \mathrm{y}))$.
    If rejected, we take the sample as $\mathrm{x}^{i+1} = \mathrm{x}^i$.
end for
```

– **Algorithm 27.4 from Barber**

# Appendix: Hamiltonian Monte-Carlo

The variable **x** has the interpretation of position and the auxiliary variable **y** has the interpretation of momentum.

Typically, **y** has the same dimension as **x** so there is one momentum variable for each space variable.

The Hamiltonian dynamics, i.e movement along a contour of $H$, can be implemented in a more sophisticated way than (15) via so-called Leapfrog discretization

- see, for example, Bishop for details.

In order to implement the algorithm we need to specify the parameters $L$ and $\epsilon_0$

- success of algorithm is quite sensitive to these choices
- improved versions of Hamiltonian MC choose these parameters adaptively
- and these versions are implemented in the new and popular STAN software
    - developed mainly by a team at Columbia University!

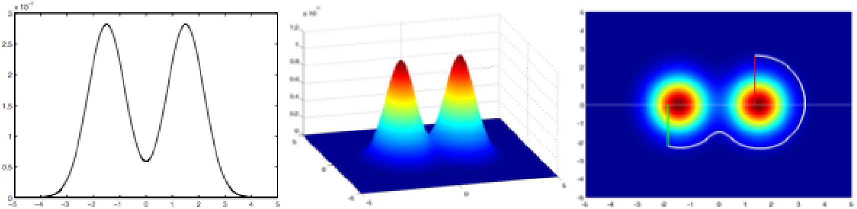


**Figure 27.9 from Barber**: Hybrid Monte Carlo. (a): Multi-modal distribution $p(x)$ for which we desire samples. (b): HMC forms the joint distribution $p(x)p(y)$ where $p(y)$ is Gaussian. (c): This is a plot of (b) from above. Starting from the point $x$, we first draw a $y$ from the Gaussian $p(y)$, giving a point $(x, y)$, given by the green line. Then we use Hamiltonian dynamics (white line) to traverse the distribution at roughly constant energy for a fixed number of steps, giving $x', y'$. We accept this point if $H(x', y') > H(x, y')$ and make the new sample $x'$ (red line). Otherwise this candidate is accepted with probability $\exp(H(x', y') - H(x, y'))$. If rejected the new sample $x'$ is taken as a copy of $x$.

# Empirical Bayes

| Claims $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Counts $y_x$ | 7840 | 1317 | 239 | 42 | 14 | 4 | 4 | 1 |
| Formula (19) | .168 | .363 | .527 | 1.33 | 1.43 | 6.00 | 1.25 | - |
| Gamma MLE | .164 | .398 | .633 | .87 | 1.10 | 1.34 | 1.57 | - |

Table displays counts $y_x$ of number of claims $x$ made in a single year by 9461 automobile insurance policy holders. Robbins' formula (19) estimates the number of claims expected in a succeeding year, for instance 0.168 for a customer in the $x = 0$ category. Parametric maximum likelihood analysis based on a gamma prior gives less noisy estimates.

Table displays one year's worth of claims data for a European insurance company. There were 9461 policy holders of whom 7840 made 0 claims, 1317 made 1 claim, 239 made 2 claims etc.

**Goal:** Estimate $\#$ of claims each policy holder will make next year.

# Empirical Bayes

Let $X_k$ denote the number of claims made in a single year by policy holder $k$.

Assume $X_k$ is Poisson with parameter $\theta_k$ so that

$$P(X_k = x) = p_{\theta_k}(x) := \frac{e^{-\theta_k}\theta_k^x}{x!}, \qquad x = 0, 1, 2, \ldots. \tag{16}$$

Assume the $\theta_k$'s are random with prior $g(\theta)$.

Consider now an individual customer with observed number of claims $x$. Then we have (why?)

$$\mathbb{E}[\theta \,|\, x] = \frac{\int_0^\infty \theta p_\theta(x) g(\theta)\, d\theta}{\int_0^\infty p_\theta(x) g(\theta)\, d\theta}. \tag{17}$$

Note that (17) also yields the expected number of claims made by the customer next year since (why?) $\mathbb{E}[\theta \,|\, x] = \mathbb{E}[X \,|\, x]$.

So (17) is what the insurance company needs to answer its question if it already knows the prior $g(\cdot)$.

**e.g.** If the company assumes $g$ is Gamma$(\nu, \sigma)$ with $\nu$ and $\sigma$ known, then no problem calculating (17)

- but how would we choose "good" values of $\nu$ and $\sigma$?

# Robbins' Approximation

A typical Bayesian approach would in fact assume they are unknown and would therefore place a hyper-prior (with known parameters) on $(\nu, \sigma)$.

In that case considerably more work required to compute $g$ and calculate (17).

Alternatively we can be a little clever! Using (16) and (17) we have

$$\begin{aligned}\mathbb{E}[\theta \mid x] &= \frac{\int_0^\infty \left[e^{-\theta}\theta^{x+1}/x!\right] g(\theta)\, d\theta}{\int_0^\infty \left[e^{-\theta}\theta^{x}/x!\right] g(\theta)\, d\theta} \\ &= \frac{(x+1)\, \int_0^\infty \left[e^{-\theta}\theta^{x+1}/(x+1)!\right] g(\theta)\, d\theta}{\int_0^\infty \left[e^{-\theta}\theta^{x}/x!\right] g(\theta)\, d\theta} \\ &= (x+1)\frac{f(x+1)}{f(x)} \qquad (18)\end{aligned}$$

where

$$f(x) = \int_0^\infty p_\theta(x) g(\theta)\, d\theta$$

is the marginal density of $X$.

# Robbins' Approximation

Clear from (18) that to answer the insurance company's question we only need $f(\cdot)$ and not $g(\cdot)$.

But we have a lot of data and can easily estimate $f(\cdot)$ directly to obtain Robbins' approximation

$$\begin{aligned}\widehat{\mathbb{E}}[\theta \,|\, x] &= (x+1)\frac{\hat{f}(x+1)}{\hat{f}(x)} \\ &= (x+1)\frac{y_{x+1}}{y_x} \qquad (19)\end{aligned}$$

with $y_x$ denoting $\#$ of observations with $x$ claims.

We see the values of $\hat{\mathbb{E}}[\theta \,|\, x]$ in the third row of the table.

Values at end of third row go awry because (19) becomes unstable at that point due to small count numbers in the data for policies that had 5 or more claims.

Can help resolve this issue by using a parametric empirical Bayesian approach in contrast to the non-parametric approach outlined above.

# Parametric Empirical Bayes

Now assume prior $g(\cdot)$ is Gamma$(\nu, \sigma)$ with

$$g(\theta) = \frac{\theta^{\nu-1} e^{-\theta/\sigma}}{\sigma^{\nu} \Gamma(\nu)}, \qquad \theta \geq 0$$

with $(\nu, \sigma)$ unknown.

Instead of placing a (hyper-) prior on $(\nu, \sigma)$ can estimate them from the data by explicitly computing (how?) the marginal density $f(x)$ which now has parameters $\nu$ and $\sigma$.

Then simply compute the MLE's $\hat{\nu}$ and $\hat{\sigma}$ to obtain

$$\hat{\mathbb{E}}[\theta \mid x] = (x+1) \frac{f_{\hat{\nu},\hat{\sigma}}(x+1)}{f_{\hat{\nu},\hat{\sigma}}(x)} \tag{20}$$

as our estimator – now see fourth row of table.