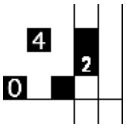
Steven R. Dunbar
Department of Mathematics
203 Avery Hall
University of Nebraska-Lincoln
Lincoln, NE 68588-0130
http://www.math.unl.edu
Voice: 402-472-3731
Fax: 402-472-8466

Topics in Probability Theory and Stochastic Processes Steven R. Dunbar

---

The Gibbs Sampler



Rating

Mathematically Mature: may contain mathematics beyond calculus with proofs.

Section Starter Question

Key Concepts

1. The Gibbs sampler is an algorithm for generating random variables from a marginal distribution indirectly, without having to calculate the density. Gibbs sampling is based only on elementary properties of Markov chains.

2. The simple case of a 2×2 table with multinomial sampling clearly illustrates the Markov chain nature of the process.

3. The simple case of a bivariate normal table with Gibbs sampling clearly illustrates Gibbs sampling for continuous distributions.

4. A simple Bayesian model for a spam filter illustrates typical Gibbs sampling.

5. Suppose $f(x_1, x_2, \ldots, x_N)$ is a probability distribution in which the variables represent parameters of a statistical model. Gibbs sampling obtains point and interval estimates for these parameters.

6. Suppose $X \sim N(\mu, 1/\tau)$ with μ and τ unknown. Based on a reasonably sized sample, Gibbs sampling obtains the posterior distributions of μ and τ.

7. Hierarchical Bayesian models naturally describe the connections between data, observed parameters and other unobserved parameters. A simple three-level hierarchical model uses Bayes' rule to bind together data, X, a parameter to be estimated, λ, and an additional hyperparameter, β.

8. In image degradation, a version of the Metropolis algorithm called **Gibbs sampling** the configuration maximizes $\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right]$, called the **maximum a posteriori estimate**.

Vocabulary

1. The **Gibbs sampler** is a technique for generating random variables from a marginal distribution indirectly, without having to calculate the density.

2. **Gibbs sampling** obtains a Gibbs sequence iteratively by alternately generating values from

$$X'_j \sim f(x \mid Y'_j = y'_j)$$
$$Y'_{j+1} \sim f(y \mid X'_j = x'_j).$$

3. The functional correspondence between the Beta families for the prior and the posterior and the Binomials for the likelihood makes them a **conjugate pair**.

4. **Hierarchical Bayesian models** naturally describe the connections between data, observed parameters and other unobserved parameters, sometimes called **latent variables**. A simple three-level hierarchical model uses Bayes' rule to bind together data, X, a parameter to be estimated, λ, and an additional hyper-parameter, β.

5. The configuration maximizing the probability of a configuration given a randomly changed or blurred configuration, $\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right]$, is the **maximum a posteriori estimate**.

6. A version of the Metropolis algorithm called **Gibbs sampling**. maximizes $\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right]$, called the **maximum a posteriori estimate**.

7. A probability distribution whose conditional probabilities depend on only the values in a neighborhood system is called a **Gibbs distribution** and is part of a larger notion called a **Markov random field**.

Mathematical Ideas

General Theory of Gibbs Sampling

The **Gibbs sampler** is an algorithm for generating random variables from a marginal distribution indirectly, without having to calculate the density. This subsection illustrates the algorithm by exploring several examples. In such cases, Gibbs sampling is based only on elementary properties of Markov chains.

Given a joint density $f(x, y_1, \ldots, y_p)$, the goal is to find characteristics of the marginal density

$$f_X(x) = \int \cdots \int f(x, y_1, \ldots, y_p)\ \mathrm{d}y_1 \ldots\ \mathrm{d}y_p,$$

such as the mean or variance. Often the integrations are extremely difficult to do, either analytically or numerically. In such cases the Gibbs sampler provides an alternative method for obtaining $f_X(x)$.

Gibbs sampling effectively generates a sample $X_1, \ldots, X_m \sim f(x)$ without requiring $f(x)$. By simulating a large enough sample, the mean, variance, or any other characteristic of $f(x)$ can be calculated to the desired degree of accuracy.

First consider the two-variable case. Starting with a pair of random variables (X, Y), the Gibbs sampler generates a sample from $f(x)$ by sampling instead from the conditional distribution $f(x \mid y)$ and $f(y \mid x)$, distributions often already known in statistical models. This is done by generating a *Gibbs sequence* of random variables

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots, Y_k', X_k'. \tag{1}$$

The initial value $Y_0' = y_0'$ is specified and the rest of the sequence is obtained

iteratively by alternately generating values from

$$X'_j \sim f(x \mid Y'_j = y'_j)$$
$$Y'_{j+1} \sim f(y \mid X'_j = x'_j).$$

The generation of the sequence is **Gibbs sampling**. Under reasonable general conditions, the distribution of X'_k converges to $f_X(x)$, the true marginal of X, as $k \to \infty$. Thus for k large enough, the last observation X'_k is effectively a sample point from $f_X(x)$.

Bivariate Binomial Example

This section discusses Gibbs sampling in detail for the simplest case of a 2×2 table. Suppose X and Y are each marginally Bernoulli random variables with joint distribution

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} p_1 & p_2 \\ p_3 & p_4 \end{pmatrix} \end{array}$$

or in terms of the joint probability distribution

$$\begin{pmatrix} f_{X,Y}(0,0) & f_{X,Y}(1,0) \\ f_{X,Y}(0,1) & f_{X,Y}(1,1) \end{pmatrix} = \begin{pmatrix} p_1 & p_2 \\ p_3 & p_4 \end{pmatrix}.$$

For the distribution, the marginal distribution of x is given by

$$f_X = (f_X(0), f_X(1)) = (p_1 + p_3, p_2 + p_4),$$

a Bernoulli distribution with success probability $p_2 + p_4$. The conditional probabilities can be expressed in two matrices

$$A_{y|x} = \begin{pmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{pmatrix}$$

and

$$A_{x|y} = \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix}$$

where $A_{y|x}$ has the conditional probabilities of Y given X and $A_{x|y}$ has the conditional probabilities of X given Y.

As an example, to generate the marginal distribution of X uses the X' sequence from (**??**). From X'_0 to X'_1 goes through Y'_0, so the iteration sequence is $X'_0 \to Y'_1 \to X'_1$ and $X'_0 \to X'_1$ is a two-stage Markov chain, with transition probability

$$\mathbb{P}\left[X'_1 \mid X'_0\right] = \sum_y \mathbb{P}\left[X'_1 \mid Y'_1 = y\right] \times \mathbb{P}\left[Y'_1 = y \mid X'_0\right].$$

For example, to go from $x = 1$ to $x = 0$ takes the dot product of the second row of $A_{y|x}$ with the first column of $A_{x|y}$. Generally, $\mathbb{P}\left[X'_1 \mid X'_0\right]$ is the right matrix multiplication of $A_{y|x}$ by $A_{x|y}$, so

$$A_{x|x} = A_{y|x} A_{x|y}$$

is the transition probability matrix for the X' sequence. The matrix that gives $\mathbb{P}\left[X'_k = x_k \mid X'_0 = x_0\right]$ is $(A_{x|x})^k$. Letting $f_k = (f_k(0), f_k(1))$ denote the marginal probability distribution of X'_k then $f_k = f_0 A^k_{x|x} = f_0 (A^{k-1}_{x|x}) A_{x|x} = f_{k-1} A_{x|x}$. By the Fundamental Theorem for Markov Chains, $f_k \to f$ as $k \to \infty$ with stationary distribution f satisfying $f A_{x|x} = f$. If the Gibbs sequence converges, the f satisfying $f A_{x|x} = f$ must be the marginal distribution of X. In this small example, it is straightforward to check that $f_X = (p_1 + p_3, p_2 + p_4)$ satisfies $f_X A_{x|x} = f_X$. So stopping the iteration scheme at a large enough value of k gives approximately f_X. The larger the value of k, the better the approximation. No general guidance on choosing such k is available. However, one possible approach is to monitor density estimates from m independent Gibbs sequences, and choosing the first point at which these densities agree to a satisfactory degree.

The algebra for the 2×2 case immediately works for any $n \times m$ joint distribution of X's and Y's. Analogously define the $n \times n$ transition matrix $A_{X|X}$ whose stationary distribution will be the marginal distribution of X. If either (or both) of X and Y are continuous, then the finite dimensional arguments will not work. However, with suitable assumptions, all of the theory still goes through, so the Gibbs sampler still produces a sample from the marginal distribution of X. The conditional density of X_1 given X_0 is $f_{x_1|x_0}(x_1 \mid x_0) = \int f_{X_1|Y_1}(x_1 \mid y) f_{Y_1|X_0}(y \mid x_0)\, \mathrm{d}y$. Then, step by step, write the conditional densities of $X'_2 \mid X'_0$, $X'_3 \mid X'_0$, $X'_4 \mid X'_0$, ... Similar to the k-step transition matrix $(A_{x|x})^k$, derive an "continuous transition probability matrix" with entries satisfying the relationship

$$f_{X'_k|X'_0}(x \mid x_0) = \int f_{X'_k|X'_{k-1}}(X \mid t) f_{X'_{k-1}|X'_0}(t \mid x_0)\, \mathrm{d}t,$$

Figure 1: Bivariate normal probability density with marginals.

the continuous version of the right matrix multiplication. As $k \to \infty$ it again follows that the stationary distribution is the marginal density of X, the density to which $f_{X'_k|X'_0}$, converges.

Gibbs Sampling from the Bivariate Normal

Recall that the probability density function for the bivariate normal distribution with, for simplicity, means 0 and variance 1 for the variables and correlation ρ between the two variables is

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \mathrm{e}^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}.$$

A standard exercise is to show that the marginal densities are in fact

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$

and

$$f_y(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}.$$

Another standard exercise is to show the conditional probability densities are

$$f_{x|y}(x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \mathrm{e}^{-(x-y)^2/(2(1-\rho^2))}$$

and

$$f_{y|x}(y) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \mathrm{e}^{-(y-x)^2/(2(1-\rho^2))}.$$

That is, $X_i \mid X_j \sim N(x_j, (1-\rho^2))$. Figure **??** illustrates the bivariate normal density with $\rho = 0.8$. The scaling of the z-axis is approximately equal to the x and y axes, so that the density appears less "hill-like" than is often illustrated. The marginal densities are projected onto coordinate planes at $x = 2$ and $y = 2$.

Because the marginals are known, using the Gibbs sampler is not necessary to simulate them. However, Gibbs sampling using this elementary

example is illustrative as in Figure **??**. Apply the algorithm for 1000 steps and to allow for convergence to the stationary distribution, the first 500 steps are discarded, using only the last 500 steps. The first two subgraphs show the frequencies of the sampled marginal distributions, along with the theoretical densities in red. The correspondence is appears close. Better than a visual comparison to the densities is to use a Q-Q plot comparing the sample quantiles to the quantiles of the standard normal distribution. The straight line correspondence of the quantiles over several standard deviations demonstrates the excellent match of simulation to the theoretical density. The third row shows the scatter of the bivariate samples, with the characteristic elliptical distribution The second figure in the third row shows the autocorrelation of the sample points by connecting successive points with segments. The last two plots illustrate the "white noise"-like aspect of the marginal distributions, as expected. Altogether, these reinforce the intuition that the Gibbs sample has produced a satisfactory sample of the marginal distributions.

A Gibbs sampler for a Spam Filter

Consider a Bayesian Inference model for an email spam filter. The goal is to estimate the prevalence ψ of email spam without having to directly count the instances of email spam. The model has the following random variables and events:

- $[S = 1]$, the event that the email is in fact spam, a positive outcome occurring with probability $\mathbb{P}[S = 1] = \psi$,
- $[S = 0]$, the event that the email is *not* spam, a negative outcome with $\mathbb{P}[S = 0] = 1 - \psi$,
- $[R = 1]$, the event that the email is correctly marked spam, a *true positive*
- $[R = 0]$, the event that the email is *incorrectly* marked spam, a *false negative*.

Then define the following parameters:

- *sensitivity* $\eta = \mathbb{P}[R = 1 \mid S = 1] = 0.90$ (so the false positive rate is $\alpha = 0.10$),

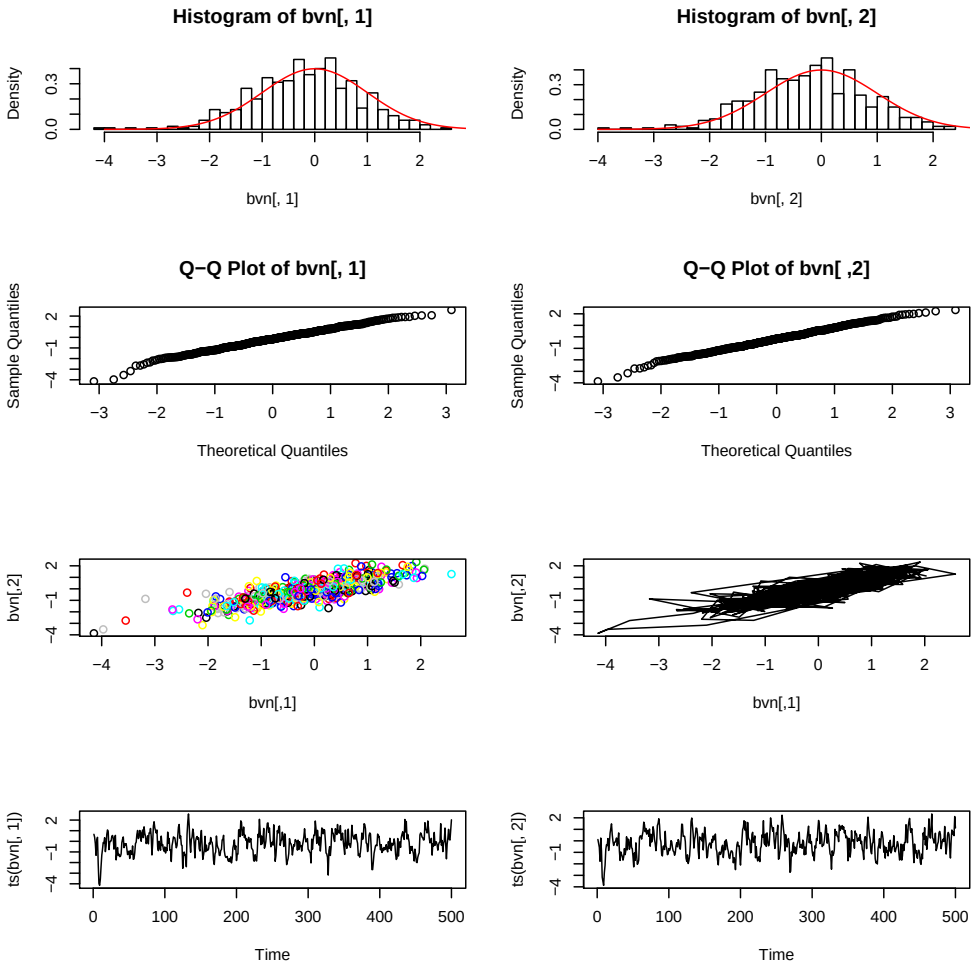


Figure 2: Outcomes from Gibbs sampling of a bivariate normal distribution, illustrating the quality of the simulation.

- *specificity* $\theta = \mathbb{P}[R = 0 \mid S = 0] = 0.95$ (so the false negative rate is $\beta = 0.05$),
- *prevalence* $\psi = \mathbb{P}[S = 1]$.

Then

$$\begin{aligned}\tau = \mathbb{P}[R = 1] &= \mathbb{P}[R = 1 \mid S = 1]\,\mathbb{P}[S = 1] + \mathbb{P}[R = 1 \mid S = 0]\,\mathbb{P}[S = 0] \\ &= \psi\eta + (1 - \psi)(1 - \theta).\end{aligned}$$

Rewrite the equation to solve for the prevalence ψ, usually the parameter of interest:

$$\psi = \frac{\tau + \theta - 1}{\eta + \theta - 1}.$$

Assume as an example, that the filter marks $r = 233$ emails as spam out of a total of $n = 1000$ emails. Then the rate of rejection $\tau = r/n = 0.233$. Likewise the estimate of prevalence is $\psi = 0.21529$. Since this is a Bernoulli random variable, the traditional frequentist confidence interval is $\tau \pm 1.96\sqrt{\tau(1-\tau)/n}$ which becomes $[0.20680, 0.25920]$. Using these in the equation for the prevalence ψ given an interval $\psi \in [0.184, 0.246]$.

Unfortunately, if n is relatively small, this can give nonsensical estimates. For example, $\eta = 0.99$, $\theta = 0.97$, $r = 5$ from $n = 250$ gives $\tau = 0.02$ but $\psi = -0.01042$.

Before starting the Gibbs sampling, first use simple Bayesian inference starting from Bayes' Formula

$$\text{Posterior} \propto \text{ Prior } \cdot \text{ Likelihood}$$

or in probability notation

$$\mathbb{P}[\tau \mid r] \propto \mathbb{P}[\tau] \cdot \mathbb{P}[r \mid \tau].$$

Start with a non-informative prior where τ is uniformly distributed on $(0, 1)$, but use a Beta density

$$\mathbb{P}[\tau = x] = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

with initial shape parameters $\alpha = 1$ and $\beta = 1$ so the density is uniform. Recall that $\mathbb{P}[r \mid \tau]$ is binomial, so

$$\begin{aligned}\mathbb{P}[\tau \mid r] &\propto \tau^{a_0-1}(1 - \tau)^{\beta_0-1} \times \tau^r(1 - \tau)^{n-r} \\ &\propto \tau^{\alpha_0+r-1}(1 - \tau)^{\beta_0+n-r-1}\end{aligned}$$

The right side is essentially the form of the Beta distribution $\text{Beta}(\alpha_n, \beta_n)$. This functional correspondence between the Beta families for the prior and the posterior and the Binomials for the likelihood makes them a **conjugate pair**.. The mean of the Beta distribution with shape parameters α and β is $\alpha/(\alpha+\beta)$ and the variance is $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$.

For the spam classification problem

$$\alpha_n = \alpha_0 + r = 1 + 233 = 234$$
$$\beta_n = \beta_0 + n - r = 1 + 1000 - 233 = 768.$$

This allows using the resulting Beta density posterior for $p(\tau \mid r)$ to derive the mean and confidence interval for τ. Using the facts about the mean and variance above, the posterior density mean is $\alpha/(\alpha+\beta) = 0.23353$. The 95% confidence interval using the Beta cdf $\text{Beta}(p, \alpha, \beta)$ defined by $\text{Beta}(p_-, \alpha_0, \beta_0) = 0.025$ and $\text{Beta}(p_+, \alpha_0, \beta_0) = 0.975$ is (0.207860.26021). This is approximately the same as the frequentist confidence interval.

Next use a Gibbs sampler to get point and interval estimates for ψ. Unlike the estimate for τ where the count r of the spam-marked emails is available, a Gibbs sampler is useful because it is not feasible to directly count the spam emails. Again using Bayes Formula

$$p = \mathbb{P}[S = 1 \mid R = 1] = \mathbb{P}[S = 1, R = 1]/\mathbb{P}[R = 1] = \mathbb{P}[R = 1 \mid S = 1]\,\mathbb{P}[S = 1]/\mathbb{P}[R = 1] = \psi\eta/\tau$$

and

$$1 - p = \mathbb{P}[S = 1 \mid R = 0] = \psi(1-\eta)/(1-\tau).$$

Simulate the *latent counts*

$$X \mid (r, \psi) \sim \text{Binom}(r, p)$$
$$Y \mid (r, \psi) \sim \text{Binom}(n - r, 1 - p)$$
$$\psi \mid (X, Y) \sim \text{Beta}(\alpha_n, \beta_n)$$

where $V + X + Y$, $\alpha_n = \alpha_0 + V$ and $\beta_n = n - V$. Use one value of ψ to find the next in the typical Gibbs iteration. This is in effect a Markov Chain for which the limiting distribution is the posterior of ψ. At each iteration, use the prior distribution and the data. The continued and recursive use of this information will cause the convergence of the simulated values toward the appropriate posterior distribution. Figure **??** illustrates the results of this Gibbs Sampler method.

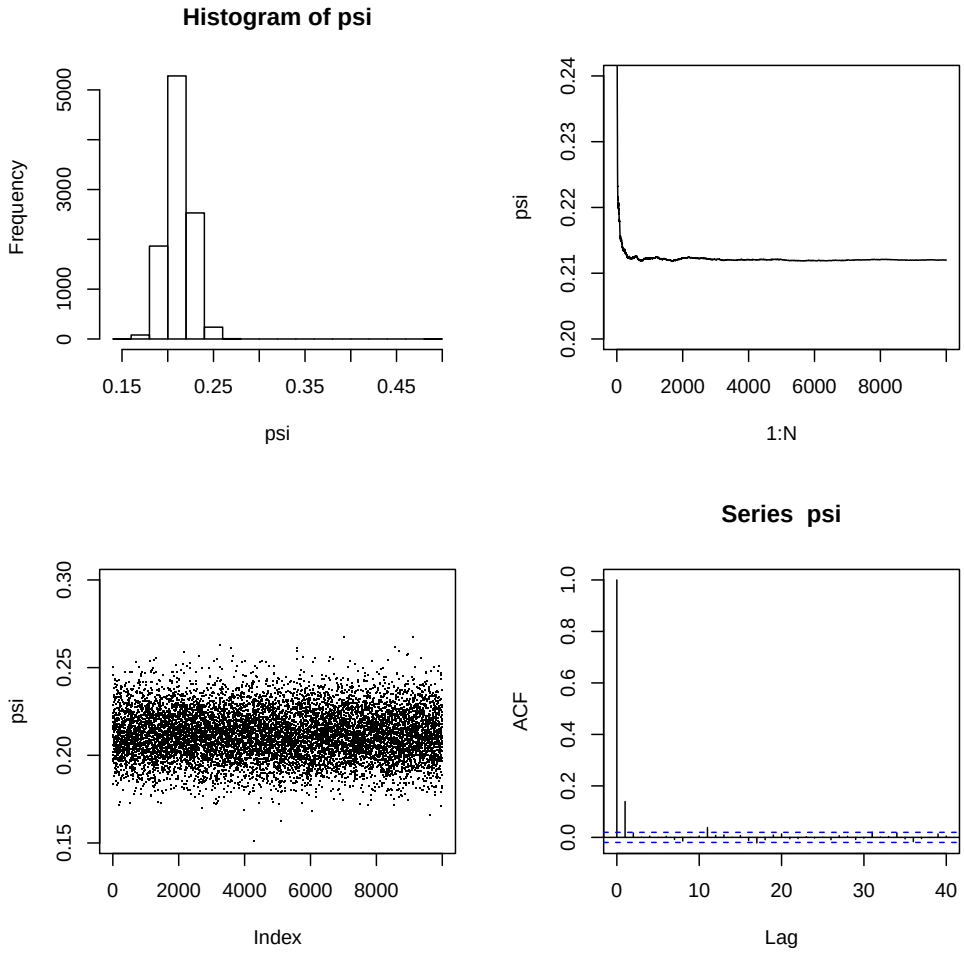


Figure 3: Results of the Gibbs sampler applied to the spam filter.

Gibbs Sampling in Statistics

Suppose $f(x_1, x_2, \ldots, x_N)$ is a probability distribution in which the variables represent parameters of a statistical model. The goal is to get point and interval estimates for these parameters. To fit this into the Gibbs sampling framework, assume that all the single-variable conditional probability densities

$$f(x_i \mid x_j, j \neq i)$$

are available, that is, are a type for which samples can be obtained using standard algorithms. Examples of available distributions include the uniform, the normal, the gamma, the Poisson, and any finite distribution. To generate a sequence of samples, select $\overrightarrow{x}^0 = (x_1^0, x_2^0, \ldots, x_N^0)$ arbitrarily and then create $\overrightarrow{x}^1 = (x_1^1, x_2^1, \ldots, x_N^1)$ as follows:

1. Generate a sample x_1^1 from $f(x_1 \mid x_2^0, x_3^0, \ldots x_N^0)$.
2. Generate a sample x_2^1 from $f(x_2 \mid x_1^1, x_3^0, x_4^0 \ldots x_N^0)$.
3. Generate a sample x_3^1 from $f(x_3 \mid x_1^1, x_2^1, x_4^0, \ldots x_N^0)$.
4. ...

N. Generate a sample x_N^1 from $f(x_N \mid x_1^1, x_2^1, \ldots x_{N-1}^1)$.

One cycle, similar to a raster scan of an image, produces a new value $\overrightarrow{x}^1$. Repeating this process M times produces

$$\overrightarrow{x}^0, \overrightarrow{x}^1, \overrightarrow{x}^2, \ldots, \overrightarrow{x}^M$$

which approximates a sample from the probability distribution $f(x_1, x_2, \ldots, x_N)$.

Using this sample, almost any property of the probability distribution can be investigated. For example, focusing on only the fist component of each $\overrightarrow{x}^k$ produces a sample

$$x_1^0, x_1^1, x_1^2, \ldots, x_1^M$$

from the marginal probability distribution of the first component, formally given by the integral

$$f(x_1) = \int_{x_2} \cdots \int_{x_N} f(x_1, x_2, \ldots, x_N) \, \mathrm{d}x_N \ldots \mathrm{d}x_2 \, .$$

In this way, Gibbs sampling can be thought of as a multi-dimensional numerical integration algorithm. The expected value of the first component x_1,

$$\mathbb{E}[x_1] = int_{x_1} x_1 f(x_1)\ \mathrm{d}x_1$$

is estimated by the average of the sample $x_1^0, x_1^1, x_1^2, \ldots, x_1^M$. A 95% confidence interval for x_1 can be taken directly from the sample.

Gibbs Sampling for Normal Parameters

As another example of the Gibbs sampler, suppose that $X \sim N(\mu, 1/\tau)$ with μ and τ unknown. Based on a reasonably sized sample, the goal is to get the posterior distributions of μ and τ using the Gibbs sampler. Here μ is the population mean, and τ, called the population precision, is the reciprocal of the variance, n is the sample size, $\bar{x}$ is the sample mean and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample variance. Then make a sequence of iterations $i = 1, \ldots N$; with a sample $\mu^{(i)}$ from $f(\mu \mid \tau^{(i-1)}, \text{data})$ (see below for the definition) and sample $\tau^{(i)}$ from $f(\tau \mid \mu^{(i-1)}, \text{data})$ (see below for the definition). Then the theory behind Gibbs sampling ensures that after a sufficiently large number of iterations, T, the set $\left\{(\mu^{(i)}, \tau^{(i)}) : i = T+1, \ldots, N\right\}$ can be seen as a random sample from the joint posterior distribution. The priors are that $f(\mu, \tau) = f(\mu) \times f(\tau)$ with $f(\mu) \propto 1$ and $f(\tau) \propto 1/\tau$. With these definitions and assumptions standard theory shows the sample mean from a normal distribution given the population precision has a conditional posterior distribution

$$(\mu \mid \tau^{(i-1)}, \text{data}) \sim N\left(\bar{x}, \frac{1}{n\tau}\right).$$

Also from standard theory, the precision, (the reciprocal of the variance) given the mean has a conditional posterior distribution

$$(\tau \mid \mu, \text{data}) \sim \text{Gamma}\left(\frac{n}{2}, \frac{2}{(n-1)s^2 + n(\mu - \bar{x})^2}\right).$$

The derivation goes like this:

$$\text{Posterior} \propto \ \text{Prior} \ \times \ \text{Likelihood}$$

or in probability notation

$$p(\tau \mid \mu, \text{data}) \propto p(\tau) \times p(\mu, \text{data} \mid \tau).$$

Then using $\frac{1}{\sigma^n} = \tau^{n/2}$

$$\begin{aligned}
p(\tau \mid \mu, \text{data}) &\propto \tau^{-1} \times \tau^{n/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \\
&\propto \tau^{-1} \times \tau^{n/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu + \mu - \bar{x})^2\right) \\
&\propto \tau^{-1} \times \tau^{n/2} \exp\left(-\frac{\tau}{2}\left[\sum_{i=1}^{n}(x_i - \mu)^2\right.\right. \\
&\qquad \left.\left. +2\sum_{i=1}^{n}(x_i - \mu)(\mu - \bar{x}) + \sum_{i=1}^{n}(\mu - \bar{x})^2\right]\right) \\
&\propto \tau^{-1} \times \tau^{n/2} \exp\left(-\frac{\tau}{2}\left[\sum_{i=1}^{n}(x_i - \mu)^2 + \sum_{i=1}^{n}(\mu - \bar{x})^2\right]\right) \\
&\propto \tau^{n/2-1} \times \exp\left(-\tau\left[\frac{2}{n-1}s^2 + n(\mu - \bar{x})^2\right]\right)
\end{aligned}$$

(Double check last step, may have some fractions wrong.) The last expression is now recognizable as a Gamma distribution with shape parameter $\frac{n}{2}$ and rate (inverse scale parameter) $\frac{2}{n-1}s^2 + n(\mu - \bar{x})^2$. Having this distribution is convenient for implementing the Gibbs sampler. (As a side remark, in many author's derivations for Bayesian statistics, the prior for τ is taken to be Gamma(α, β) which leads to a similar looking Gamma posterior including the additional parameters α and β, but for Gibbs sampling, this is unnecessary because the theory shows that the sampling will converge even with the simpler choice.) (Remark: make sure this is true.)

The R script below implements the Gibbs sampler for these parameters, with $n = 30$, a typical value, and postulating $\bar{x} = 15$, and $s^2 = 3$. The number of iterations is $N = 11{,}000$ with a transient period of 1000 iterations. Figure **??** shows the results of the of Gibbs sampler for the parameters of a normal distribution.

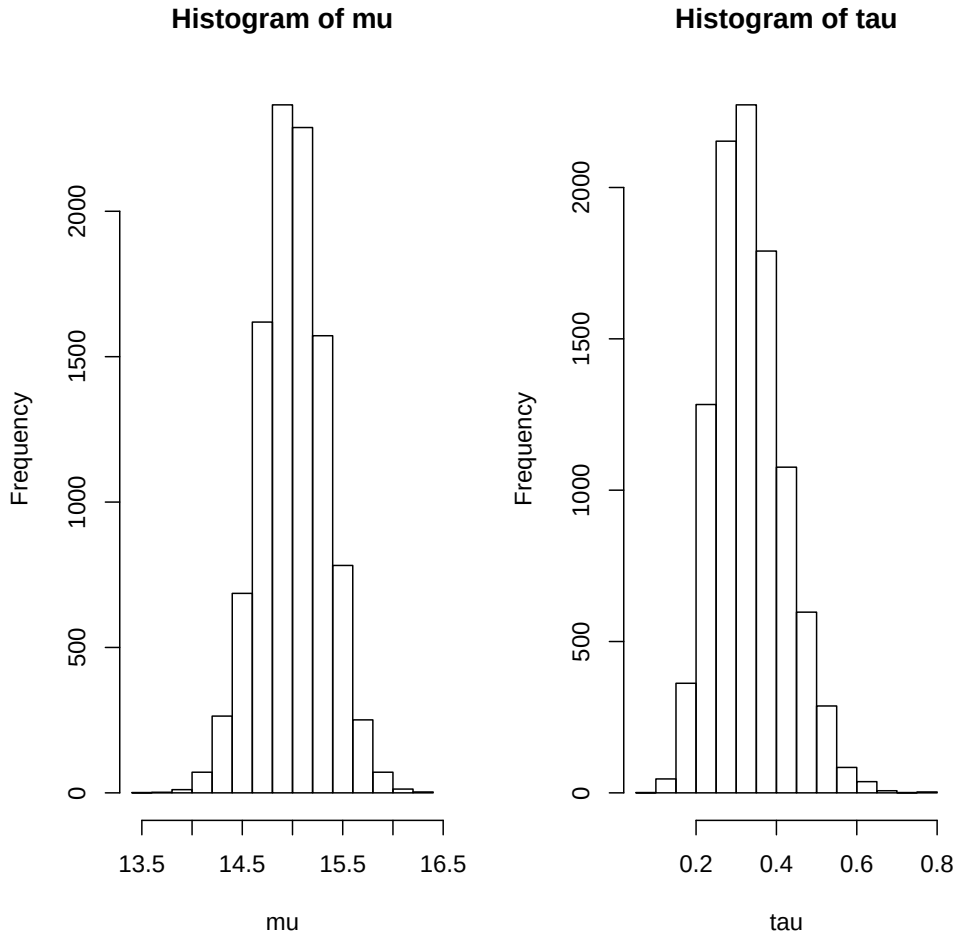


Figure 4: Results of the Gibbs sampler for the parameters of a normal distribution.

Bayesian Hierarchical Models

Apply this formalism to a **Bayesian hierarchical model**. Hierarchical Bayesian models naturally describe the connections among data, observed parameters and other unobserved parameters, sometimes called **latent variables**. A simple three-level hierarchical model uses Bayes' rule to bind together data, X, a parameter to be estimated, λ, and an additional hyper-parameter, β. Both λ and β can be vectors. These are connected in the following way:

1. At the first level, X is described by its likelihood function $f(X \mid \lambda)$, the probability of observing X conditioned on λ.

2. At the next level, λ is modeled by a probability density function, $g(\lambda \mid \beta)$, conditioned on the parameter β.

3. At the third level, the hyper-parameter β is modeled with another density function $h(\beta)$. The choice of $h(\beta)$ reflects the modeler's prior beliefs about the likely values of β.

The three density functions are combined with Bayes' rule, producing a probability density function for λ and β conditioned on the data X

$$F(\lambda, \beta|X) \propto f(X \mid \lambda)g(\lambda \mid \beta)h(\beta).$$

The constant of proportionality is the reciprocal of

$$\int_{\lambda} int_{\beta} f(X \mid \lambda)g(\lambda \mid \beta)h(\beta) \, \mathrm{d}\beta \, \mathrm{d}\lambda$$

which is independent of the parameters λ and β, though dependent on the data X. The integrals, (or sums, in the case of discrete distributions) are over all values λ and β. In most cases, the integral or sum is impossible to evaluate. However, as before, the Metropolis-Hastings algorithm avoids this expression.

As a specific example, consider a model of water pump failure rates. The data, X, are given by pairs (s_i, t_i) for $i = 1, 2, \ldots, 10$. Each pair represents failure information for an individual pump. For each pump, assume the number of failures s_i in time t_i is given by a Poisson distribution with parameter $\lambda_i t_i$, that is

$$f_i(s_i \mid \lambda_i) = \frac{(\lambda_i t_i)^{s_i} \mathrm{e}^{-\lambda_i t_i}}{s_i!}, \quad i = 1, 2, \ldots, 10.$$

Assuming the failures occur independently, the product gives the likelihood function for $\overrightarrow{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{10})$:

$$f(X \mid \overrightarrow{\lambda}) = \prod_{i=1}^{10} \frac{(\lambda_i t_i)^{s_i} \mathrm{e}^{-\lambda_i t_i}}{s_i!}.$$

The traditional frequentist approach is to use $\bar{\lambda_i} = S_i/t_i$ as the point estimate of λ_i for $i = 1, 2, \ldots, 10$. The Bayesian approach is to assume that the individual λ_i's are linked together by a common distribution. A natural choice is a gamma distribution with parameters α and β, so that the density for the ith parameter is

$$g_i(\lambda_i \mid \alpha, \beta) = \prod_{i=1}^{10} \frac{\lambda_i^{\alpha-1} \mathrm{e}^{-\lambda_i/\beta}}{\beta^{\alpha}\Gamma(\alpha)}.$$

The remaining hyper-parameter β is described by an inverse gamma distribution with parameters γ and δ, so that

$$h(\beta) = \frac{\delta^{\gamma} \mathrm{e}^{-\delta/\beta}}{\beta^{\gamma+1}\Gamma(\gamma)}.$$

The parameters γ and δ are selected to make the top-level inverse gamma reasonably *diffuse*. A diffuse distribution tries to convey as little prior information as possible about the parameters. As an extreme case of a non-informative distribution is the uniform distribution on the parameter space.

The resulting posterior joint density for the parameters $\lambda_1, \lambda_2, \ldots, \lambda_{10}$ along with the scale parameter β is

$$F(\lambda_1, \lambda_2, \ldots, \lambda_{10}, \beta \mid X) \propto \left[\prod_{i=1}^{10} \frac{(\lambda_i t_i)^{s_i} \mathrm{e}^{-\lambda_i t_i}}{s_i!}\right] \left[\prod_{i=1}^{10} \frac{\lambda_i^{\alpha-1} \mathrm{e}^{-\lambda_i/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\right] \left[\frac{\delta^{\gamma} \mathrm{e}^{-\delta/\beta}}{\beta^{\gamma+1}\Gamma(\gamma)}\right].$$

For $i = 1, 2, \ldots, 10$, the density for λ_i conditioned on the other parameters is proportional to

$$\lambda_i^{s_i+\alpha-1} \mathrm{e}^{-\lambda_i(t_i+1/\beta)}.$$

The constant of proportionality is obtained by absorbing all factors independent of λ_i. The form of the density for λ_i shows that $\mathbb{P}\left[\lambda_i \mid \lambda_j, j \neq i, X, \beta\right]$ is

a gamma distribution with parameters $s_i + \alpha - 1$ and $1/(t_i + 1/\beta)$. Since the gamma distribution is available, Gibbs sampling can be applied at this step. The density for β, conditioned on the other parameters, is proportional to

$$\frac{\mathrm{e}^{(\sum_{i=1}^{10} \lambda_i + \delta)\beta}}{\beta^{10\alpha+\gamma+1}}$$

showing that $\mathbb{P}[\beta \mid \lambda_1, \lambda_2, \ldots, \lambda_{10}, X]$ is an inverse gamma distribution with parameters $g + 10\alpha$ and $\sum_{i=1}^{10} \lambda_i + \delta$. This too is an available distribution.

This model is an example of a *conjugate hierarchical model*, that is, one whose intermediate distributions, in this case, those for the λ_i and β are similar to the original distributions in the hierarchical model. This fits with the Gibbs sampling requirement that these distributions be available.

Gibbs Sampling for Digital Images

A simple model of a digital image consists of pixel elements arranged on a rectangular lattice with N sites. Each pixel takes a value from a set $S = \{1, 2, \ldots, K\}$ of levels, such as grayscale or color levels. An image is a configuration $\omega \in \Omega$ with an assignment of a level to each of the N sites. Even modestly sized images result in immensely large configuration spaces, for a 100×100 binary image, $|\Omega| = 2^{10000}$.

Consider a model for image degradation with additive noise, modeled by N independent identically distributed random variables $\mathcal{N} = \{\eta_1, \eta_2, \ldots \eta_N\}$. Specifically, take noise with the η_i normally distributed with mean 0 and variance σ^2, that is $\eta_i \sim N(0, \sigma^2)$. Letting ω^{blurred} indicate the degraded or blurred image, $\omega^{\text{blurred}} = \omega + \mathcal{N}$. Since the values of ω^{blurred} are real numbers, the resulting image is determined by rounding each value to the nearest value in S.

The relationship between the original image and the degraded version is probabilistic, given any image ω, there is some probability a particular ω^{blurred} is the degraded version of ω. Image reconstruction looks at the problem the other way around; given ω^{blurred}, there is some probability ω is the original image. This leads to an application of Bayes' Rule. The *posterior distribution* for ω conditioned on ω^{blurred} is

$$\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right] = \frac{\mathbb{P}\left[\omega^{\text{blurred}} \mid \omega\right] \mathbb{P}[\omega]}{\mathbb{P}\left[\omega^{\text{blurred}}\right]}.$$

The goal is to find the configuration maximizing $\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right]$, called the **maximum a posteriori estimate**. The technique is to formulate a new version of the Metropolis algorithm with Gibbs sampling.

By the Law of Total Probability the denominator is

$$\mathbb{P}\left[\omega^{\text{blurred}}\right] = \int\limits_{\omega \in \Omega} \mathbb{P}\left[\omega^{\text{blurred}} \mid \omega\right] \mathbb{P}\left[\omega\right] \, \mathrm{d}\omega \, .$$

This integral (or sum) is over all $\omega \in \Omega$ and does not depend on ω. This is reminiscent of the partition function and recalling the Metropolis algorithm we ignore it.

The likelihood function $\mathbb{P}\left[\omega^{\text{blurred}} \mid \omega\right]$ is

$$\mathbb{P}\left[\omega^{\text{blurred}} \mid \omega\right] \propto \prod_{i=1}^{N} \mathrm{e}^{-\frac{(\omega_i^{\text{blurred}} - \omega_i)^2}{2\sigma^2}} = \mathrm{e}^{-\frac{1}{2\sigma^2} \sum\limits_{i=1}^{N} (\omega_i^{\text{blurred}} - \omega_i)^2}$$

where any constant of proportionality will be absorbed into the denominator above.

An image has patterns, i.e. contiguous regions of similar pixel values. This is reminiscent of the Ising model, where after magnetization the lattice has high-degree long-range correlation between pixels, that is, image-like features. On the other hand, if neighboring values are uncorrelated, the result is the visual equivalent of noise, reminiscent of the Ising model at high temperatures.

This suggests using the Boltzmann probability distribution with the Ising energy function as the prior distribution on images, that is

$$\mathbb{P}\left[\omega\right] \propto \mathrm{e}^{-E_{\text{ising}}(\omega)/kT}$$

where

$$E_{\text{ising}}(\omega) = -J \sum_{i=1}^{N} \sum_{k=1}^{4} \omega_{i,j} \omega_{\langle i,j \rangle [k]}$$

is the nearest neighbor affinity. To retain the idea of correlated pixel values, let $kT/J = 1 < T_c$, the critical temperature below which a phase transition occurs.

Putting all the parts together, the posterior distribution, $\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right]$ is

$$\begin{aligned}\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right] &\propto \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\omega_i^{\text{blurred}}-\omega_i)^2}\mathrm{e}^{-E_{\text{ising}}} \\ &\propto \mathrm{e}^{-\left[\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\omega_i^{\text{blurred}}-\omega_i)^2+E_{\text{ising}}\right]}\end{aligned}$$

Viewing this from a statistical mechanics perspective leads to an analog of an energy function

$$\begin{aligned}E_{\text{image}}(\omega \mid \omega^{\text{blurred}}) &= \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\omega_i^{\text{blurred}}-\omega_i)^2 + E_{\text{ising}} \\ &= \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\omega_i^{\text{blurred}}+\omega_i)^2 - \sum_{\langle i,j\rangle}\omega_i\omega_j\end{aligned}$$

where $\langle i,j\rangle$ indicates the set of nearest neighbor sites j for site i. (Continue to check signs on this energy function.)

Finding the most probable original image ω given ω^{blurred} is thus equivalent to minimizing $E_{\text{image}}(\omega \mid \omega^{\text{blurred}})$. The first term is a positive potential energy penalty for straying too far from the data ω^{blurred} while the second term represents a negative potential energy reflecting the desire to align neighboring pixel values making them conform to the prior notion of a generic image. The optimal solution balances between these two conflicting constraints.

Creating an energy function to minimize leads to a contrast in methods. The Ising model starts with an objective function and interprets it as a energy function, using this to convert it to a probability from a physical interpretation. The image reconstruction model starts with a probabilistic situation with a Bayesian structure, leading to an energy function.

To implement Gibbs sampling, the probability of ω_i conditioned on all the other sites depends on only the sites in the nearest neighborhood set. Suppressing the dependence on ω^{blurred}, this means

$$\begin{aligned}\mathbb{P}\left[\omega_i \mid \omega_j, j \neq i\right] &= \mathbb{P}\left[\omega_i \mid \omega_j, j \in \langle i,j\rangle\right] \\ &\propto \mathrm{e}^{-E_i(\omega_i \mid \omega_j, j\in\langle i,j\rangle)}\end{aligned}$$

where

$$E_i(\omega_i \mid \omega_j, j \in \langle i,j \rangle) = \frac{1}{2\sigma^2}(\omega_i^{\text{blurred}} - \omega_i)^2 - \sum_{\langle i,j \rangle} \omega_i \omega_j.$$

(Continue to check signs on this energy function.) A probability distribution whose conditional probabilities depend on only the values in a neighborhood system is called a **Gibbs distribution** and is an example of a larger notion called a *Markov random field*.

A standard way to implement Gibbs sampling for images is to use a sequence of raster scans, in order by rows or columns, guaranteeing all sites are visited many times. At a selected site i, select $\omega_i = k$ with probability

$$\mathbb{P}\left[\omega_i = k\right] \propto \mathrm{e}^{-\frac{1}{2\sigma^2}(\omega_i^{\text{blurred}} - k)^2 - \sum\limits_{\langle i,j \rangle} k\omega_j}.$$

Repeating this with many raster scans results in a sequence of images that approximates a sample from the posterior distribution. Note the connection to the previous simple Gibbs sampling algorithm, generating a sample from $f(x)$ by sampling iteratively from the conditional distribution $f(x, \mid y)$ then $f(y \mid x)$, creating

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots, Y_k', X_k'$$

except that now there are many more conditional distributions, one for each pixel.

Gibbs sampling fits into the Hastings generalization of the Metropolis algorithm in the following way: In the second step of the Metropolis-Hastings algorithm, the probabilities α_{ij} are all equal to 1. However, the transitions are no longer time-independent, since each depends on the site choice. As a result, the proof is somewhat more involved than the original proofs of convergence given by Hastings. Gibbs sampling will produce a sequence representing a sample from $\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right]$. The full algorithm also includes a "temperature" parameter T to create a simple form of simulated annealing.

Theorem 1. *Assume*

1. *an image with N pixels,*
2. *T_k is a any decreasing sequence of temperatures such that*
   (a) *$T_k \to 0$ as $k \to \infty$,*
   (b) *$T_k \geq N\Delta / \ln(k)$ for all sufficiently large k and constant Δ.*

Then starting at $\omega^{(0)} = \omega^{blurred}$*, the Gibbs sampling sequence* $\omega^{(k)}$ *for* $k = 0, 1, 2, \ldots$ *converges in distribution to the distribution which is uniform on the minimum vales of* $E_{image}(\omega)$ *and* 0 *otherwise.*

In other words, following a prescribed annealing schedule, Gibbs sampling must, in theory, produce a maximum a posteriori estimate of $\mathbb{P}\left[\omega \mid \omega^{\text{blurred}}\right]$.

Even though this result guarantees convergence to the most likely image, the rate of convergence is slow. Theoretically, for a 100×100 lattice with $N = 10^4$ pixels, using the theorem requires $\mathrm{e}^{20{,}000}$ steps to go from $T = 4$ to $T = 0.5$. In practice, it takes about $300 - 1000$ raster scans to produce acceptable results.

For a two-color ($k = 0$ or 1) image Gibbs sampling with annealing is especially straightforward to implement using the ideas of the section on Gibbs sampling. At the pixel ω_i define

$$E^k = \frac{1}{2\sigma^2}(k - \omega_i^{\text{blurred}})^2 + k \sum_{\langle i,j \rangle} \omega_j.$$

Set $\omega_i = k$ with probability

$$\frac{\mathrm{e}^{-E^k/T}}{\mathrm{e}^{-E^0/kT} + \mathrm{e}^{-E^1/kT}}.$$

(Continue to check signs on this energy function. Note also the multiple uses of the symbol k here along with possible typos.)

Section Ending Answer

Sources

The sections on the general theory of the Gibbs sampler and the bivariate binomial are adapted from the article "Explaining the Gibbs Sampler" by

George Casella and Edward George! [?]. The section on Gibbs Sampling from the Bivariate Normal is adapted from "A simple Gibbs sampler" by Darren Wilkinson [?]. The subsection on A Gibbs Sampler for a Spam Filter is adapted from "Three Simple Applications of Markov Chains and Gibbs Sampling" by Gui Larangeira [?]. The subsection on theory and Bayesian hierarchical models is adapted from "The Evolution of Markov Chain Monte Carlo Methods" by Richey [?].

Algorithms, Scripts, Simulations

Algorithm

Scripts

R R script for Gibbs sampler for marginals from bivariate normal.

```
gibbs<-function (n, rho)
{
        mat <- matrix(ncol = 2, nrow = n)
        x <- 0
        y <- 0
        mat[1, ] <- c(x, y)
        for (i in 2:n) {
                x <- rnorm(1, rho * y, sqrt(1 - rho^2))
                y <- rnorm(1, rho * x, sqrt(1 - rho^2))
                mat[i, ] <- c(x, y)
        }
        mat
}

f  <- function(x) { return( (1/sqrt(2 * pi)) * exp(-x^2/
    2) )}

N  <-  1000
rho  <-  0.8

bvngs <- gibbs(N, rho)
```

```r
bvn <- bvngs[(N/2 + 1):N, ]

par(mfrow=c(4,2))
hist(bvn[,1], freq=FALSE, 40)
curve(f, -4, 4, add=TRUE, col="red")
hist(bvn[,2], freq=FALSE, 40)
curve(f, -4, 4, add=TRUE, col="red")
qqnorm(bvn[ , 1], main="Q-Q Plot of bvn[, 1]")
qqnorm(bvn[ , 2], main="Q-Q Plot of bvn[ ,2]")
plot(bvn, col=1:500)
plot(bvn, type="l")
plot(ts(bvn[ , 1]))
plot(ts(bvn[ , 2]))
par(mfrow=c(1,1))
```

R script for Gibbs sampler for normal parameters.

```r
# summary statistics of sample
n    <- 30
ybar <- 15
s2   <- 3

# sample from the joint posterior (mu, tau | data)
mu       <- rep(NA, 11000)
tau <- rep(NA, 11000)
T        <- 1000    # burnin
tau[1] <-  1
## tau[1]  <- 1  # initialisation
for(i in 2:11000) {
    mu[i]  <- rnorm(n = 1, mean = ybar, sd = sqrt(1 / (n
   * tau[i - 1])))
##    sigmasq[i] <- sigmasq[i-1] * rchisq(n = 1, n-1) / (
   n - 1)
    tau[i] <- rgamma(n = 1, shape = n / 2, scale = 2 / ((
   n - 1) * s2 + n * (mu[i] - ybar)^2))
}
mu  <- mu[-(1:T)]   # remove burnin
tau <- tau[-(1:T)] # remove burnin

par(mfrow=c(1,2))
hist(mu)
hist(tau)
par(mfrow=c(1,1))
```

R ℝ script for Gibbs sampler for estimating spam prevalence.

```
N <- 10000              # Number of iterations
Nb <- 2000; N1 <- N+1   # Burn-in

psi <- numeric(N)
psi[1] <- .5            # Initial value

alpha.0  <- 1.0
beta.0   <- 1.0
eta <- 0.99
theta <- 0.97
r <- 233
n  <- 1000

for(i in 2:N) {       # Gibbs Sampler Loop
tau<-psi[i-1]*eta+(1-psi[i-1])* (1-theta)
X <- rbinom(1, r, psi[i-1]* eta/tau)
Y <- rbinom(1, n-r, psi[i-1]*(1-eta)/(1-tau))
psi[i] <- rbeta(1, alpha.0+X+Y, beta.0+n-X-Y)
}

gspsi  <- mean(psi[Nb:N])

par(mfrow=c(2,2))
hist(psi)
plot(1:N,cumsum(psi)/(1:N),type="l",ylab= "psi", ylim=c
    (0.20,0.24))
plot(psi,type='p',pch='.',ylim=c(0.15,0.30))
acf(psi)
```

Problems to Work for Understanding

1: For the probability density function for the bivariate normal distribution with means 0 and variance 1 for the variables and correlation ρ between the two variables is

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \mathrm{e}^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}$$

show that the x marginal is

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}.$$

Reading Suggestion:

Outside Readings and Links:

1.

2.

3.

4.

I check all the information on each page for correctness and typographical errors. Nevertheless, some errors may occur and I would be grateful if you would alert me to such errors. I make every reasonable effort to present current and

accurate information for public use, however I do not guarantee the accuracy or timeliness of information on this website. Your use of the information from this website is strictly voluntary and at your risk.

I have checked the links to external sites for usefulness. Links to external websites are provided as a convenience. I do not endorse, control, monitor, or guarantee the information contained in any external website. I don't guarantee that the links are active at all times. Use the links here with the same caution as you would all information on the Internet. This website reflects the thoughts, interests and opinions of its author. They do not explicitly represent official positions or policies of my employer.

Information on this website is subject to change without notice.

Steve Dunbar's Home Page, `http://www.math.unl.edu/~sdunbar1`

Email to Steve Dunbar, `sdunbar1 at unl dot edu`

Last modified: Processed from LaTeX source on February 5, 2021