Computational Methods in Stochastics

Lecture V

MCMC and Bayesian Inference

This is strictly based on Ch 9 of Wilkinson.

(Bayesian inference: Ch 9 in the online book.)

MCMC & Bayesian Inference

Definition: A Markov chain Monte Carlo (MCMC) method for the simulation of a distribution f is any method producing an *ergodic* Markov chain $\left(X^{(t)}\right)$ whose stationary distribution is f.
Pragmatic definition of ergodic: sample & time averages are equal ↔ stationarity, equilibrium.

Goal: Statistical inference for the (model) parameters on the basis of experimental data.

<u>Bayesian Inference</u>

Discrete problems
We have various *hypotheses* $H_i, i = 1, 2, \dots, n$ on the measured stochastic process X. H_i form a **partition of the sample space**: $S = \bigcup_{1=1}^{n} H_i$, $H_i \cap H_j = \emptyset, \forall i \neq j$, and $\mathrm{P}(H_i) > 0, \forall i$.

MCMC & Bayesian Inference

We observe some measurement *outcome* $X = x$ and **we are interested in the probabilities of the hypotheses *conditional* on the outcome**, $P(H_i|X = x)$. To compute these probabilities we need *prior* probabilities of the hypotheses, $P(H_i)$.

To update our prior beliefs about the hypotheses $P(H_i)$ we use Bayes Theorem to compute our *posterior beliefs* based on the occurrence of $X = x$:

$$P(H_i|X = x) = \frac{P(X = x|H_i)P(H_i)}{\sum_{j=1}^{n} P(X = x|H_j)P(H_j)}, \qquad i = 1, \dots, n.$$

$P(X = x|H_i) = L(H_i; x)$ are *likelihoods*. **The likelihood** is not a probability mass function PMF for H_i: it **does not sum to 1**.

MCMC & Bayesian Inference

Continuous and mixed problems

Bayes Theorem $$\pi(\theta|X=x)=\frac{\pi(\theta)L(\theta;x)}{\int_\theta P(X=x|\theta')\pi(\theta')d\theta'}$$

For discrete outcome, the likelihood $L(\theta;x)=P(X=x|\theta)$ is a function of θ for given fixed x. ($L(\theta;x)$ is not a probability density.)

For a continuum of hypotheses **the incalculable denominator** $\int_\theta P(X=x|\theta')\pi(\theta')d\theta'$ simply represents a constant of proportionality, and we write

$$\pi(\theta|X=x)\propto\pi(\theta)L(\theta;x).$$

Slogan:
"The posterior is proportional to the prior times the likelihood."

MCMC & Bayesian Inference

And of course the notation has to be different with different authors ☹
In the online book: Likelihood - $L(x;\theta)$

MCMC & Bayesian Inference

Example (discrete outcome). For a particular gene transcription, events occur according to **a Poisson process** at **the rate** θ per minute. Prior to carrying out an experiment, an expert specifies his **opinion regarding** $\boldsymbol{\theta}$ as $\theta \sim Ga(a, b)$ (gamma distribution) with $a = 2, b = 1$. Counts of the number of transcripts are gathered from n separate one-minute intervals to get data $x = (x_1, x_2, \dots, x_n)^T$.

The likelihood for θ:

$$L(\theta; x) = P(x|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

$$\Rightarrow L(\theta; x) \propto \prod_{i=1}^{n} \theta^{x_i} e^{-\theta} = e^{\sum_{i=1}^{n} x_i} e^{-n\theta}$$

***L* is seen to depend on data only through** n **and** $\bar{x} = (1/n) \sum_{1=1}^{n} x_i$, so n and $\bar{x}$ are said to be *sufficient statistics* for the likelihood function.

MCMC & Bayesian Inference

The priori $\pi(\theta) \propto Ga(a,b) = \theta^{a-1}e^{-b\theta}$.

The posteriori

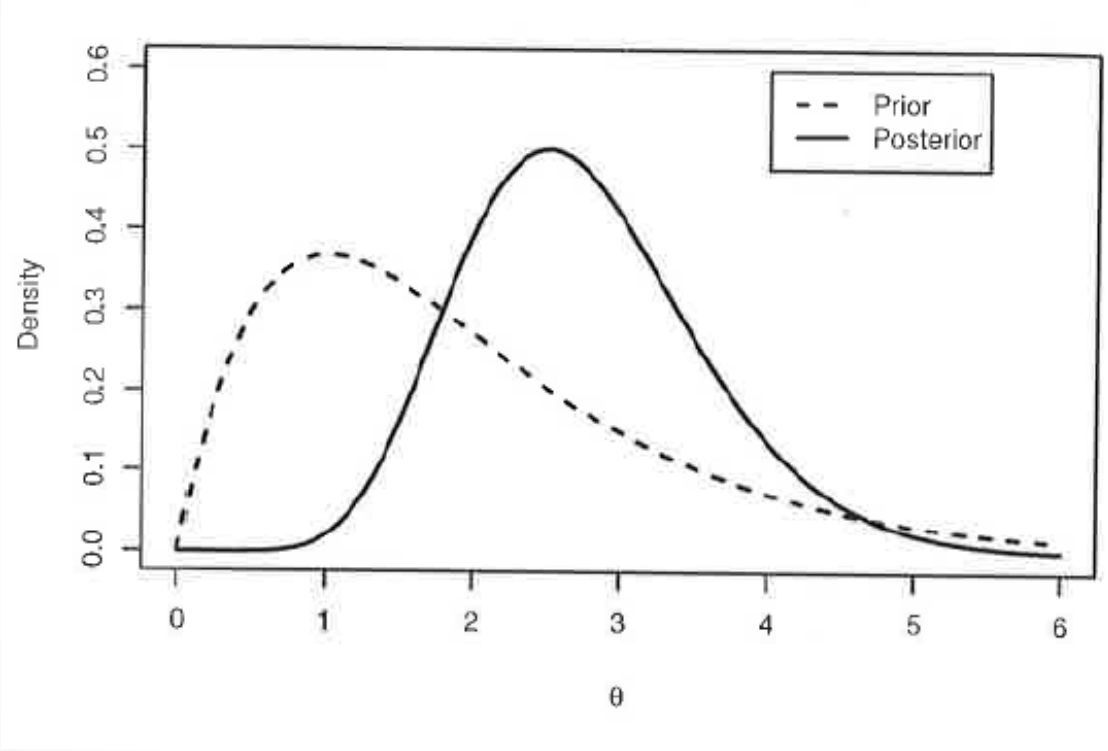
$$\pi(\theta|x) \propto \pi(\theta)L(\theta;x) \propto e^{a+\sum_{i=1}^{n} x_i - 1}e^{-(b+n)\theta}$$

$$\Rightarrow \theta \sim Ga\left(a + \sum_{i=1}^{n} x_i, b + n\right).$$

Starting with a gamma prior results in a gamma posteriori; such problems are said to be *conjugate*. Gamma is conjugate for the Poisson likelihood.

MCMC & Bayesian Inference

For example, observation $x = (4, 2, 3) \Rightarrow \theta \sim Ga(11, 4)$ posteriori.



(For continuous X, in the likelihood PMF → PDF.)

Note: In *statistical inference* we are interested in the distribution of the parameters θ (e.g. mean and variance) describing the distribution of the data.

MCMC & Bayesian Inference

Bayesian computation

The previous example includes in principle everything within Bayesian inference: **the posterior is the conditional distribution for the parameters given the data**. In nontrivial problems, computational effort is quite a bit more substantial.

Problem 1: choosing the constant of proportionality such that the density integrates to 1.

- for non-standard densities, one needs to integrate the product of the likelihood and the prior (the *kernel* of the posterior) over the support of θ, which may be infinite and/or multidimensional. (MAP vs. MLE, see 9.1.3)

Problem 2: In a multi-dimensional parameter space we want to know what the marginal distribution of each component looks like. → A hard numerical integration problem.

MCMC & Bayesian Inference

Example (see the detailed derivations in SMSB, Section 9.1.2 – the first edition; 10.1.2 the new edition)

We have a collection of observations, X_i, which we believe to be iid (independent identically distributed) normal with unknown mean and precision τ ($= 1/\sigma^2$): $X_i|\mu,\tau \sim N(\mu, 1/\tau)$.

The **likelihood** for a single observation:

$$L(\mu,\tau;x_i) = f(x_i|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}}\exp\left\{-\frac{\tau}{2}(x_i-\mu)^2\right\}.$$

$\rightarrow$ the likelihood for n independent observations $x = (x_1,\ldots,x_n)^{\mathrm{T}}$:

$$L(\mu,\tau;x) = f(x|\mu,\tau) = \prod_{i=1}^{n}\sqrt{\frac{\tau}{2\pi}}\exp\left\{-\frac{\tau}{2}(x_i-\mu)^2\right\}.$$

MCMC & Bayesian Inference

$$\Rightarrow L(\mu, \tau; x) \propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}[(n-1)s^2 + n(\bar{x}-\mu)^2]\right\},$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

To proceed with the Bayesian analysis we need **prior distributions** for the parameters θ: (μ, τ).

There exists a conjugate analysis based on the specifications $\tau \sim Ga(a, b)$ and $\mu|\tau \sim N\left(c, \frac{1}{d\tau}\right)$, but in this μ and τ are not independent.

MCMC & Bayesian Inference

Alternatively, one may specify $\tau \sim Ga(a, b)$ and $\mu|\tau \sim N\left(c, \frac{1}{d}\right)$, but then conjugacy is lost and the analysis becomes intractable.

$$\pi(\mu, \tau|x)$$
$$\propto \tau^{\alpha+\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2}[(n-1)s^2 + n(\bar{x}-\mu)^2] - \frac{d}{2}(\mu - c)^2 - b\tau\right\}$$

In other words, the posterior $\pi(\mu, \tau|x)$ will not factorise, because in it μ and τ are not independent ("are not independent *a posteriori*"). → There's no way of working out the marginal posterior distributions for μ and τ.

Consequently, we **need a way to understand posterior densities without being able to analytically integrate the posterior density**.

This is where **Markov Chain Monte Carlo (MCMC)** algorithms like the Gibbs sampler and Metropolis- Hastings method come into play.

MCMC & Bayesian Inference

The Gibbs Sampler

The Gibbs sampler can be used for simulating from multivariate distributions *when one is able to simulate from conditional distributions*.

In the previous example of a normally distributed random sample the posterior was found to be

$$\pi(\mu, \tau|x)$$
$$\propto \tau^{\alpha+\frac{n}{2}-1}\exp\left\{-\frac{\tau}{2}[(n-1)s^2 + n(\bar{x}-\mu)^2] - \frac{d}{2}(\mu - c)^2 - b\tau\right\}$$

This problem is said to be *semi-conjugate,* because by picking out terms in the variable of interest and regarding everything else as a constant of proportionality we get $\pi(\mu, \tau|x)$ and $\pi(\tau|\mu, x)$ in standard forms.

MCMC & Bayesian Inference

$$\tau|\mu, x \sim Ga\left(a + \frac{n}{2}, b + \frac{1}{2}[(n-1)s^2 + n(\bar{x} - \mu)^2]\right),$$

$$\mu|\tau, x \sim N\left(\frac{cd + n\tau\bar{x}}{n\tau + d}, \frac{1}{n\tau + d}\right).$$

If we can simulate normal and gamma quantities, we can simulate from the full conditionals. → We need a way to **simulate** from the joint density - and marginals – **based only on the ability to sample from the full conditionals**.

MCMC & Bayesian Inference

Sampling from bivariate densities (prelude to the Gibbs algorithm)

For a bivariate density,

A. $\pi(x, y) = \pi(x)\pi(y|x)$

B. $\pi(x, y) = \pi(y)\pi(x|y)$

To simulate from $\pi(x, y)$:

A: Simulate $X = x$ from $\pi(x)$, then simulate $Y = y$ from $\pi(y|x)$.

B: Simulate $Y = y$ from $\pi(y)$, then simulate $X = x$ from $\pi(x|y)$.

Supposing we can simulate from **one** of the marginals, the scheme is:

1. Get (x, y) from the bivariate density by first simulating $X = x$ from $\pi(x)$, then $Y = y$ from $\pi(y|x)$. We have (x, y).
2. $Y = y$ must be from $\pi(y) \rightarrow$ get x' from $\pi(x'|y)$. We have (x', y).
3. x' is from $\pi(x) \rightarrow$ get y' from $\pi(y|x')$. We have (x', y').
4. Keep going…

MCMC & Bayesian Inference

The previous scheme defines a **bivariate Markov chain**. Obviously, $\pi(x, y)$ is **stationary**, since consequent samples are drawn from it ad infinitum.

The transition kernel for this Markov chain is

$$p\big((x, y), (x', y')\big) = \pi(x', y'|x, y) = \pi(x'|x, y)\pi(y'|x', x, y) = \pi(x'|y)\pi(y'|x').$$

MCMC & Bayesian Inference

The Gibbs sampling algorithm

The density of interest: $\pi(\theta)$, where $\theta = (\theta_1, \ldots, \theta_d)^{\mathrm{T}}$.

The full conditionals:

$$\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d) = \pi(\theta_i|\theta_{-i}) = \pi_i(\theta_i), \qquad i = 1, \ldots, d$$

The Gibbs sampler: (notation, θ_{-i} : all but θ_i)

1. Initialise the counter to $j = 1$ and the state to
   $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_d^{(0)}\right)^{\mathrm{T}}$.
2. Obtain a new value $\theta^{(j)}$ from $\theta^{(j-1)}$ by
   $$\theta_1^{(j)} \sim \pi\left(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_d^{(j-1)}\right)$$
   $$\theta_2^{(j)} \sim \pi\left(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}\right)$$
   $$\vdots$$
   $$\theta_d^{(j)} \sim \pi\left(\theta_d|\theta_1^{(j)}, \ldots, \theta_{d-1}^{(j)}\right).$$
3. Change j to $j + 1$ and return to step 2.

MCMC & Bayesian Inference

Please note that the notation is a bit misleading.

$$\pi(\theta_i|\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_d) = \pi(\theta_i|\theta_{-i}) = \pi_i(\theta_i), \qquad i = 1, \dots, d$$

is actually

$$\pi(\theta_i|\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_d, x) = \pi(\theta_i|\theta_{-i}, x) = \pi_i(\theta_i), i = 1, \dots, d$$

In other words, these are probabilities for a distribution parameter conditional on the value of other parameters **and the data.** However, the data is fixed, so one omits it from the notation and in Gibbs is represented by e.g. mean and variance. And once more: θ_i are **parameters of the distribution describing the data, and we want to find the marginal distributions of these parameters**. →

MCMC & Bayesian Inference

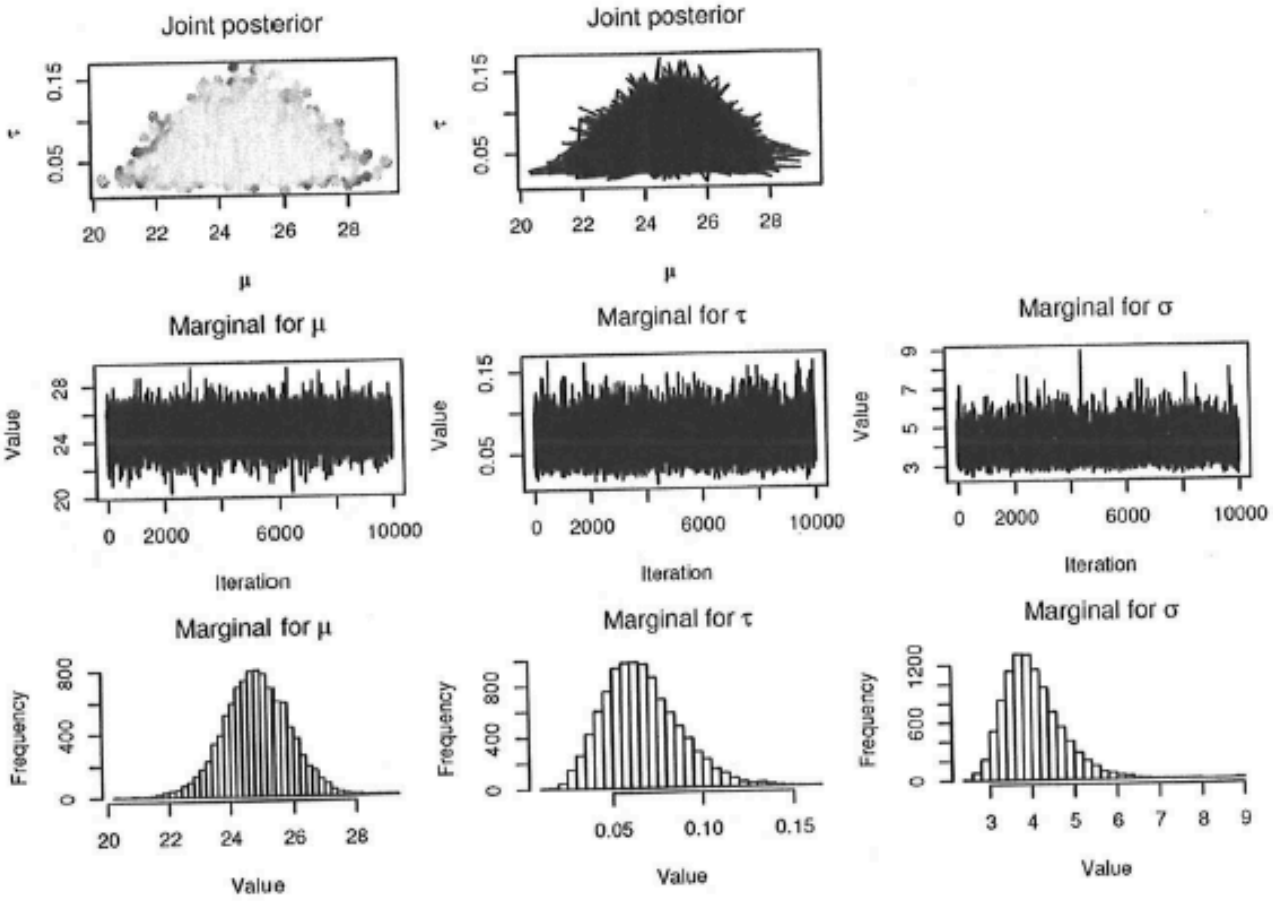


Figure 9.4 *Figure showing the Gibbs sampler output resulting from running the example code in Figure 9.3. The top two plots give an indication of the bivariate posterior distribution. The second row shows trace plots of the marginal distributions of interest, indicating a rapidly mixing MCMC algorithm. The final row shows empirical marginal posterior distributions for the parameters of interest.*

MCMC & Bayesian Inference

The procedure defines a **homogeneous Markov chain**: each simulated value depends only on the previous simulated value and not the iteration counter j.

In the book the distribution of the chain is shown to be **stationary**, i.e. $\pi(\phi) = \int_S p(\theta, \phi)\pi(\theta)d\theta$, where $p(\theta, \phi) = \prod_{i=1}^{d} \pi(\phi_i|\phi_1, \dots, \phi_{i-1}, \theta_{i+1}, \dots, \theta_d)$ is the transition kernel.
Note: ϕ_i is the updated value for the component i, that is, ϕ_i replaces θ_i in the update.

This fixed-sweep Gibbs sampler is **not reversible**, so detailed balance cannot be used to check for reversibility.

MCMC & Bayesian Inference

Reversible Gibbs samplers

It is the **fixed sweep** that makes the previous Markov chain irreversible. Each component update is reversible. To remedy this, one can pick components at random or, even more simply, scan the components first in order and then in the reversed order.

Since reversibility is not a requirement of a useful algorithm and the fixed-sweep Gibbs has better convergence properties and is easiest to implement, fixed sweep is often used.

MCMC & Bayesian Inference

The Metropolis-Hastings algorithm

When one cannot simulate full conditionals, i.e. the prior, one can **propose an initial distribution**. If $\pi(\theta)$ is the density of interest and $q(\theta, \phi)$ is the *proposal distribution*, we can construct the following algorithm:

1. Initialise the counter to $j = 1$ and the state to $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_d^{(0)}\right)^{\mathrm{T}}$.
2. Generate a proposal value ϕ using the **kernel** $q(\theta^{(j-1)}, \phi)$.
3. Evaluate the **acceptance probability** $\alpha(\theta^{(j-1)}, \phi)$ of the proposed move, where $\alpha\left(\theta^{(j-1)}, \phi\right) = \min\left\{1, \frac{\pi(\phi)q(\phi,\theta)}{\pi(\theta)q(\theta,\phi)}\right\}$.
4. Put $\theta^{(j)} = \phi$ with probability $\alpha(\theta^{(j-1)}, \phi)$, else put $\theta^{(j)} = \theta^{(j-1)}$.
5. Change j to $j + 1$ and return to step 2.

The Markov chain defined above is reversible.

MCMC & Bayesian Inference

Any distribution $q(\cdot)$can be used for simulating the proposals. Typically, $N(\mu, 1)$ is used, where the mean μ is chosen as the current state θ. So, note that in this procedure $q(\cdot)$ changes at each step. When using $q(\cdot) = N(\cdot)$, the Metropolis-Hastings samples the distribution like a random walker: the distance covered grows as $\sqrt{k}$, where k labels the steps. There are ways to make the sampling more efficient, for example HMC.

Note: The goal may not be the same as previously in Gibbs - to obtain distributions for θ_i. Instead, in this version of Metropolis-Hastings the distribution $\pi(\theta)$ is simulated, which means that *here* θ are the samples from the distribution, $\theta \sim \pi(\theta)$. The notation is confusing: previously θ was strictly interpreted as parameters of the target distribution, here this does not necessarily hold. (Of course Gibbs sampling could be used for direct sampling as well (meaning $X = \theta$). And, of course, M-H is used such that θ are parameters of a distribution.)

MCMC & Bayesian Inference

What's the use?

So, what use do we have of these distributions of θ then? One can, for example "predict" future $\tilde{x}$: $p(\tilde{x}|x) = \int p(\tilde{x}, \theta) d\theta$.

In the case $\theta = (\mu, \tau)$:

$$\pi(\tilde{x}|x) = \int \pi(\tilde{x}, \mu, \tau|x) d\mu d\tau = \int \pi(\tilde{x}|\mu, \tau, x)\pi(\mu, \tau|x) d\mu d\tau =$$

$$= \int \pi(\tilde{x}|\mu, \tau)\pi(\mu, \tau|x) d\mu d\tau.$$

So, use this for generating $\tilde{x}$.

How does one choose the transition kernel? →

MCMC & Bayesian Inference

The Metropolis method

(Here again θ, ϕ are **not** parameters for distribution π.)

This is the simplest case of Metropolis-Hastings, historically preceding it. Here **the proposal is symmetric**, $q(\theta, \phi) = f(|\theta - \phi|) = q(\phi, \theta)$.

The acceptance probability then simplifies to

$$\alpha(\theta, \phi) = \min\left\{1, \frac{\pi(\phi)}{\pi(\theta)}\right\}.$$

Proposed moves that take the chain to a region of higher density are always accepted and moves that take the chain to a region of lower density are accepted with probability equalling the ratio of the density of the proposed state and the density of the present state.

MCMC & Bayesian Inference

Random walk chains

The proposed value ϕ at stage j is $\phi = \theta^{(j-1)} + w_j$, where w_j are iid random variables, independent of the state of the chain. Accept/reject according to $\alpha(\theta, \phi)$.

If the w_j have density $f(\cdot)$, from which it is easy to simulate, we can simulate an *innovation* w_j and set the *candidate* point to $\phi = \theta^{(j-1)} + w_j$.

If the transition kernel $q(\theta, \phi) = f(\phi - \theta)$ is symmetric about zero, then we have a symmetric chain. → The acceptance probability does not depend on $f(\cdot)$.

How should $f(\cdot)$ be chosen?

Simple distribution: E.g. uniform or normal. Parameters (e.g. variance) of $f(\cdot)$ should also be judiciously chosen to sample efficiently.

Rule of thumb: Acceptance rate of appr. 30 % works.

MCMC & Bayesian Inference

The variation of the innovation must be selected appropriately for good mixing: Too small $\sigma^2 \rightarrow$ high acceptance but movement is small. Too large $\sigma^2 \rightarrow$ moves about but low acceptance.

Here, innovations from $U(-\alpha, \alpha)$. (So, α in the caption is **not** the acceptance ratio.)

$\rightarrow$

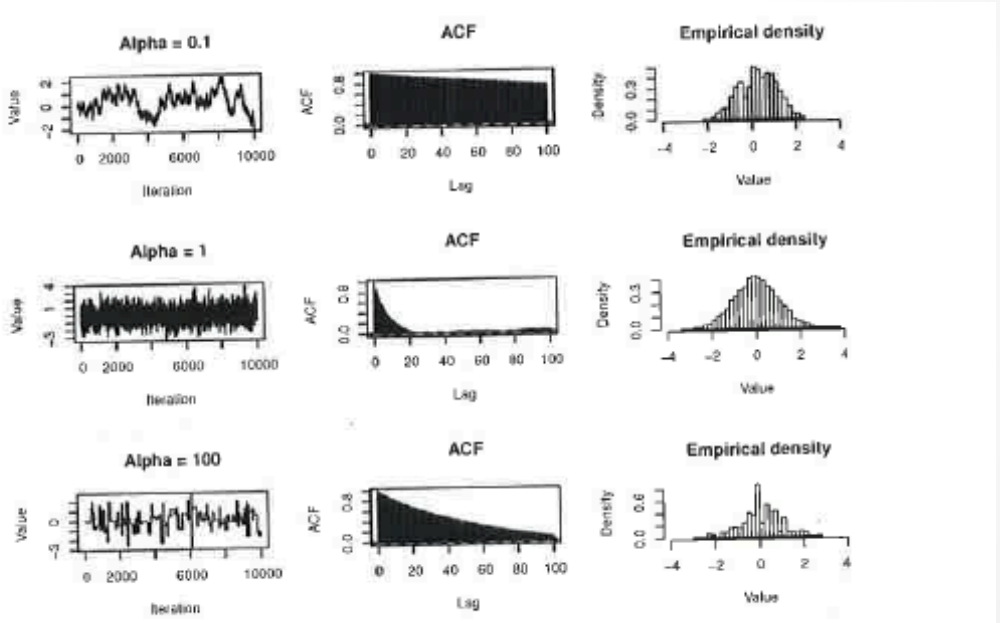


Figure 9.6 *Output from the Metropolis sampler given in Figure 9.5. The top row shows the result of running the chain with* $\alpha = 0.1$, *corresponding to a chain that is too cold. The middle row shows the results for* $\alpha = 1$. *This* α *is close to optimal, and the ACF plot shows auto-correlations in the sampled values decaying away rapidly to zero. The final row shows the results for* $\alpha = 100$, *representing a chain that is too hot, with many rejected proposed moves.*

MCMC & Bayesian Inference

Independence chains

Here the transition kernel does not depend on previous state, i.e. $q(\theta, \phi) = f(\phi)$ for some density $f(\cdot)$.

move towards larger $\pi(\cdot)/f(\cdot)$

The acceptance probability is $\downarrow$

$$\alpha(\theta, \phi) = \min\left\{1, \frac{\pi(\phi)}{\pi(\theta)} \Big/ \frac{f(\phi)}{f(\theta)}\right\} \left(= \min\left\{1, \frac{\pi(\phi)/f(\phi)}{\pi(\theta)/f(\theta)}\right\}\right).$$

Note that there is dependence between θ and ϕ via $\alpha(\theta, \phi)$.

So, the more similar $f(\cdot)$ is to $\pi(\cdot)$, the larger is α. Here we want to **maximise** α, because the sampling is "direct".

In the context of Bayesian inference, choosing the prior density as the proposal density:

$$\alpha(\theta, \phi) = \min\left\{1, \frac{L(\phi; x)}{L(\theta; x)}\right\}.$$

MCMC & Bayesian Inference

Bayesian inference

When using the Metropolis-Hastings sampler in a "real" Bayesian inference problem, meaning you infer for a parameter vector θ given some data x generated from a probability model of the form $\pi(x|\theta)$, you factorise the joint distribution as

$$\pi(\theta, x) = \pi(x|\theta)\pi(\theta)$$

We compute the **posterior distribution** $\pi(\theta|x) \propto \pi(\theta, x)$. We need not care about the proportionality (normalisation constant), because in Metropolis-Hastings only ratios $\pi(\phi)/\pi(\theta)$ appear.

$\pi(x|\theta)$ is something you use your understanding (oftentimes called 'domain expertise') to come up with.

MCMC & Bayesian Inference

Then we construct a Metropolis-Hastings scheme that targets $\pi(\theta|x)$. We need a proposal kernel $q(\theta, \theta^*)$, which can be arbitrary:. This proposes a move from θ to θ^*, which we either accept or reject with probability $\alpha(\theta, \theta^*) = \min\{1, A\}$, where…

$$A = \frac{\pi(\theta^*)\pi(x|\theta^*)q(\theta^*, \theta)}{\pi(\theta)\pi(x|\theta)q(\theta, \theta^*)}.$$

(Compare this with the previous $A = \frac{\pi(\phi)q(\phi,\theta)}{\pi(\theta)q(\theta,\phi)}$.)

"Missing data is parameters in Bayesian: anything you want to infer." (This is the latent variable stuff.)

In **stan** you can either type distributions in or use library distributions.

MCMC & Bayesian Inference

Bayesian inference for latent variable models

We may infer **missing data** (called "data augmentation") by using actual observations. Denoting the actual data by y and supposing we can deduce missing data x indirectly, we write

$$\pi(\theta, x, y) = \pi(\theta)\pi(x|\theta)\pi(y|x, \theta).$$

Now we use this joint distribution as the basis of inference.

In Metropolis-Hastings we have

$$A = \frac{\pi(\theta^*)\pi(y|\theta^*)q(\theta^*, \theta)}{\pi(\theta)\pi(y|\theta)q(\theta, \theta^*)}, \quad \text{where } \pi(y|\theta) = \int_X \pi(y|x, \theta)\pi(x|\theta)dx.$$

Marginalising over x is impossible → there are techniques to tackle this, but this is way beyond the scope here.

MCMC & Bayesian Inference

Epilogue

As you can see, Metropolis-Hastings has much in common the envelope method. Just like in the envelope method the **proposal distribution should satisfy** $f(\cdot) \geq \pi(\cdot)$ over the relevant range.

In comparison to envelope sampling, the acceptance ratio of independence Metropolis-Hastings is better. For proof see e.g. **Robert, Casella: Monte Carlo Statistical methods** (Springer). This book contains more proofs and details you ever care to know, but is formal and presents straightforward stuff in a complicated manner.