Bayesian Tools for Optimization

Vanja Dukic
David Bortz

Department of Applied Mathematics
University of Colorado, Boulder, CO 80309-0526

Based in part on personal lecture notes, and lecture notes and books by Christian Robert, Emmanuel Lesaffre, Hedibert Lopes and Dani Gamerman

August 12, 2016
SAMSI

Outline

1. Statistical philosophies
   - Bayesian vs. Frequentist Inference
2. Bayesian inference
   - Illustrating Example #1
   - Important Distributions for Bayesians
   - Specific Illustrating Example #2
3. Bayesian Computation
4. Markov Chains
5. Basic MCMC methods

Statistical philosophies

- Two main points of view: Frequentist and Bayesian
- Bayesian philosophy: everything in the world is a measure (uncertain)
- Frequentist philosophy: based on long-run frequency principle
- Bayesian: statements about parameters (eg, most likely values) conditional on the observed data
- Frequentist: statements about parameter estimators, conditional on all possible (unobserved) data samples from the experiment ("repeated sampling")

Parameters

- Bayesian statistics: parameters are considered to be random variables (because they are unknown, not because they are truly random).
- Uncertainty about a parameter == Distribution of a parameter
- Uncertainty shrinks with more data, allows for probabilistic statements about parameter values given evidence thus far (observed data thus far)
- Frequentist statistics: parameters treated as fixed but unknown quantities. No probabilistic statements about the parameter values given evidence thus far.

Estimation

- Bayesians: work with the whole distribution of the parameter given the data. This distribution of the parameter given the data is called the "POSTERIOR DISTRIBUTION".
- Objective: summarize the posterior distribution via its maximum (optimization), expected value (integration), or credible intervals (both)
- Frequentists: want to estimate the parameter by its most likely value (MLE) or some other suitable estimator (sample moments)
- Objective: quantify the variability via the estimator's behavior in repeated samples. Frequentists interpretations depend on the asymptotic theory (in infinitely many repeated samples).

Hypothesis Testing

- Bayesians: compare the probabilities of parameter subspaces corresponding to the hypotheses in question, given the observed data.
- Frequentists: gather evidence against the null hypothesis by considering how extreme the observed sample would be if the null were true, and many repeated samples were generated under that null.
- Frequentists assess this "extremism" by p-values or confidence intervals. P-values measure what fraction of times in repeated sampling they would have observed a dataset that is as distant or even more distant from the null than the one at hand.

Main Ingredients

- Six main ingredients in a Bayesian Model:
    - priors
    - likelihood
    - posterior distribution
    - conditional distributions
    - marginal distributions
    - posterior predictive distribution
- Bayes Theorem

Bayesian Model

Definition

A Bayesian statistical model is made of a parametric statistical model,

$$(\mathcal{X}, f(x|\theta)),$$

and a prior distribution on the parameters,

$$(\Theta, \pi(\theta))$$

.

Ingredients

- Observations $x_1, \ldots, x_n$ are generated from a probability distribution $f_i(x_i|\theta_i, x_1, \ldots, x_{i-1}) = f_i(x_i|\theta_i, x_{1:i-1})$

$$x = (x_1, \ldots, x_n) \sim f(x|\theta), \qquad \theta = (\theta_1, \ldots, \theta_n)$$

  Associated Likelihood

$$\ell(\theta|x) = f(x|\theta)$$

- *Uncertainty* in the parameters θ is modeled through a *probability* distribution π on Θ, called the *prior distribution*
- *Inference* based on the distribution of θ conditional on x, $\pi(\theta|x)$, called *posterior distribution*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

Bayes Theorem

Theorem

Let events $A_1, \dots, A_k$ *form a partition of the space* S *such that* $\Pr(A_j) > 0$ *for all* j *and let* B *be any event such that* $Pr(B) > 0$*. Then for* $i = 1, \dots, k$*:*

$$Pr(A_i|B) = \frac{Pr(A_i)\Pr(B|A_i)}{\sum_k Pr\left(A_k\right)\Pr(B|A_k)}$$

Bayes Theorem

- Bayes' Theorem in continuous setting:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

- The prior and the likelihood are combined to derive the posterior distribution.
- The denominator of the RHS is known as the marginal distribution of the data. This often gets discarded, as it is known (or assumed) to be a fixed constant which is irrelevant in optimization and simulation.
  - NB: Can do this only if the integral is finite (ie the likelihood is integrable with respect to the prior)!
  - NB2: In general, for purposes other than maximization and credible intervals, the constant can't be ignored (eg. model comparison, expected values)

Example

- Billiard ball W is rolled on a line of length one, with uniform probability of stopping anywhere: W stops at location p.
- Second ball O the rolled n times under the same assumptions.
- Let X denote the number of times the ball O stopped on the left of W
- Given X, what inference can we make on p?

Example

$$p \sim \mathcal{U}([0,1]) \quad \text{and} \quad X \sim \mathcal{B}(n,p)$$

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$P(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp$$

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp$$

Example

then

$$P(a < p < b|X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp}$$
$$= \frac{\int_a^b p^x (1-p)^{n-x} dp}{B(x+1, n-x+1)},$$

where $B(x+1, n-x+1) = \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n)}$. Thus,

$$p|x \sim \mathcal{B}e(x+1, n-x+1)$$

Main Distributions

- Given $f(x|\theta)$ and $\pi(\theta)$, several distributions of interest are:
- the *joint distribution* of $(0, x)$,

$$\varphi(\theta, x) = f(x|\theta)\pi(\theta);$$

- the *marginal distribution* of x,

$$\begin{aligned} m(x) &= \int \varphi(\theta, x) d\theta \\ &= \int f(x|\theta)\pi(\theta) d\theta; \end{aligned}$$

Main Distribution

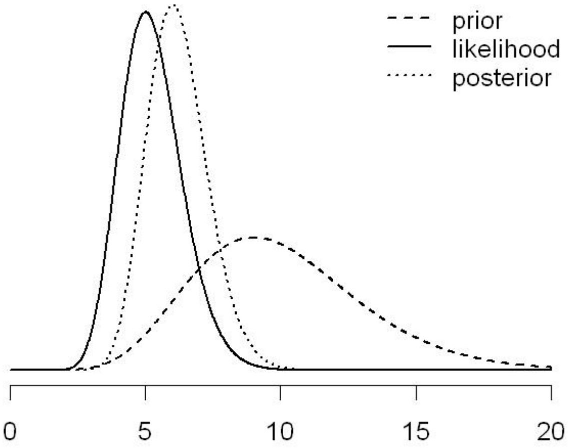
- the *posterior distribution* of θ,

$$\begin{aligned}\pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d]\theta} \\ &= \frac{f(x|\theta)\pi(\theta)}{m(x)};\end{aligned}$$

- the *predictive distribution* of y, when $y \sim g(y|x)$,

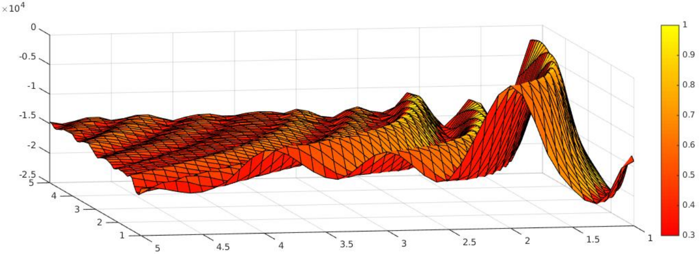
$$g(y|x) = \int g(y|\theta, x)\pi(\theta|x)d\theta.$$

Bayes Theorem



Bayesian theorem: an illustration

Multi Modal Likelihood Example



Example of a complex likelihood function

Example

Consider $x|\theta \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(0, 10)$.

$$\begin{aligned}\pi(\theta|x) \propto f(x|\theta)\pi(\theta) &\propto \exp\left(-\frac{(x-\theta)^2}{2} - \frac{\theta^2}{20}\right) \\ &\propto \exp\left(-\frac{11\theta^2}{20} + \theta x\right) \\ &\propto \exp\left(-\frac{11}{20}\{\theta - (10x/11)\}^2\right)\end{aligned}$$

and

$$\theta|x \sim \mathcal{N}\left(\frac{10}{11}x, \frac{10}{11}\right)$$

HPD (highest posterior density)

- Natural confidence region

$$
\begin{aligned}
C &= \{\theta; \pi(\theta|x) > k\} \\
&= \left\{\theta; \left|\theta - \frac{10}{11}x\right| > k'\right\}
\end{aligned}
$$

- Highest posterior density (HPD) region

Bayesian Inference

- All inference is based on the posterior distribution
- Summaries of the posterior:
    - Many estimators can be defined
- Most popular estimator is the posterior mean
    - Many good properties from the decision theoretic point of view
- Posterior mode
    - Similar to the MLE, many useful optimization tools
- Posterior median
    - Good summary of the most frequent a posteriori value
    - Often used for non-symmetric posterior distributions
- Credible intervals and regions
    - Regions of the parameter space with the desired probability mass
    - Credible intervals vs confidence intervals: probability that the parameter lives in this interval rather than the probability that the CI covers the true value
    - Central and HPD intervals

Quantities of Interest

- Posterior modes: $\max_\theta \pi(\theta)$;
- Posterior moments: $E_\pi[g(\theta)]$;
- Density estimation: $\hat{\pi}(g(\theta))$;
- Bayes factors: $f(x|M_0)/f(x|M_1)$
- Decision: $\max_d \int U(d,\theta)\pi(\theta)d\theta$

How do we find priors?

- informative/subjective:
  - summaries of available knowledge
  - expert opinion
- ''non-informative''
- flat (over finite space)
- improper
- vague and conjugate
- objective
- the higher the parameter is in the modeling hierarchy, the harder it is to come up with priors

Prior elicitation

- **There exist no "non-informative" priors.** Each prior contains some amount of information about a parameter. So the term "non-informative" is misleading.
- Nowadays there is less emphasis on the actual choice of the non-informative prior. It is ok when it represents weak prior information and is equivalent to negligible "extra data".
- Usually, in inference, we also examine "sensitivity" of model results to priors under different prior assumptions.

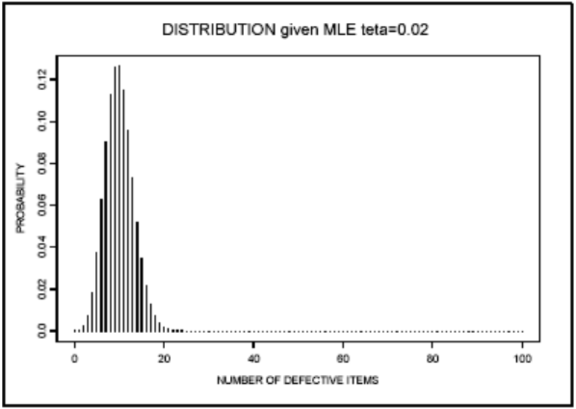
Example

- A start up company (A) sells goods to company B
- According to the internal procedure, around 5% of the goods may be defective.
- Company A has sent in the past two large lots of B, but with about 10% defective parts.
- Company B gives company A a last chance and orders a lot of 50. Now, 1 of these items is defective.
- *What is the best guess of the proportion of defective items for another lot of 500?*

Example

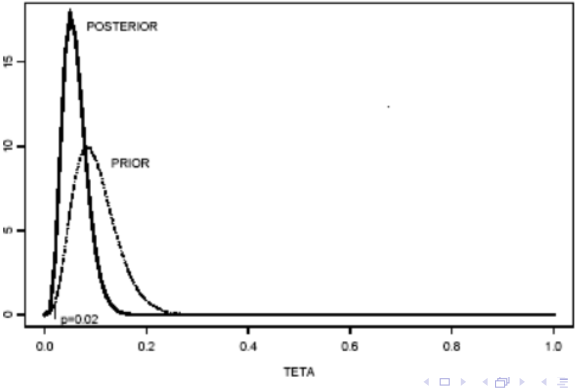
- **Classical prediction** is based on $\hat{\theta} = \frac{1}{50} = 0.02$

$\Rightarrow$ **estimate in the new lot of size 500 = 10 deffective items**



Example

- **Bayesian prediction**
- **Prior** distribution of company B, before the lot of 50 arrived
  - quite concentrated around 0.10, since they were large lots
  - not sure "how" concentrated. Take Beta(p, q) with $p = 5$, $q = 45$
- **Sample** of $N = 50$ with $x = 1$ defective
- **Posterior** $p(\theta|\mathbf{x}) = \text{Beta}(p + x, q + N - x) = \text{Beta}(6, 94)$



Example

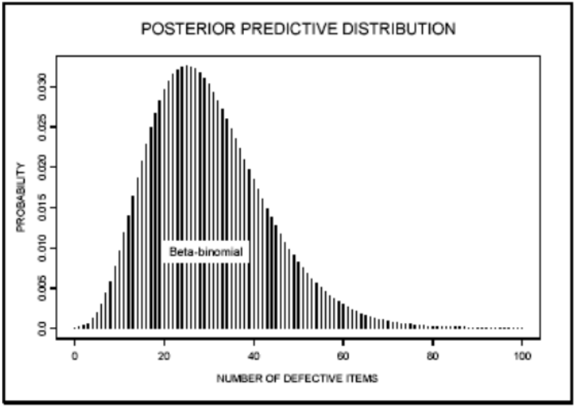
- **Predictive distribution** of # defective parts in new sample of 500:
    - # of defective parts in new sample of size M: z

$$f(z|x) = \int_0^1 \binom{M}{z} \theta^z (1-\theta)^{(M-z)} \frac{\theta^{x+p-1}(1-\theta)^{(N-x+q-1)}}{B(x+p, N-x+q)} d\theta$$

$$= \binom{M}{z} \frac{B(z+x+p, M-z+N-x+q)}{B(x+p, N-x+q)}$$

Example

Posterior mean $= 30!$



Overview of Computational algorithms

1. All algorithms can be classified based on their purpose
   - optimization/maximization (solving equations $d/dx = 0$)
   - marginalization (integration)
   - simulation (generating a random sample from the posterior)
2. ... or, based on whether a stream of random numbers is used
   - Monte Carlo (MC)
   - non-Monte Carlo (non-MC) algorithms
3. ... or, based on whether data augmentation (DA) is performed (augmentation and non-augmentation algorithms)
   - Augmenting data with additional data or parameter values is a basis for EM, Data Augmentation and Gibbs sampling algorithms. We augment the observed data with "stuff" (actual missing data values, parameter values, or sufficient statistics values) to simplify the computation and sampling. The idea is to replace one complicated move with an iterative series of simple moves.

Optimization algorithms

Example Optimization algorithms (iterative):

- Methods using Hessians
  - Newton methods
    - Sequential quadratic programming (small-medium scale constrained problems)
    - Levenberg-Marquardt (NLS)
- Methods using gradients
  - Quasi-Newton methods (medium-large problems, e.g. N<1000)
  - Conjugate gradient methods (large problems)
  - Interior point methods (constrained optimization)
  - Gradient descent aka steepest descent (recent results for big data).
  - Stochastic optimization (uses random gradient approximation)
- Methods only using functions
  - Interpolation methods
  - Pattern search methods (Nelder-Mead, Hooke-Jeeves, etc.)
  - Grid search
    - brute force algorithm
    - “works” only in low dimensions
    - starting values for more sophisticated algorithms

Marginalization/integration

- Analytically
- Trapezoid rule
- Quadrature methods
- Laplace's method
- Monte Carlo integration
- Importance Sampling
- Acceptance/rejection
- Data Augmentation
- MCMC

Simulation algorithms

Generates a random sample from a distribution:

- Acceptance/rejection
- Data augmentation
- Gibbs sampler
- Metropolis-Hastings
- Other MCMC

This sample of random draws from a given distribution can then be used for optimization, obtaining marginals (integration), for their moments, credible intervals, etc.

Random vs. deterministic

- Random = requiring stream of pseudo-random numbers (seed-dependent):
  - Markov chain Expectation-Maximization (MCEM)
  - Data Augmentation (DA)
  - Metropolis-Hastings (MH)
  - Markov chain Monte Carlo (MCMC)
  - Acceptance/rejection (AR)
  - Importance sampling (IS)
- Non-random = no random number generator involved
  - Expectation-Maximization (EM)
  - Newton-Raphson (NR)
  - Laplace...

Random vs. Deterministic, cont'd

- For regular, unimodal posteriors or likelihoods, deterministic algorithms work efficiently
- However, complicated posteriors (or likelihoods) - for example distributions with regions of low probability, with many modes (multi-modal) - will generally not be represented well with the output from deterministic algorithms. We need an algorithm that can efficiently explore the interesting regions of posterior or likelihood.

Augmentation-based

- EM, MCEM, DA/Gibbs, other MCMC
- Instead of working directly with the observed likelihood or posterior, one "pads" the observed data with additional stuff. In the EM that stuff is missing data (or varieties of missing data). In DA and Gibbs, we add parameter values (parameters are treated as missing data).
- For example, in MCMC instead of working with $p(\theta_1, \theta_2, \theta_3, \ldots, \theta_k | y)$, Gibbs augments the data with values of parameters $\theta_2, \theta_3, \ldots \theta_k$ (i.e., fixes their values), and works with lower-dimensional $p(\theta_1 | \theta_2, \ldots, \theta_k, y)$, then rotates and iterates.

MCMC

- Provide a *correlated* sample from an approximation to the posterior
- Require Markov chain convergence analysis
- Require dealing with the correlation problem (to obtain an approximately uncorrelated sample)
- Can explore the entire posterior space (in theory). In practice, we don't know what the posterior looks like and therefore can't tell whether the sampler has explore all of it or not. Need techniques to assure good exploration ("mixing").

Markov Chains[1]

- Markov chains
- Stationary distributions and Ergodicity
- Inefficiency factor and effective sample size
- Central limit theorem

[1] Based on Gamerman and Lopes (2007) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman & Hall/CRC.

Markov Chains

- A *Markov chain with the state space S*is a stochastic process where given the present state, past and future states are independent, i.e.

  $$\Pr(\theta^{(n+1)} \in A | \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \dots, \theta^{(0)} \in A_0)$$

  equals

  $$\Pr(\theta^{(n+1)} \in A | \theta^{(n)} = x)$$

  for all sets $A_0, \dots, A_{n-1}, A \subset S$ and $x \in S$.
- This defines the transition function, or kernel $P(x, A)$ of the Markov chain
- When $P(x, A)$ does not depend on n, the chain is *homogeneous*
  1. for all $x \in S$, $P(x, \cdot)$ is a probability distribution over S;
  2. for all $A \subset S$, the function $x \mapsto P(x, A)$ can be evaluated

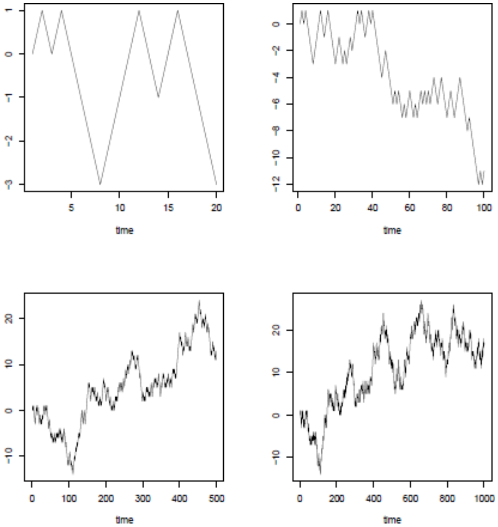
Example: Random walk

- Consider a particle jumping left and right independently, with successive jumps, w_i, governed by a probability function f over the integers and $\theta^{(n)}$ representing its position at time n, $n \in N$.
- Initially, $\theta^{(0)}$ is distributed according to some distribution $\pi^{(0)}$.
- The positions can be related as

$$\theta^{(n)} = \theta^{(n-1)} + w_1 + w_2 + \cdots + w_n$$

  where the w_i are independent random variables with probability function f.
- $\{\theta^{(n)} : n \in N\}$ is a Markov chain in Z.

Example: Random walk



$$\Pr\{\theta^{(n)} = \theta^{(n-1)} + i\} = 1/2, \text{ for } i = -1, 1 \text{ and } \theta^{(0)} = 0.0$$

Discrete state spaces

If S is finite with r elements, $S = \{x_1, x_2, \ldots, x_r\}$, a transition matrix P with (i, j)th element given by $P(x_i, x_j)$ can be defined as

$$P = \begin{pmatrix} P(x_1 x_1) & \cdots & P(x_1, x_r) \\ \vdots & & \vdots \\ P(x_r, x_1) & \cdots & P(x_r, x_r) \end{pmatrix}.$$

Transition probability from state x to state y over m steps, denoted by $P^m(x, y)$, is given by the probability of a chain moving from state x to state y in exactly m steps. It can be obtained for $m \geq 2$ as

$$\begin{aligned} P^m(x, y) &= \Pr(\theta^{(m)} = y | \theta^{(0)} = x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} \Pr(y, x_{m-1}, \ldots, x_1 | x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} \Pr(y, |x) \ldots \Pr(x_1 | x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} P(x_{m-1}, y) \cdots P(x, x_1) \end{aligned}$$

Chapman-Kolmogorov equations

$$
\begin{aligned}
P^{n+m}(x,y) &= \sum_z \Pr(\theta^{(n+m)} = y | \theta^{(n)} = z, \theta^{(0)} = x) \\
&\times \Pr(\theta^{(n)} = z | \theta^{(0)} = x) \\
&= \sum_z P^n(x,z) P^m(z,y)
\end{aligned}
$$

and (more generally)

$$P^{n+m} = P^n P^m.$$

- Marginal distribution of the chain at time n:

$$\pi^{(n)} = (\pi^{(n)}(x_1), \ldots, \pi^{(n)}(x_r))$$

Then,

$$\pi^{(n)}(y) = \sum_{x \in S} P^n(x,y) \pi^{(0)}(x)$$

or, in matrix notation

$$\pi^{(n)} = \pi^{(0)} P^n$$

Stationary distributions

- A fundamental problem for Markov chains is the study of the asymptotic behavior of the chain as the number of iterations $n \to \infty$.
- A key concept is that of a *stationary distribution* π. A distribution π is said to be a stationary distribution of a chain with transition probabilities $P(x, y)$ if

  $$\sum_{x \in S} \pi(x) P(x, y) = \pi(y), \ \forall y \in S$$

  or in matrix notation as $\pi P = \pi$.
- If the marginal distribution at any given time n is π, then the marginal distribution at time $n + 1$ is $\pi P = \pi$.
- Once the chain reaches a stage where π is its distribution, all subsequent distributions are π.
- π is also known as the *equilibrium distribution*.

Ergodicity

- A chain is said to be *geometrically ergodic* if $\exists\lambda \in [0,1)$ and a real, integrable function $M(x)$ such that

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)\lambda^n$$

  for all $x \in S$. If $M(x) = M$, then the *ergodicity is uniform*.
- Uniform ergodicity $\Rightarrow$ geometric ergodicity $\Rightarrow$ ergodicity.
- The smallest λ, satisfying the above condition is called the *rate of convergence*.
- A very large value of $M(x)$ may slow down convergence considerably.

Ergodic theorem

Once ergodicity of the chain is established, important limiting theorems can be stated. The first and most important one is the ergodic theorem.

Theorem

The ergodic average of a real-valued function $t(\theta)$ *is the average* $\bar{t}_n = (1/n)\sum_{i=1}^{n} t(\theta^{(i)})$. *If the chain is ergodic and* $E_\pi[t(\theta)] < \infty$ *for the unique limiting distribution* π *then*

$$\bar{t}_n \xrightarrow{a.s.} E_\pi[t(\theta)] \text{ as } n \to \infty$$

which is a Markov chain equivalent of the law of large numbers.

It states that averages of chain values also provide strongly consistent estimates of parameters of the limiting distribution π despite their dependence across iterations.

Inefficiency factor

Definition

For the chain $t^{(n)} = t(\theta^{(n)})$ we have:
Autocovariance at lag k: $\gamma_k = \text{Cov}_\pi(t^{(n)}, t^{(n+k)})$
Variance of $t^{(n)}$: $\sigma^2 = \gamma_0$
Autocorrelation at lag k: $\rho_k = \gamma_k / \sigma^2$
$\tau_n^2 / n = \text{Var}_\pi(\bar{t}_n)$.

It can be shown that

$$\tau_n^2 = \sigma^2 \left(1 + 2 \sum \frac{n-k}{n} \rho_{k k=1}^{n-1} \right) \rightarrow \tau^2 = \sigma^2 \left(1 + 2 \sum_{k=1}^{\infty} \rho_k \right)$$

as $n \to \infty$.
The term between parentheses in the above equation can be called *inefficiency factor* or *integrated autocorrelation time* because it measures how far $t^{(n)}s$ are from being a random sample and how much $\text{Var}_\pi(\bar{t}_n)$ increases because of that.

Effective sample size

The inefficiency factor can be used to derive the *effective sample size*

$$n_{\text{eff}} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

which can be thought of as the size of a random sample with the same variance since

$$\text{Var}_\pi(\bar{t}_n) = \sigma^2 / n_{\text{eff}}.$$

It is important to distinguish between

$$\sigma^2 = \text{Var}_\pi[t(\theta)] \text{ and } \tau^2$$

σ^2 is the variance of $t(\theta)$ under the limiting distribution of the Markov chain,
τ^2 is the limiting sampling variance of $\bar{t}_n\sqrt{n}$.
Note that under independent sampling they are both given by σ^2.

Central limit theorem

- If a chain is uniformly (geometrically) ergodic

$$\frac{\bar{t}_n - E_\pi[t(\theta)]}{\tau_n/\sqrt{n}} \xrightarrow{d} N(0,1),$$

  as $n \to \infty$.
- The ergodic theorem provides theoretical support for the use of ergodic averages as estimates of limiting quantities.
- The above CLT equation provides support for inference, such as for example evaluation of confidence intervals.

Reversible chains

- Let $(\theta^{(n)})_{n\geq 0}$ be a homogeneous Markov chain with transition probabilities $P(x,y)$ and stationary distribution π.
- The sequence of states $\theta^{(n)}, \theta^{(n-1)}, \ldots$. in reverse order is also a Markov chain, with transition probabilities:

$$\begin{aligned} P_n^*(x,y) &= \Pr(\theta^{(n)} = y | \theta^{(n+1)} = x) \\ &= \frac{\Pr(\theta^{(n+1)} = x | \theta^{(n)} = y)\Pr(\theta^{(n)} = y)}{\Pr(\theta^{(n+1)} = x)} \\ &= \frac{\pi^{(n)}(y)P(y,x)}{\pi^{(n+1)}(x)} \end{aligned}$$

- In general, the reverse chain need not be homogeneous.

Reversible chains

- When $P^*(x, y) = P(y, x)\ \forall x, y \in S$, the Markov chain is *reversible*, and this is equivalent to:

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

- It can be interpreted as saying that the rate at which the system moves from x to y when in equilibrium, $\pi(x)P(x, y)$, is the same as the rate at which it moves from y to x, $\pi(y)P(y, x)$.
- The above equation is referred to as the *detailed balance equation*; *balance* because it equates the rates of moves through states and *detailed* because it does it for every possible pair of states.
- For reversible Markov chains, π is always a steady-state distribution
- Reversible Markov chains are one of the pivotal ideas in designing MCMC algorithms

Markov chain Monte Carlo

Top 10 algorithms with the greatest influence on the development and practice of science and engineering in the 20th century[2]:

1. Metropolis Algorithm for Monte Carlo
2. Simplex Method for Linear Programming
3. Krylov Subspace Iteration Methods
4. The Decompositional Approach to Matrix Computations
5. The Fortran Optimizing Compiler
6. QR Algorithm for Computing Eigenvalues
7. Quicksort Algorithm for Sorting
8. Fast Fourier Transform
9. Integer Relation Detection
10. Fast Multipole Method

[2] In chronological order. Dongarra and Sullivan (2000) Guest Editors' Introduction: The Top 10 Algorithms, Computing in Science and Engineering, 2, 22-23.

Historical background

- Monte Carlo ideas floating around in the 1940's, mostly credited to Stan Ulam, John Von Neumann, and Nick Metropolis
- The main premise is that combinatorially hard questions could be approximately answered by sample statistics (Monte Carlo)
    - Eckhard (1987) Stan Ulam, John Von Neumann and the Monte Carlo method. Los Alamos Science, 15, 131-136.
    - Metropolis and Ulam (1949) The Monte Carlo method. Journal of the American Statistical Association, 44, 335-341;

1970s

- The Metropolis Algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953): Equations of state calculations by fast computing machines. Journal of Chemical Physics, 21, 1087-1091.)
- Hastings and his student Peskun showed that Metropolis and the more general Metropolis-Hastings algorithm are particular instances of a larger family of algorithms.
- Hastings (1970): Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, 97-109.
- Peskun (1973): Optimum Monte-Carlo sampling using Markov chains. Biometrika, 60, 607-612.

1980s

Other seminar papers followed:

- Geman and Geman (1984): Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.
- Tanner and Wong (1987): The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82, 528-550.
- Gelfand and Smith (1990): Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

Practical MCMC

Markov chain: A sequence of random variables $\{X_n : n \in \mathbb{N}\}$ defined on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ which satisfies the property, for any $A \in \mathcal{B}(\mathbb{X})$

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

and we denote

$$P(x, A) = \mathbb{P}(X_n \in A | X_{n-1})$$

Markov chain Monte Carlo: Given a target π, design a transition kernel P such that asymptotically as $n \to \infty$

$$\frac{1}{N} \sum_{n=1}^{N} \varphi(X_n) \to \int \varphi(x) \pi(x) dx$$

- Need to simulate the Markov chain easily even if π is hard to sample from.

Transition kernels for MCMC

- Given $\pi(x)$, there is an infinite number of kernels $P(x, A)$ which have $\pi(x)$ as their invariant distribution.
- The “art” of MCMC consists of coming up with good ones.
- Convergence is ensured under very weak assumptions: namely irreducibility and aperiodicity.
- It is usually very easy to establish that an MCMC sampler converges toward π but very difficult to obtain rates of convergence.

Metropolis-Hastings

- The Metropolis-Hastings algorithm is a generic algorithm designed to sample from the probability distribution $\pi(\theta)$ known up to a normalizing constant.
- Start with a proposal distribution/kernel $q(\theta, \theta')$, a simple approximation to π that is easy to sample from, such that:

$$\int_{\Theta} q(\theta, \theta') d\theta' = 1 \text{ for any } \theta\,.$$

- The basic idea of the MH algorithm is to propose a new candidate θ', drawn from $q(\theta, \theta')$, which is based on the current state of the Markov chain, θ.
- We only accept this new sample with probability $\alpha(\theta, \theta')$, designed so the invariant distribution of the transition kernel is the target distribution $\pi(\theta)$.

MH algorithm

- Initialization:
  - Select deterministically or randomly $\theta^{(0)}$.
- iteration i; $i \geq 1$:
  - Sample $\theta' \sim q(\theta^{(i-1}, \theta')$, and compute

$$\alpha(\theta^{(i-1)}, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta', \theta^{(i-1)})}{\pi(\theta^{(i-1)})q(\theta^{(i-1)}, \theta')}\right) .$$

  - with probability $\alpha(\theta^{(i-1)}, \theta')$, set $\theta^{(i)} = \theta'$; otherwise set $\theta^{(i)} = \theta^{(i-1)}$.

Random Walk Metropolis algorithm

- The original Metropolis algorithm (1953) corresponds to the following choice for $q(\theta, \theta')$

  $$\theta' = \theta + Z \text{ where } Z \sim f;$$

  This is the *random walk proposal*.
- The distribution $f(z)$ is the distribution of the random walk's increments Z and

  $$q(\theta, \theta') = f(\theta' - \theta) \Rightarrow \alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta')f(\theta - \theta')}{\pi(\theta)f(\theta' - \theta)}\right).$$

- If $f(\theta' - \theta) = f(\theta - \theta')$ - e.g., $Z \sim N(0, \Sigma)$ - then

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

General M-H algorithm

- The Hastings' generalization (1970) corresponds to the following choice for $q(\theta, \theta')$

  $$q(\theta, \theta') = q(\theta')\,;$$

  This is the *independence proposal*.
- in this case, the acceptance probability is given by

  $$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta)}{\pi(\theta)q(\theta')}\right) = \min\left(1, \frac{\pi^*(\theta')}{q^*(\theta')}\frac{q^*(\theta)}{\pi^*(\theta)}\right)$$

  where π^* and q^* are unnormalized versions of π and q.
- Note that the ratio $\pi^*(\theta)/q^*(\theta)$ also appears in the Accept/Reject and Importance Sampling algorithms.

Why does MH work?

- To establish that the MH converges towards the required target, we need to show that
  - $\pi(\theta)$ is the invariant distribution of the Markov kernel associated to the MH algorithm.
  - The Markov chain is irreducible; i.e., one can reach any set A such that $\pi(A) > 0$.
  - the Markov chain is aperiodic; i.e., one does not visit in a periodic way the state-space.

Why does MH work?

- The transition kernel associated to the MH algorithm can be rewritten as

$$K(\theta,\theta') = \alpha(\theta,\theta')q(\theta,\theta') + \underbrace{\left(1 - \int \alpha(\theta,u)q(\theta,u)du\right)}_{\text{rejection probability}}\delta_\theta(\theta')$$

Remark: Note that the proper way to write the above is:

$$K(\theta,d\theta') = \alpha(\theta,\theta')q(\theta,\theta')d\theta' + \left(1 - \int \alpha(\theta,u)q(\theta,u)du\right)\delta_\theta(d\theta')$$

- Clearly:

$$\begin{aligned}\int K(\theta,\theta')d\theta' =& \int \alpha(\theta,\theta')q(\theta,\theta')d\theta' \\ &+ \left(1 - \int \alpha(\theta,u)q(\theta,u)du\right)\int \delta_\theta(\theta')d\theta' \\ =&1.\end{aligned}$$

Why does MH work?

- We want to show that

$$\int \pi(\theta)K(\theta,\theta')d\theta = \pi(\theta').$$

- Note that this condition is satisfied if the *reversibility property* is satisfied $\forall(\theta,\ \theta')$:

$$\pi(\theta)K(\theta,\theta') = \pi(\theta')K(\theta',\theta);$$

i.e., the probability of being in A and moving to B is equal to the probability of being in B and moving to A.

Is MH reversible?

- Note that

$$\pi(\theta)K(\theta,\theta') = \pi(\theta)\{\alpha(\theta,\theta')q(\theta,\theta')+\left(1-\int\alpha(\theta,u)q(\theta,u)du\right)\delta_\theta(\theta')\}\,.$$

- Then

$$\begin{aligned}\pi(\theta)\alpha(\theta,\theta')q(\theta,\theta') =&\pi(\theta)\min\left(1,\frac{\pi(\theta')q(\theta',\theta)}{\pi(\theta)q(\theta,\theta')}\right)q(\theta,\theta')\\ =&\min(\pi(\theta)q(\theta,\theta'),\pi(\theta')q(\theta',\theta))\\ =&\pi(\theta')\min\left(1,\frac{\pi(\theta)q(\theta,\theta')}{\pi(\theta')q(\theta',\theta)}\right)q(\theta',\theta)\\ =&\pi(\theta')\alpha(\theta',\theta)q(\theta',\theta)\,.\end{aligned}$$

Is MH reversible?

- Trivially, we also have that

$$\left(1-\int \alpha(\theta,u)q(\theta,u)du\right)\delta_\theta(\theta')\pi(\theta)$$
$$=\left(1-\int \alpha(\theta',u)q(\theta',u)du\right)\delta_{\theta'}(\theta)\pi(\theta')$$

- Thus, it follows that

$$\pi(\theta)K(\theta,\theta')=\pi(\theta')K(\theta',\theta)\,.$$

- Hence, π is the invariant distribution of the transition kernel K.

Example

- Consider the case where

$$\pi(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right) .$$

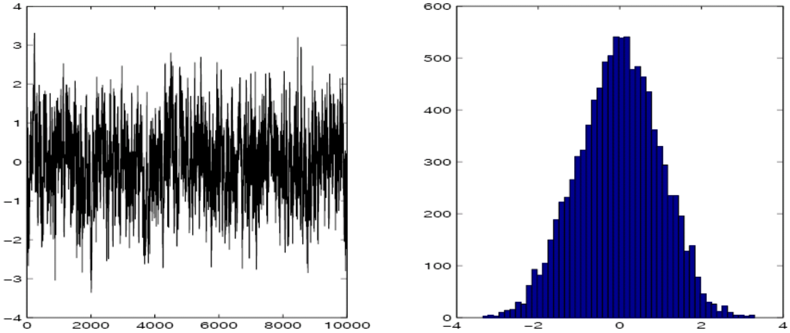
- We implement the MH algorithm for

$$q_1(\theta, \theta') \propto \exp\left(-\frac{(\theta' - \theta)^2}{2(0.2)^2}\right) .$$

- We implement the MH algorithm for

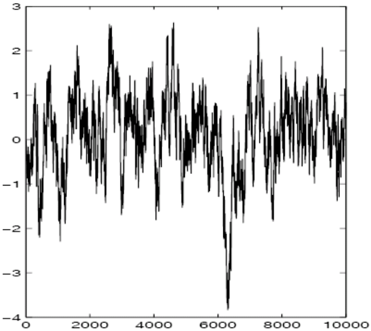
$$q_2(\theta, \theta') \propto \exp\left(-\frac{\left(\theta' - \theta\right)^2}{2(5)^2}\right) .$$

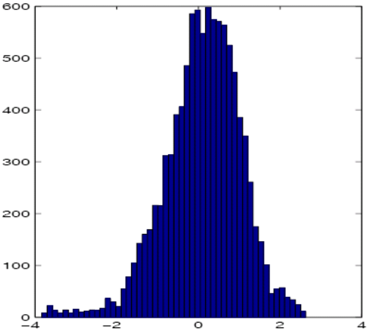
Example



MCMC output for q_1, we estimate $E(\theta) = -0.02$ and $var(\theta) = 0.99$

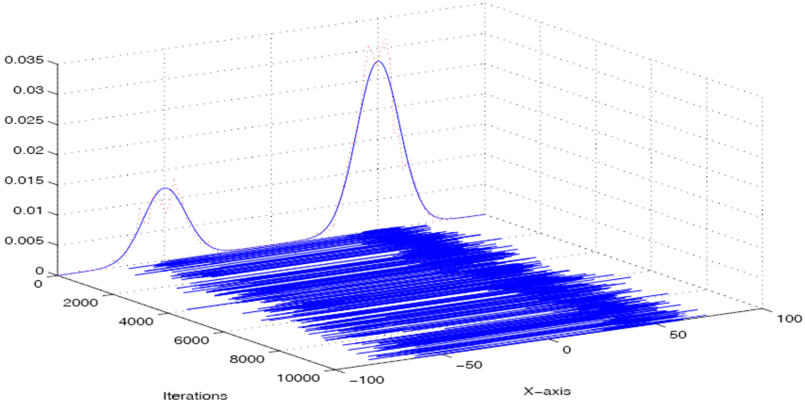
Example





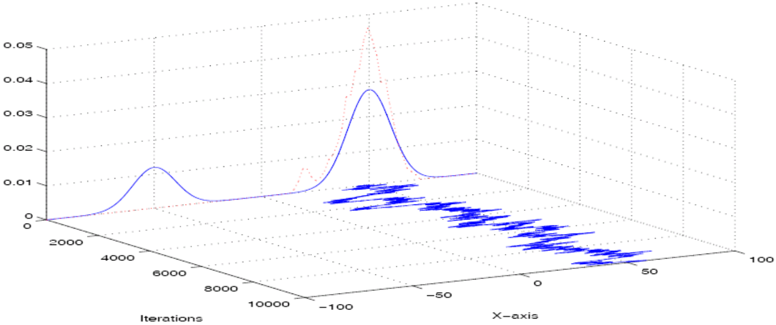
MCMC output for $q_3(\theta,\theta') \propto \exp\left(-\frac{(\theta'-\theta)^2}{2(0.02)^2}\right)$, we estimate $E(\theta) = 0.10$ and $var(\theta) = 0.92$

Example



Exploration of a bimodal distribution using a random walk MH algorithm

Example



Bad exploration of a bimodal distribution using a random walk MH algorithm; the variance of the random walk increments is too small.

Gibbs Sampler

- If $\theta = (\theta_1, \ldots, \theta_p)$ where $p > 2$, the Gibbs sampling strategy is appropriate.
- Initialization:
  - Select deterministically or randomly $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_p^{(0)}\right)$.
- Iteration i; $i \geq 1$:

  For $k = 1 : p$

  - Sample $\theta_k^{(i)} \sim \pi\left(\theta_k | \theta_{-k}^{(i)}\right)$, where

$$\theta_{-k}^{(i)} = \left(\theta_1^{(i)}, \ldots, \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, \ldots, \theta_p^{(i-1)}\right) .$$

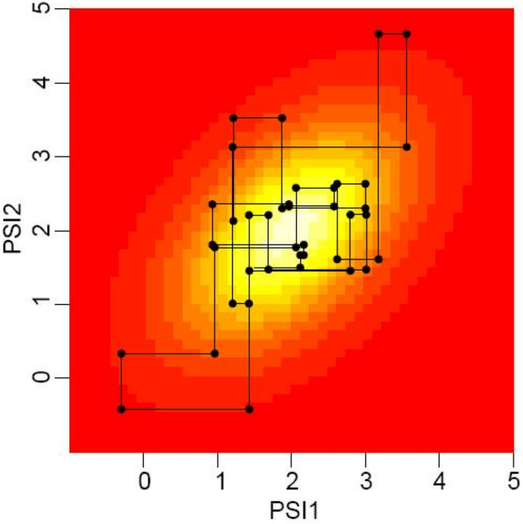
Gibbs Sampler

- The Gibbs sampler requires sampling from the full conditional distributions
$$\pi(\theta_k|\theta_{-k})\,.$$
- For many complex models, it is impossible to sample from several of these "full"conditional distributions.
- In those cases, another nested algorithm (eg, MH) is needed to sample from these lower-dimensional full conditionals.
- In other cases, the algorithm might be very inefficient because a subset of the variables can be very correlated.

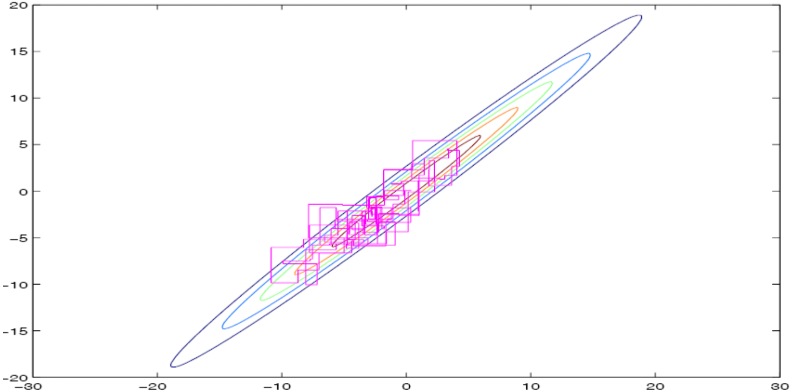
Gibbs Sampler

- Try to have as few “blocks” as possible.
- Put the most correlated variables in the same block.
- If necessary, re-parametrize the model to achieve this.
- Integrate analytically as many variables as possible: pretty algorithms can be much more inefficient than ugly algorithms.
- There is no general result telling strategy A is better than strategy B in all cases, you need experience.

Gibbs Sampler



Slow Gibbs Sampler



Example: Pumps

- Multiple failures in a nuclear plant:

Pump	1	2	3	4	5	6	7	8	9	10
Failures	5	1	5	14	3	19	1	1	4	22
Times	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.4

- Model: Failures of the i-th pump follow a Poisson process with parameter λ_i $(1 \leq i \leq 10)$.
- For a given time t_i, the number of failures up to time t_i is a Poisson $\mathcal{P}(\lambda_i t_i)$ random variable. Denote that random variable as p_i
- The unknowns consist of $\theta := (\lambda_1, \ldots, \lambda_{10}, \beta)$

Example: Pumps

- Hierarchical model

$$\lambda_i \overset{iid}{\sim} \mathcal{G}a(\alpha, \beta) \text{ and } \beta \sim \mathcal{G}a(\gamma, \delta)$$

  with $\alpha = 1.8$ and $\gamma = 0.01$ and $\delta = 1$.

- the posterior distribution is proportional to

$$\prod_{i=1}^{10} \{(\lambda_i t_i)^{p_i} \exp(-\lambda_i t_i) \lambda_i^{\alpha-1} \exp(-\beta\lambda_i)\} \beta^{10\alpha} \beta^{\gamma-1} \exp(-\delta\beta)$$

$$\propto \prod_{i=1}^{10} \{\lambda_i^{p_i+\alpha-1} \exp(-(t_i+\beta)\lambda_i)\} \beta^{10\alpha+\gamma-1} \exp(-\delta\beta)\,.$$

- The multidimensional distribution is rather complex. It is not obvious how the inverse cdf method, the rejection method or importance sampling could be used in this context.

Example: Pumps

- The conditionals have a familiar form

$$\lambda_i|(\beta, t_i, p_i) \sim \mathcal{G}a(p_i + \alpha, t_i + \beta) \text{ for } 1 \leq i \leq 10\,,$$

$$\beta|(\lambda_1, \ldots, \lambda_{10}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i)\,.$$

- Instead of directly sampling the vector $\theta = (\lambda_1, \ldots, \lambda_{10}, \beta)$ at once, one could suggest sampling it iteratively, starting for example with the $\lambda_i's$ for a given guess of β, followed by an update of β given the new samples $\lambda_1, \ldots, \lambda_{10}$.

Examples: Pumps

- Given a sample, at iteration t, $\theta^t := (\lambda_1^t, \ldots, \lambda_{10}^t, \beta^t)$ one could proceed as follows at iteration $t+1$,

1. $\lambda_i^{t+1} | (\beta^t, t_i, p_i) \sim \mathcal{G}a(p_i + \alpha, +\beta^t)$ for $1 \leq i \leq 10$,
2. $\beta^{t+1} | (\lambda_1^{t+1}, \ldots, \lambda_{10}^{t+1}) \sim \mathcal{G}a(\lambda + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i^{t+1})$.

- Instead of directly sampling in a space with 11 dimensions, one samples in spaces of dimension 1

Computer Session:

- Gibbs

Gibbs sampler

A sequence $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots\}$ is drawn from a Markov chain whose *limiting equilibrium distribution* is the posterior distribution, $\pi(\theta)$, and whose transition kernel is the product of the full conditional distributions:

Algorithm

1. $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$
2. $\theta^{(j)}$ is sampled as follows:

$$
\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1 | \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}) \\
\theta_2^{(j)} &\sim \pi(\theta_2^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}) \\
&\vdots \\
\theta_p^{(j)} &\sim \pi(\theta_p | \theta_1^{(j)}, \ldots, \theta_{p-1}^{(j)})
\end{aligned}
$$

Example 0. Simple linear regression

For $i = 1, \ldots, n$, y_i is linearly related to x_i, i.e.,

$$y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

with y's independent conditionally on $\theta = (\alpha, \beta, \sigma^2)$.
Prior distribution:

$$\begin{aligned} \alpha &\sim N(\alpha_0, \tau_\alpha^2) \\ \beta &\sim N(\beta_0, \tau_\beta^2) \\ \sigma^2 &\sim IG(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

No analytical solution to $E(g(\theta)|x, y)$

- Easy to derive full conditionals ⇒ Gibbs sampler!

Full conditionals

- $[\alpha] \sim N(m, C)$

$$\begin{aligned} m =& C(\tau_\alpha^{-2}\alpha_0 + \sigma^{-2}\sum_{i=1}^{n}(y_i - \beta x_i)) \\ C^{-1} =& \tau_\alpha^{-2} + \sigma^{-2}n \end{aligned}$$

- $[\beta] \sim N(m, C)$

$$\begin{aligned} m =& C(\tau_\beta^{-2}\beta_0 + \sigma^{-2}\sum_{i=1}^{n}(y_i - \alpha x_i)) \\ C^{-1} =& \tau_\beta^{-2} + \sigma^{-2}n \end{aligned}$$

- $[\sigma^2] \sim IG(\nu_1/2, \nu_1\sigma_2^2/2)$

$$\begin{aligned} \nu_1 &= \nu_0 + n \\ \nu_1\sigma_1^2 &= \nu_0\sigma_0^2 + \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \end{aligned}$$

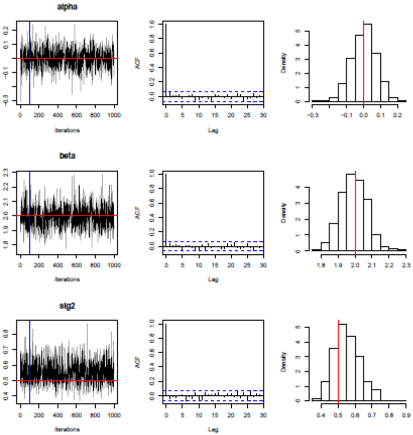
R Code (written by: Hedibert Lopes)

```
# Simulating the data
set.seed(1244)
n=100;alpha=0;beta=2;sig2=0.5;true=c(alpha,beta,sig2)
x=rnorm(n)
y=rnorm(n,alpha+beta*x,sqrt(sig2))
# Prior hyperparameters
alpha0=0;tau2a=10;beta0=0;tau2b=10;nu0=3;s02=1;nu0s02=nu0*s02
# Setting up starting values
alpha=0;beta=0;sig2=1
# Gibbs sampler
M = 1000
draws = matrix(0,M,3)
for (i in 1:M){
  var = 1/(1/tau2a+n/sig2)
  mean = var*(sum(y-beta*x)/sig2+alpha0/tau2a)
  alpha = rnorm(1,mean,sqrt(var))
  var = 1/(1/tau2b+sum(x^2)/sig2)
  mean = var*(sum((y-alpha)*x)/sig2+beta0/tau2b)
  beta = rnorm(1,mean,sqrt(var))
  sig2 = 1/rgamma(1,(nu0+n)/2,(nu0s02+sum((y-alpha-beta*x)^2)/2))
  draws[i,] = c(alpha,beta,sig2)
}
# Markov chains + marginal posterior
names = c("alpha","beta","sig2")
ind = 101:M
par(mfrow=c(3,3))
for (i in 1:3){
  ts.plot(draws[,i],xlab="iterations",ylab="",main=names[i])
  abline(v=ind[1],col=4)
  abline(h=true[i],col=2,lwd=2)
  acf(draws[ind,i],main="")
  hist(draws[ind,i],prob=T,main="",xlab="")
  abline(v=true[i],col=2,lwd=2)
}
```

R Code (written by: David Bortz)

See the separate file “mcmcSAMSI.m”

Output



Simulation: $n = 100$, $\alpha = 0$, $\beta = 2$, $\sigma^2 = 0.5$. Initial values $\alpha^{(0)} = 0$, $\beta^{(0)} = 0$ and $\sigma^{2(0)} = 1$. 1000 draws, 100 burn in.