Bayesian Techniques for Parameter Estimation

“He has Van Gogh’s ear for music,” Billy Wilder

Statistical Inference

Goal: The goal in statistical inference is to make conclusions about a phenomenon based on observed data.

Frequentist: Observations made in the past are analyzed with a specified model. Result is regarded as confidence about state of real world.

- Probabilities defined as frequencies with which an event occurs if experiment is repeated several times.
- Parameter Estimation:
    - Relies on estimators derived from different data sets and a specific sampling distribution.
    - Parameters may be unknown but are fixed

Bayesian: Interpretation of probability is subjective and can be updated with new data.

- Parameter Estimation: Parameters described as density

Bayesian Inference

Framework:

- Prior Distribution: Quantifies prior knowledge of parameter values.
- Likelihood: Probability of observing a data if we have a certain set of parameter values.
- Posterior Distribution: Conditional probability distribution of unknown parameters given observed data.

Joint PDF: Quantifies all combination of data and observations

$$p(\theta, y) = p(y|\theta)\pi_0(\theta)$$

Bayes' Relation: Specifies posterior in terms of likelihood, prior, and normalization constant

$$\pi(\theta|y) = \frac{p(y|\theta)\pi_0(\theta)}{p_Y(y)} = \frac{p(y|\theta)\pi_0(\theta)}{\int_{\mathbb{R}^p} p(y|\theta)\pi_0(\theta)d\theta}$$

Problem: Evaluation of normalization constant typically requires high dimensional integration.

Bayesian Inference

Uninformative Prior: No *a priori* information parameters

e.g., $\pi_0(\theta) = 1$ with limits

Informative Prior: Use conjugate priors; prior and posterior from same distribution

Evaluation Strategies: $\int_{\mathbb{R}^p} p(y|\theta)\pi_0(\theta)d\theta$

- Analytic integration --- Rare
- Classical quadrature; e.g., p = 2
- Monte Carlo quadrature Techniques
- Markov Chains

Bayesian Inference

Example: Y_i: Result from i^{th} coin toss

$$Y_i = \begin{cases} 0 & \text{tails} \\ 1 & \text{heads} \end{cases}$$

θ: Probability of getting heads

Consider probability of observing series of tosses $y = [y_1, y_2, \cdots, y_N]$ conditioned on probability θ:

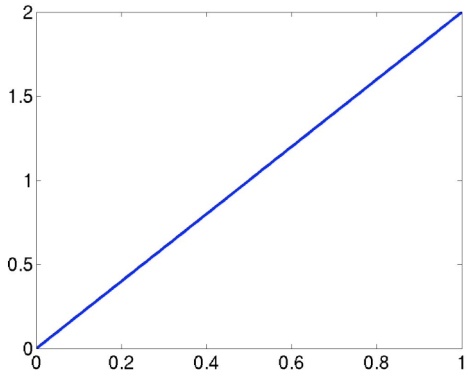
$$\begin{aligned} P(y_1, \cdots, y_N|\theta) &= \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum y_i}(1-\theta)^{1-\sum y_i} \qquad N_1\text{: Number of heads} \\ &= \theta^{N_1}(1-\theta)^{N_0} \qquad N_0\text{: Number of tails} \end{aligned}$$

Uninformative prior yields

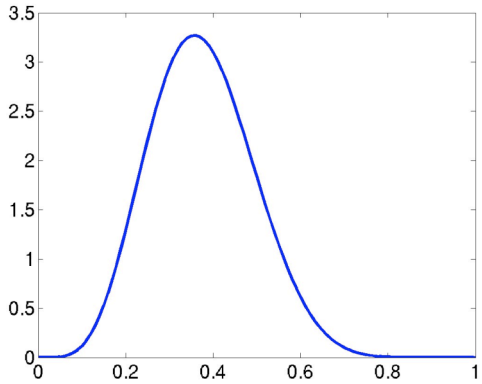
$$\pi(\theta|y) = \frac{\theta^{N_1}(1-\theta)^{N_0}}{\int_0^1 \theta^{N_1}(1-\theta)^{N_0} d\theta} = \frac{(N+1)!}{N_0!N_1!}\theta^{N_1}(1-\theta)^{N_0}$$

Bayesian Inference

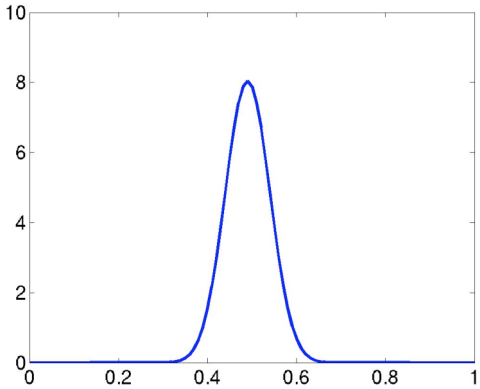
Example:



1 Head, 0 Tails



5 Heads, 9 Tails



49 Heads, 51 Tails

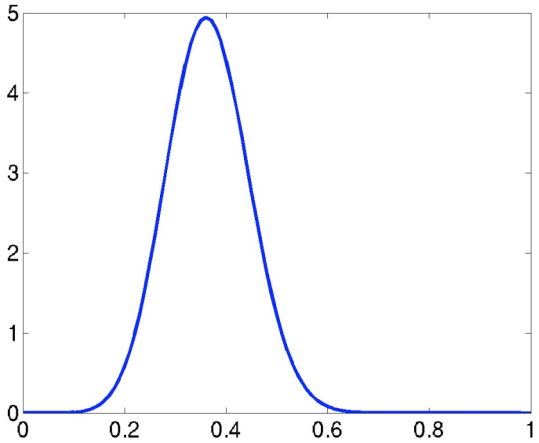
Note: For $N = 1$, frequentist theory would give probability 1 or 0

Bayesian Inference

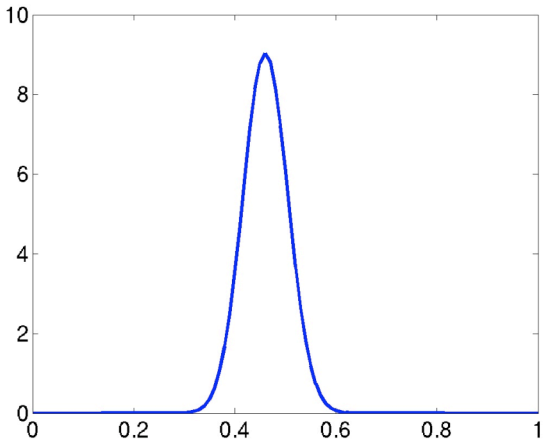
Example: Now consider

$$\pi_0(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\theta-\mu)^2/2\sigma^2}$$

with $\mu = .3$ and $\sigma = .1$.



5 Heads, 5 Tails



50 Heads, 50 Tails

Note: Poor informative prior incorrectly influences results for a long time.

Parameter Estimation Problem

Likelihood: Suppose that ε_j has a parameter-dependent density $p_\theta(\varepsilon) = p(\varepsilon;\theta)$. The associated likelihood is

$$p(Y|\theta) = \prod_{j=1}^{n} p(Y_j - y(t_j;q);\theta)$$

where θ is the parameter for ε_j.

Assumption: Suppose $\varepsilon_j \sim N(0,\sigma_0^2)$ so that $Y_j \sim N(y(t_j;q_0),\sigma_0^2)$. Then

$$p(Y|q) = \frac{1}{(2\pi)^{n/2}\sqrt{|V|}} e^{-\frac{1}{2}(Y-y(t;q))^T V^{-1}(Y-y(t;q))}$$

For iid errors,

$$p(Y_j|q) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(Y_j-y(t_j;q))^2/2\sigma^2}$$

$$\Rightarrow p(Y|q) = \prod_{j=1}^{n} p(Y_j|q) = \frac{1}{(2\pi\sigma)^{n/2}} e^{-\frac{1}{2}SS_q/\sigma^2}$$

where

$$SS_q = \sum_{j=1}^{n} (Y_j - y(t_j;q))^2$$

Note:

- Solutions q_{MLE}, σ^2_{MLE} that maximize equation are maximum likelihood estimators for q_0, σ_0^2.
- Equivalent to OLS when distribution for ε_j known.

Parameter Estimation: Example

Example: Consider the spring model

$$m\ddot{y} + c\dot{y} + ky = 0$$

$$y(0) = 2\ ,\ \dot{y}(0) = 0$$

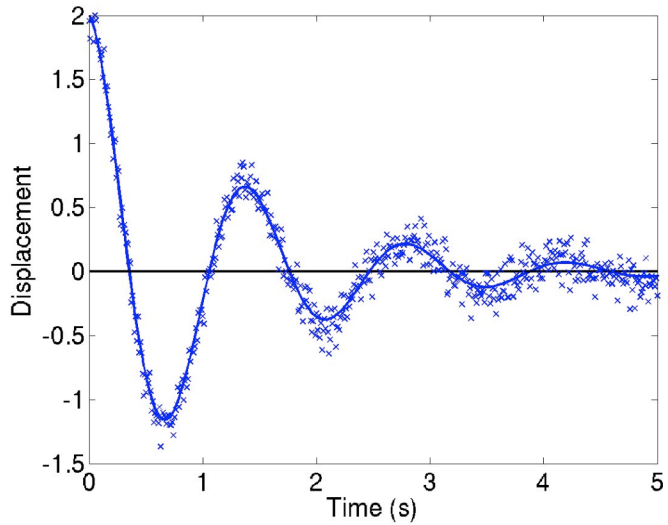
Note: $m = 2, k = 4.1, c = q_0 = 0.3$

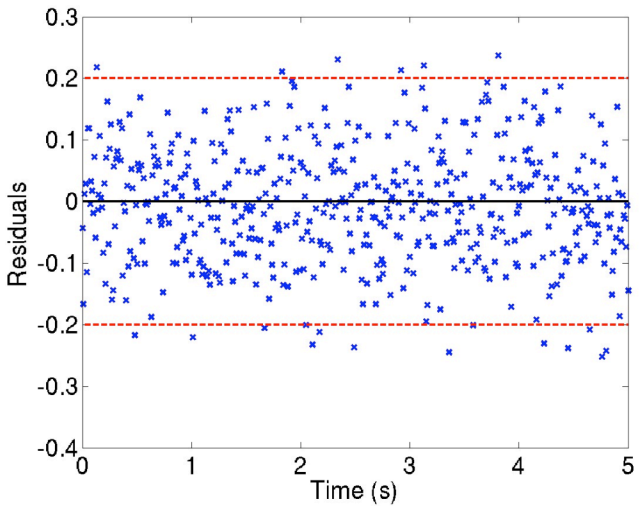
$n = 501$

which has the solution

$$y(t) = 2e^{-ct/2m} \cos\left(\frac{\sqrt{4mk - c^2}}{2m}\, t\right)$$

when $c^2 - 4km < 0$. Assume that m, k known and take $q = c$. We also assume that $\varepsilon_j \sim N(0, \sigma_0^2)$ where $\sigma_0 = 0.1$.





Parameter Estimation: Example

Ordinary Least Squares: Here

$$\chi(q) = \chi(c) = \begin{bmatrix} \frac{\partial y}{\partial c}(t_1; c) \\ \vdots \\ \frac{\partial y}{\partial c}(t_n; c) \end{bmatrix}$$

where

$$\frac{\partial y}{\partial c} = -\frac{t}{m} e^{-ct/2m} \cos\left(\frac{\sqrt{4mk - c^2}}{2m} t\right) + \frac{ct}{m\sqrt{4mk - c^2}} \sin\left(\frac{\sqrt{4mk - c^2}}{2m} t\right)$$

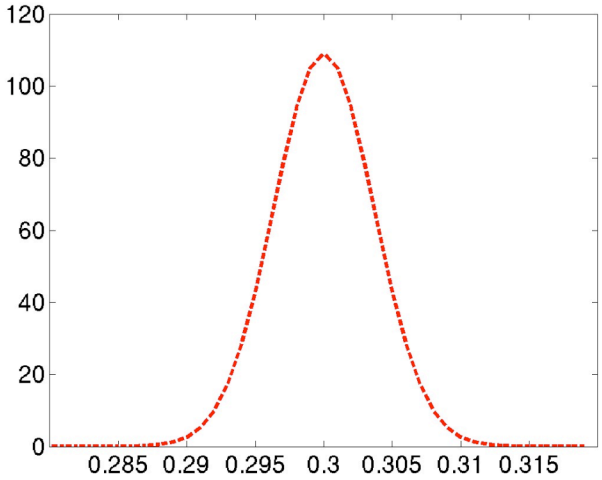
Then

$$V = \sigma_c^2 = \sigma_0^2 \left[\chi^T(c)\chi(c)\right]^{-1} = 1.3382 \times 10^{-5}$$

$$\Rightarrow \sigma_c = 0.0037$$

Recall:

$$\varepsilon_j \sim N\left(0, \sigma_0^2\right) \Rightarrow C \sim N(c_0, \sigma_0^2[\chi^T(c_0)\chi(c_0)]^{-1})$$



Sample Distribution

Parameter Estimation: Example

Bayesian Inference: The likelihood is

$$P(Y|c) = \frac{1}{(2\pi\sigma_0)^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n}(Y_j - y(t_j;q))^2/\sigma_0^2}$$

and we use the uninformed prior

$$\pi(\theta) = \pi(c) = 1 \cdot \chi_{(0,\infty)}$$

Note: Use the midpoint rule to approximate

$$\int_{\mathbb{R}} p(y|c)\pi(c)dc = \int_0^{\infty} p(y|c)dc$$

Posterior Distribution:

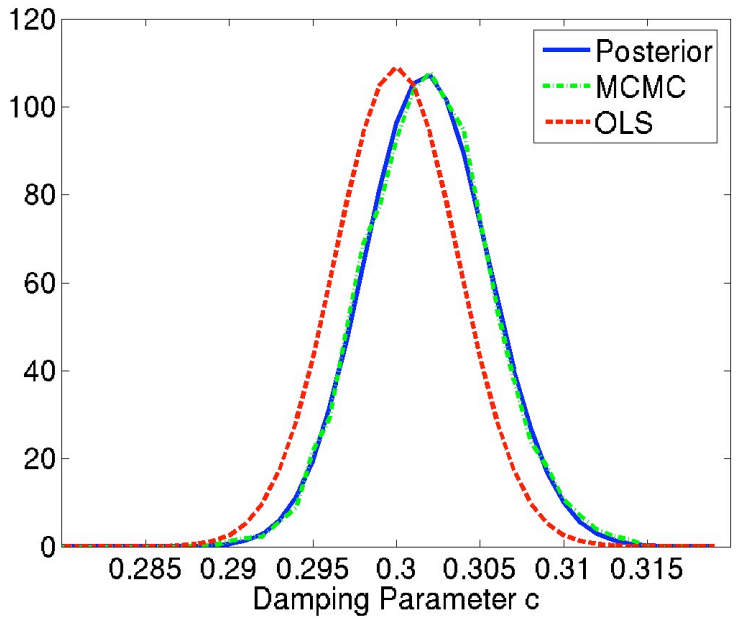
$$\pi(c|y) = \frac{p(y|c)}{\int_0^{\infty} p(y|c)dc}$$

Strategy: Create Markov chain using random sampling so that created chain has the posterior distribution as its limiting (stationary) distribution.

Problem: $SS_{\min} \approx 5.18$

$$\Rightarrow e^{-SS_{\min}/2\sigma_0^2} \approx 3.3 \times 10^{-113}$$

Solution: Reformulate but slow even for one parameter!



Markov Chains

Definition: Sequence of random variables $X_1, X_2, \cdots$ that satisfy Markov property: X_{n+1} depends only on X_n; that is

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, \cdots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

where x_i is the state of the chain at time i.

Note: A Markov chain is characterized by three components: a state space, an initial distribution, and a transition kernel.

State Space: Range of X_i: Set of all possible values

Initial Distribution: (Mass)

$$p^0 = [p_1^0, p_2^0, \cdots, p_n^0] \quad , \quad p_i^0 = P(X_0 = x_i)$$

Transition Probability: (Markov Kernel)

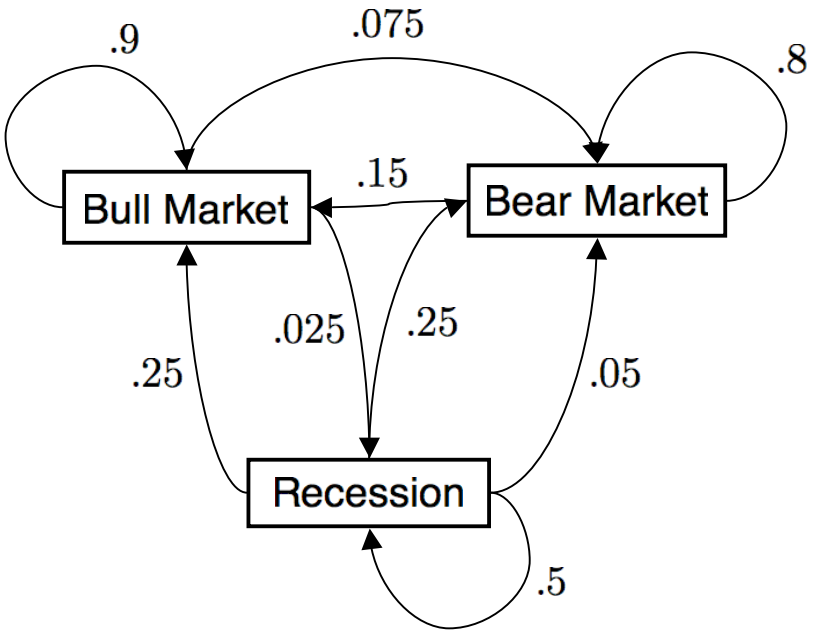
$$p_{ij} = P(X_{n+1} = x_j | X_n = x_i)$$

$$p_{ij}^{(n)} = P(X_{m+n} = x_j | X_m = x_i) \quad (n\text{-step transition probability})$$

$$P = [p_{ij}] \quad , \quad P_n = [p_{ij}^{(n)}]$$

Markov Chains

Example:

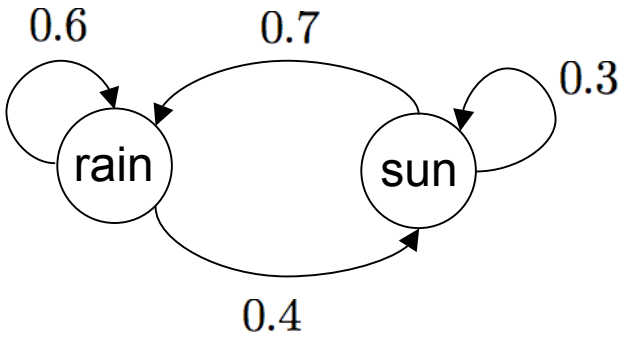


Chapman-Kolmogorov Equations: For any k such that $0 < k < n$,

$$p_{ij}^{(n)} = \sum_{r \in S} p_{ir}^{(k)} p_{rj}^{(n-k)}$$

Markov Chains: Limiting Distribution

Example: Raleigh weather -- Tomorrow's weather conditioned on today's weather



$$P = \begin{bmatrix} .6 & .4 \\ .7 & .3 \end{bmatrix} \quad S = \{\text{rain, sun}\}$$

- Distribution at Step n: $p^n = p^0 P^n$
- Note: Rows must sum to unity

Question:

(i) Can we say anything about $\lim_{n\to\infty} X_n$? Not really

(ii) What about $\lim_{n\to\infty} p^n = \pi$? Convergence in Distribution

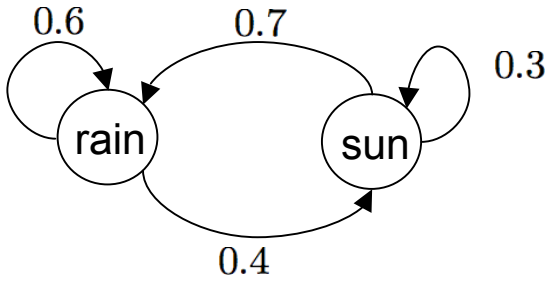
Note: If limit exists,

$$\pi = \lim_{n\to\infty} p^0 P^n = \lim_{n\to\infty} p^0 P^{n+1} = (\lim_{n\to\infty} p^0 P^n) P = \pi P$$

Definition: This is the limiting distribution (invariant measure)

Markov Chains: Limiting Distribution

Example: Raleigh weather



$$P = \begin{bmatrix} .6 & .4 \\ .7 & .3 \end{bmatrix}$$

Solve

$$\pi P = \pi \quad , \quad \sum \pi_i = 1$$

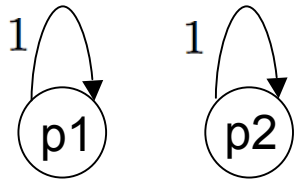
$$\Rightarrow [\pi_r \ , \ \pi_s] \begin{bmatrix} .6 & .4 \\ .7 & .3 \end{bmatrix} = [\pi_r \ , \ \pi_s] \quad , \quad \pi_r + \pi_s = 1$$

$$\Rightarrow \pi_r = .6364 \ , \ \pi_s = .6363$$

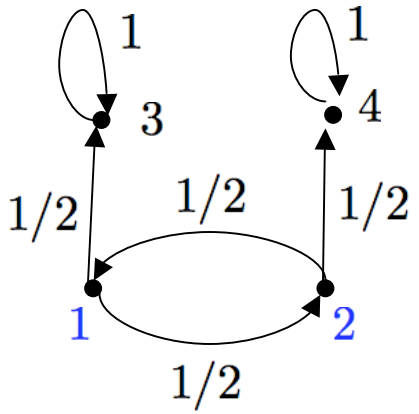
Code: stationary.m

Irreducible Markov Chains

Reducible Markov Chain:



$$p^0 = [p1\ ,\ p2] = \pi$$



$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

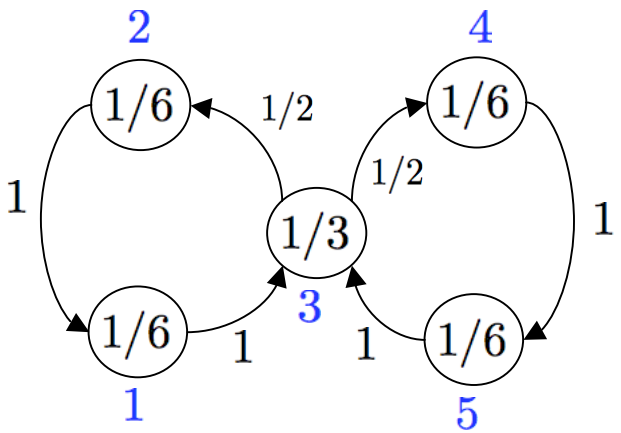
Note: Limiting distribution not unique if chain is reducible.

Irreducible: A Markov chain is *irreducible* if any state x_j and be reached from any state x_i in a finite number of steps; that is

$p_{ij}^{(n)} > 0$ for all states in finite n

Periodic Markov Chains

Example:



$$P = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\pi = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}$$

Note: Chain returns to state 1 at steps $3, 6, 9, \cdots$ so Period = 3

Note: Probability mass "cycles" through chain so no convergence

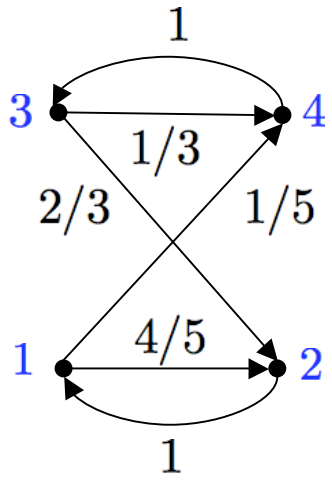
Periodicity: A Markov chain is *periodic* if parts of the state space are visited at regular intervals. The period *k* is defined as

$$\begin{aligned} k &= \gcd\left\{n | p_{ii}^{(n)} > 0\right\} \\ &= \gcd\left\{n | P(X_{m+n} = x_i | X_m = x_i) > 0\right\} \end{aligned}$$

- The chain is aperiodic if $k = 1$.

Periodic Markov Chains

Example:



$$P = \begin{bmatrix} 0 & 4/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$p^0 = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

$$p^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

Stationary Distribution

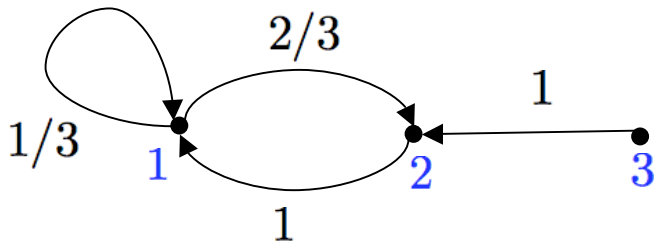
Theorem: A finite, homogeneous Markov chain that is irreducible and aperiodic has a unique stationary distribution π and the .chain will converge in the sense of distributions from any initial distribution p^0.

Recurrence (Persistence): A state x_i is recurrent (persistent) if the probability of returning to x_i is 1; that is,

$$P(X_{m+n} = x_i \text{ for some } n \geq 1 | X_m = x_i) = 1$$

- It is *transient* if probability strictly less than 1

Example: State 3 is transient



Ergodicity: A state is termed *ergodic* if it is aperiodic and recurrent. If all states of an irreducible Markov chain are ergodic, the chain is said to be *ergodic*.

Matrix Theory

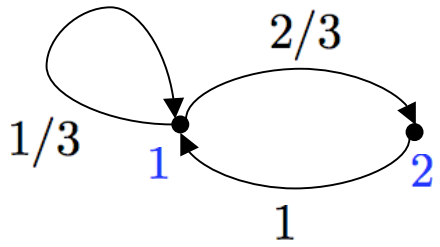
Definition: A matrix $A \in \mathbb{R}^{(n \times n)}$ is

(i) Nonnegative, denoted $A \geq 0$, if $a_{ij} \geq 0$ for all i, j

(ii) Strictly positive, denoted $A > 0$, if $a_{ij} > 0$ for all i, j

Lemma: Let P be the transition matrix of an ergodic finite Markov chain with state space S. Then for some $N_0 \geq 1, P_n > 0$ for all $n > N_0$.

Example:



$$P = \begin{bmatrix} 1/3 & 2/3 \\ 1 & 0 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 7/9 & 2/9 \\ 1/3 & 2/3 \end{bmatrix}$$

Matrix Theory

Theorem (Perron-Frobenius): For any strictly positive matrix $A > 0$, there exist $\lambda_0 > 0$ and $x_0 > 0$ such that

(i) $Ax_0 = \lambda_0 x_0$

(ii) If $\lambda \neq \lambda_0$ is any other eigenvalue of A, then $|\lambda| < \lambda_0$

(iii) λ_0 has geometric and algebraic multiplicity 1

Corollary 1: If $A \geq 0$ is a nonnegative matrix such that $A^n > 0$, then theorem also applies to A.

Proposition: Let $A > 0$ be a strictly positive $n \times n$ matrix with row and column sums

$$r_i = \sum_j a_{ij} \quad , \quad c_j = \sum_i a_{ij} \ , \ i, j = 1, \cdots, n$$

Then

$$\min_i r_i \leq \lambda_0 \leq \max_i r_i \quad , \quad \min_j c_j \leq \lambda_0 \leq \max_j c_j$$

Stationary Distribution

Corollary: Let $P \geq 0$ be the transition matrix of an ergodic Markov chain. Then there exists a unique stationary distribution π such that $\pi P = \pi$.

Proof: By Lemma and Corollary 1, P has a largest eigenvalue $\lambda_0 = 1$.

Since multiplicity is 1, unique π such that $\pi P = \pi$ and $\sum_i \pi_i = 1$.

Convergence: Express

$$UPV = \Lambda = \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

where $1 > |\lambda_2| \geq \cdots \geq |\lambda_n|, V = U^{-1}$

Note:

$$P^n = V \begin{bmatrix} 1 & & & \\ & \lambda_2^n & & \\ & & \ddots & \\ & & & \lambda_n^n \end{bmatrix} U \to V \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} U$$

Stationary Distribution

Note:

$$UP = \Lambda U \Rightarrow \begin{bmatrix} \pi_1 & \cdots & \pi_n \\ \vdots & & \vdots \\ & & \end{bmatrix} \begin{bmatrix} & & \\ & P & \\ & & \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_n \\ \vdots & & \vdots \\ & & \end{bmatrix}$$

and

$$V = U^{-1} = \begin{bmatrix} \pi_1 & \cdots & \pi_n \\ \vdots & & \vdots \\ & & \end{bmatrix} \begin{bmatrix} 1 & \cdots & \\ \vdots & & \\ 1 & \cdots & \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

Thus

$$\begin{aligned} \lim_{n\to\infty} p^n &= \lim_{n\to\infty} p^0 P^n \\ &= \lim_{n\to\infty} \begin{bmatrix} p_1^0 & \cdots & p_n^0 \end{bmatrix} \begin{bmatrix} 1 & \cdots & \\ \vdots & & \\ 1 & \cdots & \end{bmatrix} \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_n \\ \vdots & & \vdots \\ & & \end{bmatrix} \\ &= \begin{bmatrix} p_1^0 & \cdots & p_n^0 \end{bmatrix} \begin{bmatrix} 1 & \cdots & \\ \vdots & & \\ 1 & \cdots & \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_n \\ \vdots & & \vdots \\ & & \end{bmatrix} \\ &= \begin{bmatrix} \pi_1 \cdots \pi_n \end{bmatrix} \\ &= \pi \end{aligned}$$

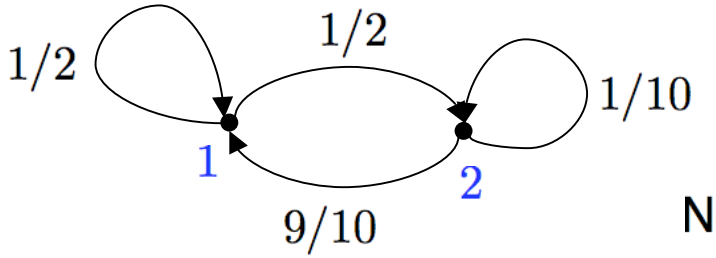
Detailed Balance Conditions

Reversible Chains: A Markov chain determined by the transition matrix $P = [p_{ij}]$ is reversible if there is a distribution π that satisfies the detailed balance conditions

$$\pi_i p_{ij} = \pi_j p_{ji}$$

Proof: We need to show that $\pi_j = \sum_i \pi_i p_{ij}$. Note that $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji}$

Example:



$$P = \begin{bmatrix} 1/2 & 1/2 \\ 9/10 & 1/10 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 9/14 & 5/14 \end{bmatrix}$$

Note: $\frac{1}{2} \cdot \frac{9}{14} = \frac{9}{10} \cdot \frac{5}{14}$ so detailed balance satisfied

Markov Chain Monte Carlo Methods

Strategy: Markov chain simulation used when it is impossible, or computationally prohibitive, to sample θ directly from

$$\pi(\theta|y) = \frac{p(\theta|y)}{\int p(\theta|y)dy}$$

or an unnormalized density $p(\theta|y)$.

- Create a Markov process whose stationary distribution is $\pi(\theta|y)$ or $p(\theta|y)$.

Note:

• In Markov chain theory, we are given a Markov chain, P, and we construct its equilibrium distribution.

• In MCMC theory, we are "given" a distribution and we want to construct a Markov chain that is reversible with respect to it.

Markov Chain Monte Carlo Methods

General Strategy:

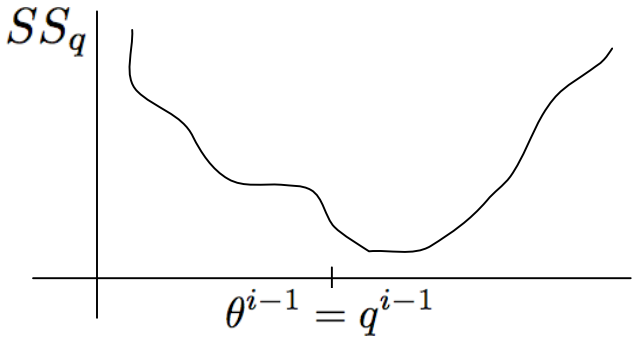
- Current value: $X_n = \theta^{i-1}$
- Propose candidate $\theta^* \sim J_i(\theta^*|\theta^{i-1})$ from proposal (jumping) distribution
- With probability $\alpha(\theta^*, \theta^{i-1})$, accept θ^*; i.e., $X_{n+1} = \theta^*$
- Otherwise, stay where you are: $X_{n+1} = \theta^{i-1}$

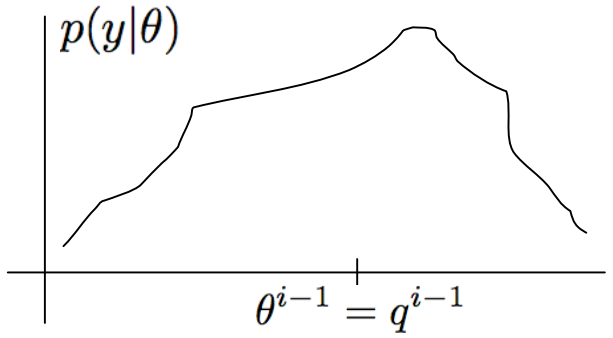
Intuition: Recall that

$$\pi(\theta|y) = \frac{p(y|\theta)\pi_0(\theta)}{\int_{\mathbb{R}^p} p(y|\theta)\pi_0(\theta)d\theta}$$

where

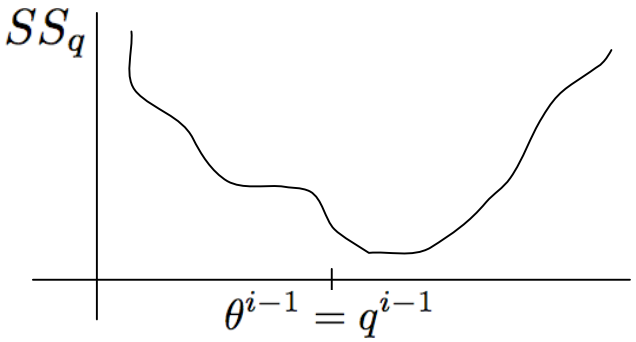
$$p(y|\theta) = p(y|q) = \frac{1}{(2\pi\sigma)^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n}[y_j - y(t_j;q)]^2/\sigma^2} = \frac{1}{(2\pi\sigma)^{n/2}} e^{-\frac{1}{2}SS_q/\sigma^2}$$

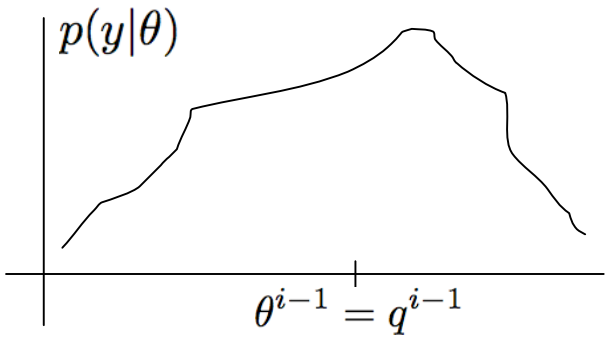




Markov Chain Monte Carlo Methods

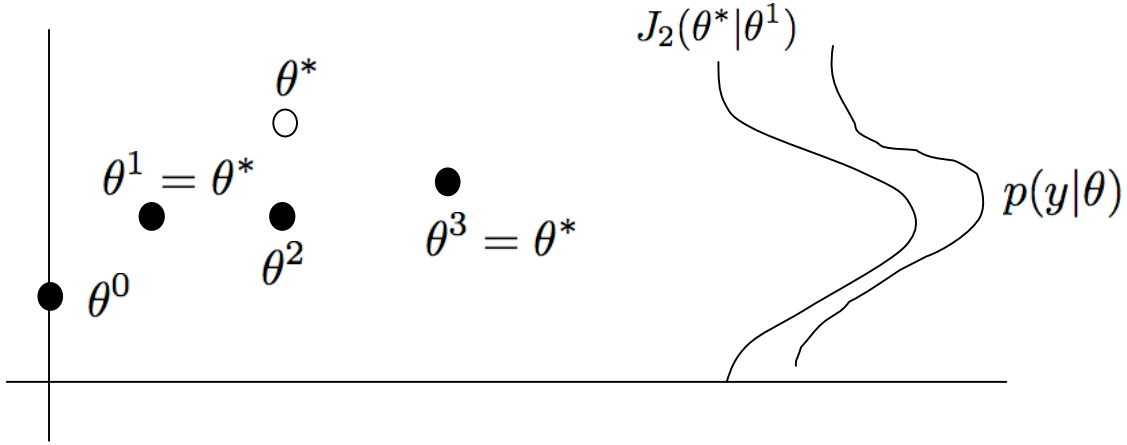
Intuition:





- Consider $r(\theta^*, \theta^{i-1}) = \frac{p(y|\theta^*)}{p(y|\theta^{i-1})}$
  - If $r < 1 \Leftrightarrow p(y|\theta^*) < p(y|\theta^{i-1})$, accept with probability $\alpha = r$
  - If $r > 1$, accept with probability $\alpha = 1$

Note: Narrower proposal distribution yields higher probability of acceptance.



Metropolis Algorithm

Metropolis Algorithm: [Metropolis and Ulam, 1949]

1. Initialization

   - Choose a starting value θ^0 for which $\pi(\theta|y) > 0$

2. For $i = 1, 2, \cdots, N$

   (a) Choose candidate θ^* from proposal (jumping) distribution $J_i(\theta^*|\theta^{i-1})$.
   For Metropolis, distribution is symmetric; i.e.,

$$J_i(\theta^*|\theta^{i-1}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\theta^* - \theta^{i-1})^2/2\sigma^2} = J_i(\theta^{i-1}|\theta^*)$$

   (b) Compute ratio

$$r(\theta^*, \theta^{i-1}) = \frac{\pi(\theta^*|y)}{\pi(\theta^{i-1}|y)} = \frac{p(y|\theta^*)\pi_0(\theta^*)}{p(y|\theta^{i-1})\pi_0(\theta^{i-1})} = \frac{p(y|\theta^*)}{p(y|\theta^{i-1})}$$

   if noninformative prior

   (c) Set

$$\theta^i = \begin{cases} \theta^* & , \text{ with probability } \alpha = \min(1, r) \\ \theta^{i-1} & , \text{ else} \end{cases}$$

Metropolis-Hastings Algorithm

Metropolis-Hastings Algorithm: $J_i(\theta^*|\theta^{i-1})$ does not have to be symmetric

- Acceptance Ratio:

$$r(\theta^*, \theta^{i-1}) = \frac{\pi(\theta^*|y)/J_i(\theta^*|\theta^{i-1})}{\pi(\theta^{i-1}|y)/J_i(\theta^{i-1}|\theta^*)} = \frac{p(\theta^*|y)J_i(\theta^{i-1}|\theta^*)}{p(\theta^{i-1}|y)J_i(\theta^*|\theta^{i-1})}$$

Examples:

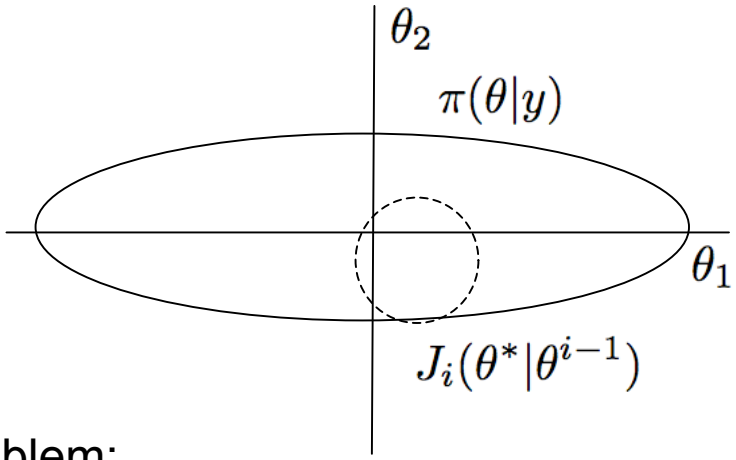
- Cauchy Distribution: $J_i(\theta^*|\theta^{i-1}) = \frac{1}{\pi(1+(\theta^*)^2)}$
- $\chi^2(n)$ Distribution: $J_i(\theta^*|\theta^{i-1}) = C(\theta^*)^{n/2-1}e^{\theta^*/2}$

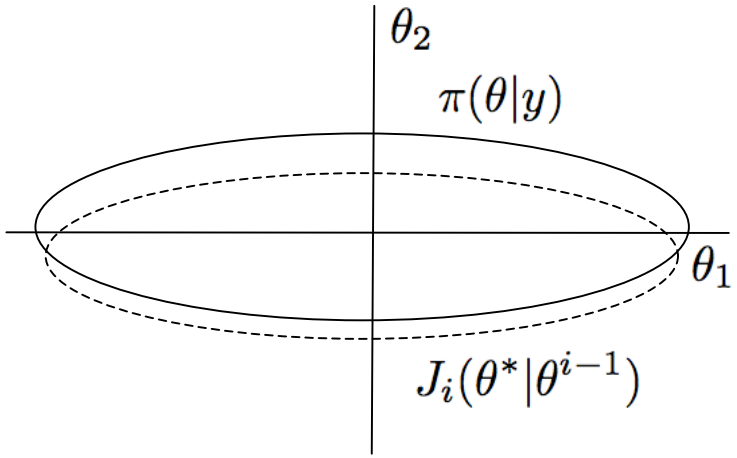
Note: Considered one of top 10 algorithms of 20th century

Proposal Distribution

Proposal Distribution: Significantly affects mixing

- Too wide: Too many points rejected and chain stays still for long periods;
- Too narrow: Acceptance ratio is high but algorithm is slow to explore parameter space
- Ideally, it should have similar “shape” to posterior (target) distribution.





Problem:

- Anisotropic posterior, isotropic proposal;
- Efficiency nonuniform for different parameters

Result:

- Recovers efficiency of univariate case

Proposal Distribution and Acceptance Probability

Proposal Distribution: Two basic approaches

- Choose a fixed proposal function
  - o Independent Metropolis
- Random walk (local Metropolis)

$$\theta^* = \theta^{i-1} + Rz$$

  - o Two (of several) choices: $z \sim N(0,1)$

(i) $R = cI \;\Rightarrow\; \theta^* \sim N(\theta^{i-1}, cI)$

(ii) $R^T R = V\ ,\ V = \sigma^2(\chi^T\chi)^{-1} \;\Rightarrow\; \theta^* \sim N(\theta^{i-1}, V)$

Acceptance Probability:

$$r(\theta^*, \theta^{i-1}) = \frac{p(y|\theta^*)}{p(y|\theta^{i-1})}$$

$$= \frac{e^{-\frac{1}{2}SS_{\theta^*}/\sigma^2}}{e^{-\frac{1}{2}SS_{\theta^{i-1}}/\sigma^2}}$$

$$= e^{-\frac{1}{2}[SS_{\theta^*} - SS_{\theta^{i-1}}]/\sigma^2}$$

Random Walk Metropolis Algorithm for Parameter Estimation

1. Set N: Number of simulations
2. Determine

$$\theta^0 = q^0 = \arg\min_{q \in Q} \sum_{j=1}^{n} [y_j - y(t_j; q)]^2 = \arg\min_{q \in Q} SS_q$$

   using favorite optimization routine

3. Set $SS_{\theta^0} = SS_{q^0} = \sum_{j=1}^{n} [y_j - y(t_j; q^0)]^2$
4. Estimate error variance

$$s^2 = \frac{1}{n-p} \sum_{j=1}^{n} [y_j - y(t_j; q^0)]^2$$

5. Construct covariance matrix estimate

$$V = s^2 \left[\chi^T(q^0) \chi(q^0) \right]^{-1} \ , \ \chi_{jk}(q^0) = \frac{\partial y(t_j; q^0)}{\partial q_k}$$

Notes:

- Rank will indicate parameter identifiability (or lack thereof)
- Adaptive techniques can be used to update V as algorithm proceeds

Random Walk Metropolis Algorithm for Parameter Estimation

6. Compute $R = \text{chol}(V)$
7. For $i = 1, 2, \cdots, N$
   (a) Sample $z \sim N(0, 1)$
   (b) Construct candidate

$$\theta^* = \theta^{i-1} + Rz$$

   (c) Sample $u_\alpha \sim U[0, 1]$
   (d) Compute

$$SS_{\theta^*} = \sum_{j=1}^{n} [y_j - y(t_j; \theta^*)]^2$$

   (e) Compute

$$\alpha(\theta^*, \theta^{i-1}) = \min\left(1, e^{-[SS_{\theta^*} - SS_{\theta^{i-1}}]/2s^2}\right)$$

   (f) If $u_\alpha < \alpha$

   Set $\theta^i = \theta^*$, $SS_{\theta^i} = SS_{\theta^*}$

   else

   Set $\theta^i = \theta^{i-1}$

   endif

Markov Chain Monte Carlo: Example

Example: Consider the spring model

$$m\ddot{y} + c\dot{y} + ky = 0$$

$$y(0) = 2\ ,\ \dot{y}(0) = 0$$

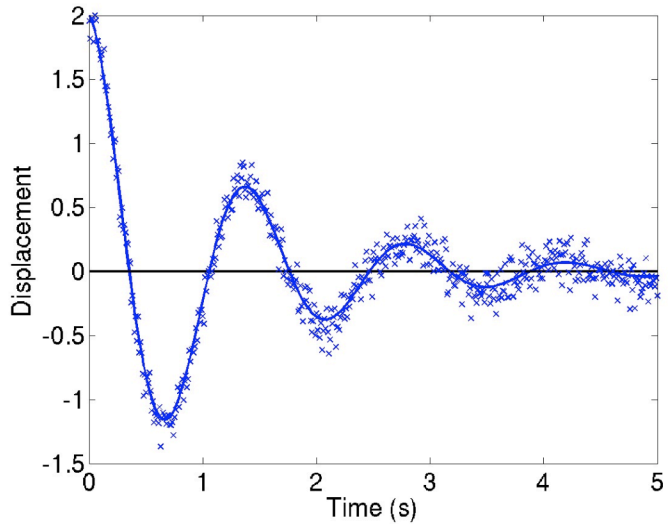
Note: $m = 2, k = 4.1, c = q_0 = 0.3$

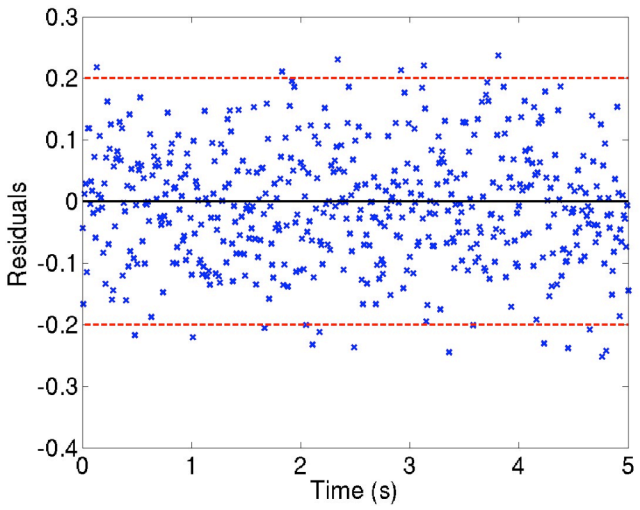
$n = 501$

which has the solution

$$y(t) = 2e^{-ct/2m} \cos\left(\frac{\sqrt{4mk - c^2}}{2m}\, t\right)$$

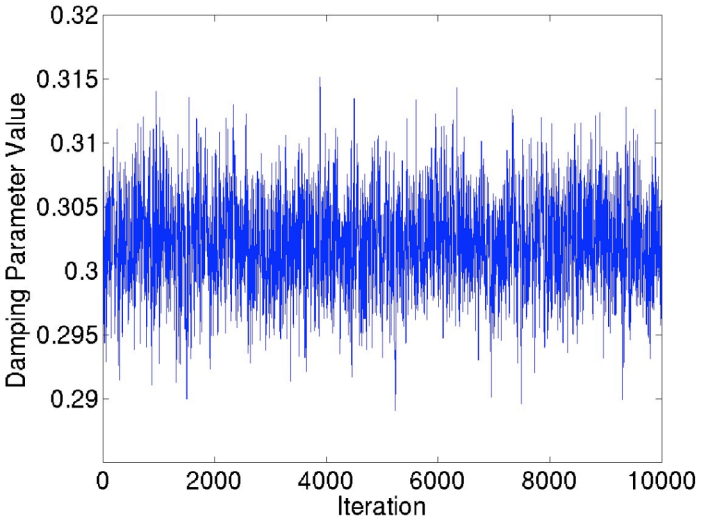
when $c^2 - 4km < 0$. Assume that m, k known and take $q = c$. We also assume that $\varepsilon_j \sim N(0, \sigma_0^2)$ where $\sigma_0 = 0.1$.

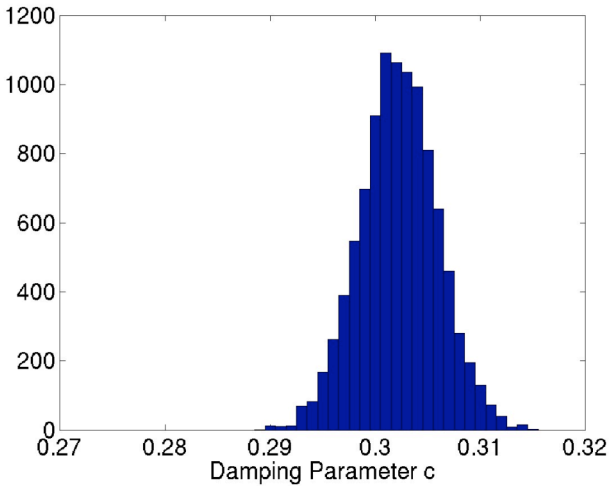


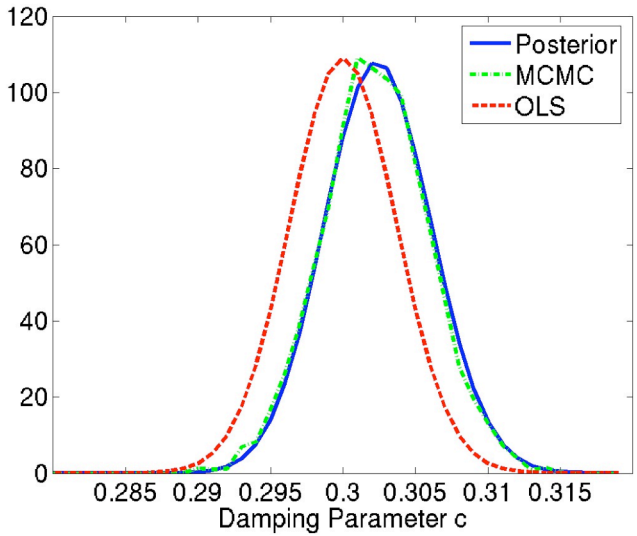


Markov Chain Monte Carlo: Example

Example: Single parameter c







Markov Chain Monte Carlo: Example

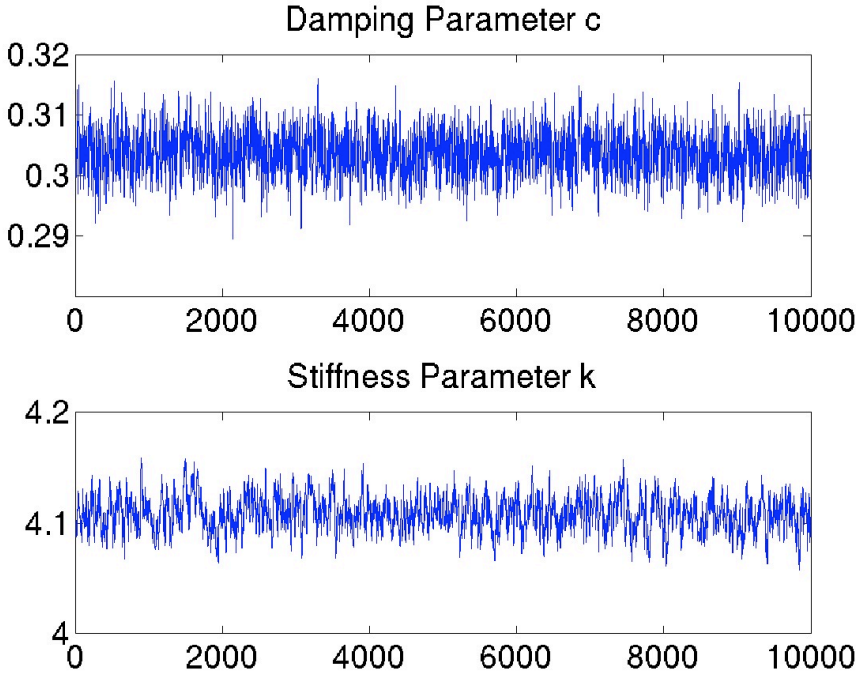
Example: Consider the spring model

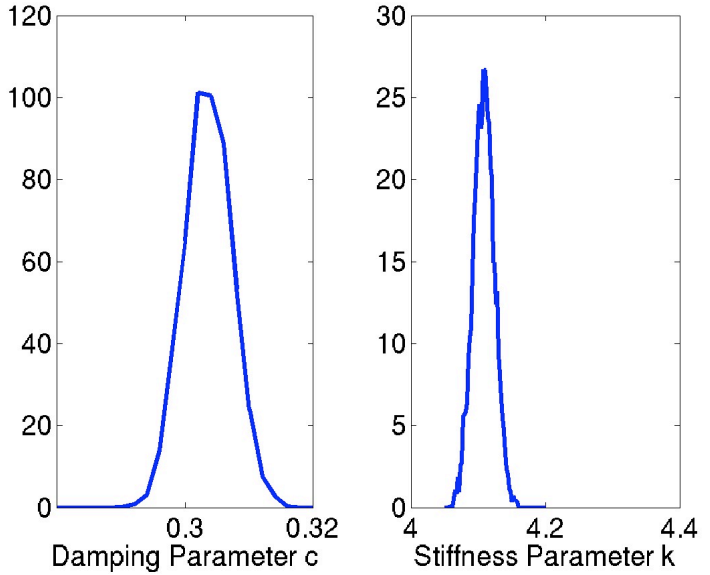
$$m\ddot{y} + c\dot{y} + ky = 0$$

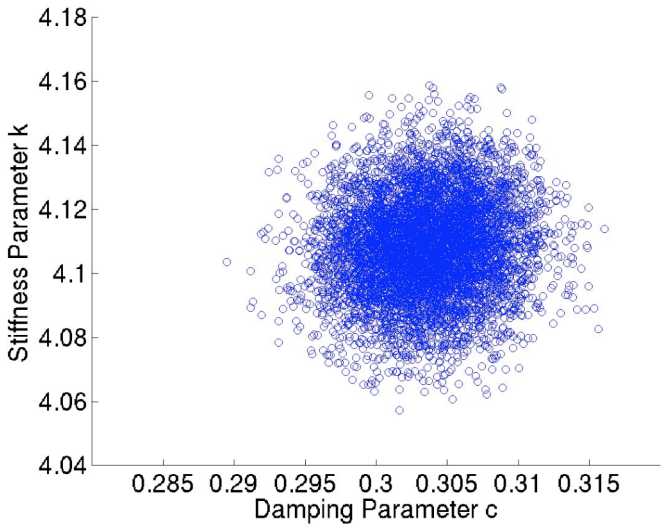
$$y(0) = 2\ ,\ \dot{y}(0) = 0$$

where we take $\theta = q = (c, k)$.

Case i: $c_0 = 0.3, k_0 = 4.1$

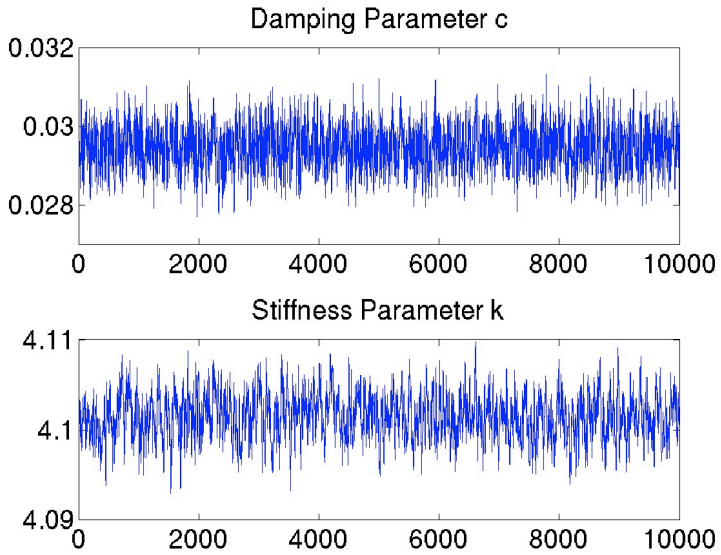


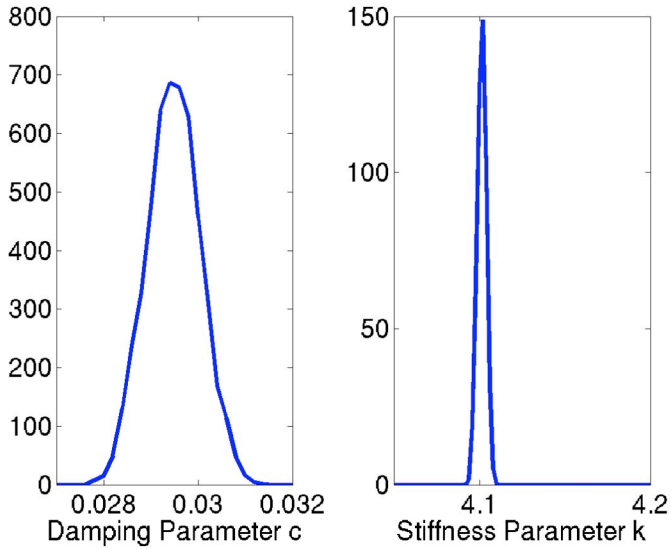




Markov Chain Monte Carlo: Example

Case i: $c_0 = 0.03, k_0 = 4.1$



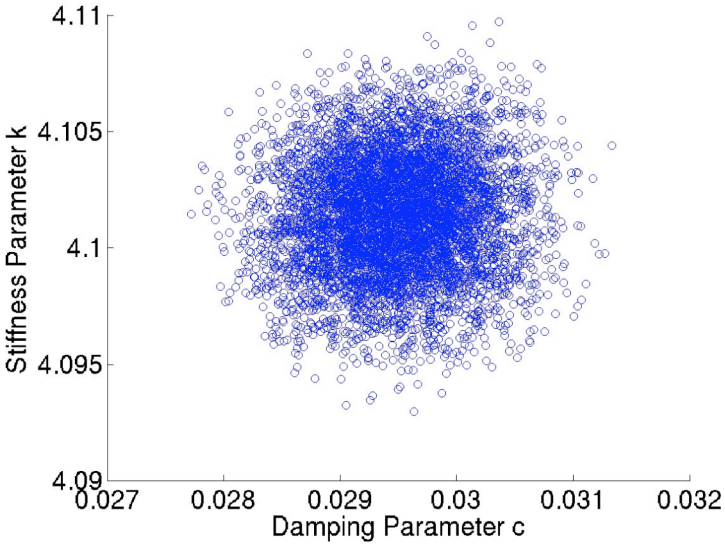


Note:

- Anisotropic posterior

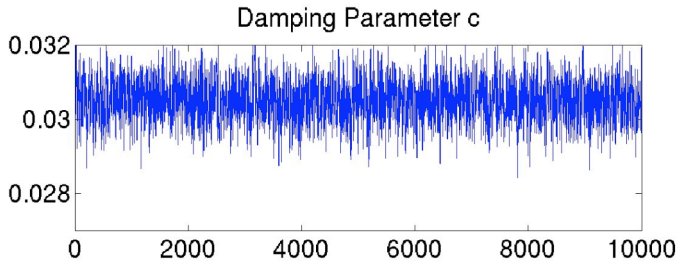
$$V = \begin{bmatrix} 0.0316 \times 10^{-5} & -0.0020 \times 10^{-5} \\ -0.0020 \times 10^{-5} & 0.1800 \times 10^{-5} \end{bmatrix}$$

- Normal with constant variance is less effective
- Little correlation between parameters

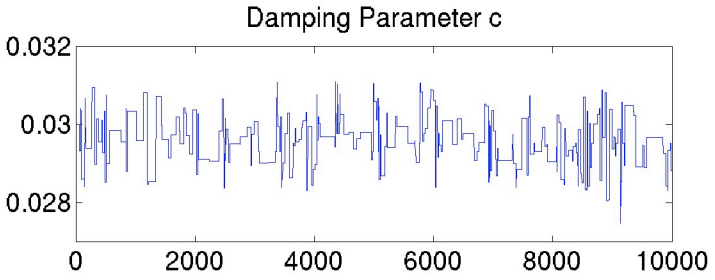


Markov Chain Monte Carlo: Example

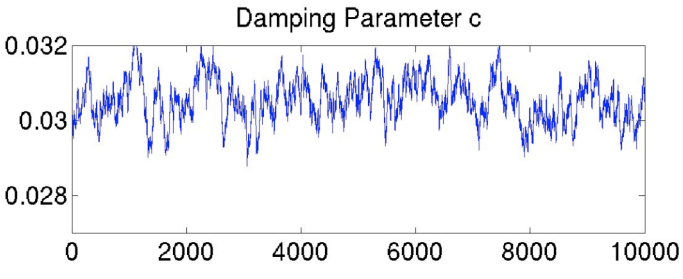
Case i: $c_0 = 0.03, k_0 = 4.1$



$$\theta^* \sim N(\theta^{i-1}, 10^{-6} I)$$



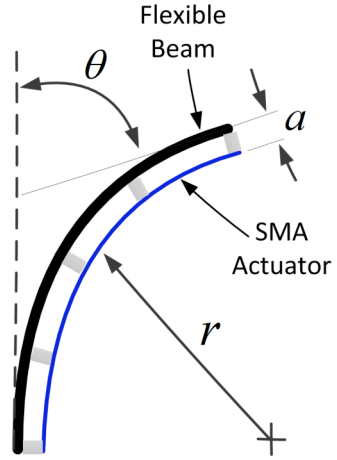
$$\theta^* \sim N(\theta^{i-1}, 10^{-4} I)$$

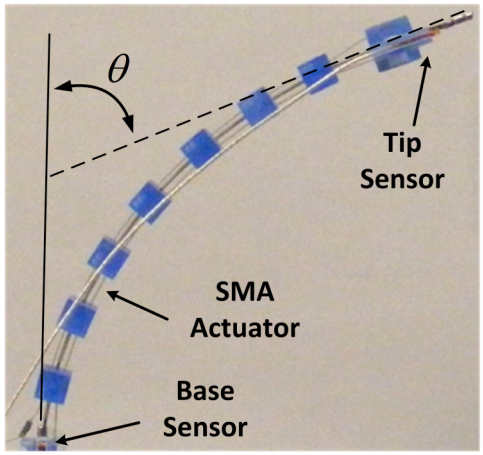


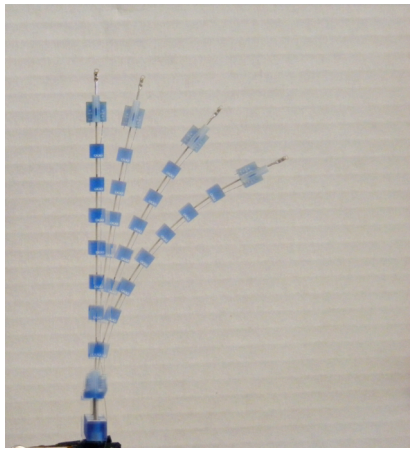
$$\theta^* \sim N(\theta^{i-1}, 10^{-8} I)$$

Markov Chain Monte Carlo: Example

Example: SMA-driven bending actuator -- talk with John Crews







Model:

$$y(t) = \theta(t) = \frac{aA_cL\sigma(t)}{EI}$$

$$\sigma(t) = \frac{EI}{a^2A_c}(\varepsilon_P - \varepsilon(t))$$

$$\varepsilon(t) = \int_0^\infty \int_\infty^\infty \bar{\varepsilon}(\sigma(t) + \sigma_I, T(t); \sigma_R)\nu_R(\sigma_R)\nu_I(\sigma_I)d\sigma_I d\sigma_R$$

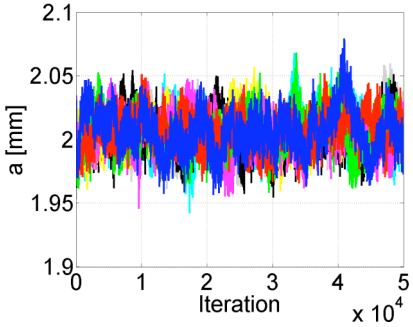
$$\bar{\varepsilon}(\sigma(t) + \sigma_I, T(t); \sigma_R) = x_A(t)\frac{\sigma(t)}{E_A} + x_{M^+}(t)\left(\frac{\sigma(t)}{E_M} + \varepsilon_T\right) + x_{M^-}(t)\left(\frac{\sigma(t)}{E_M} - \varepsilon_T\right)$$

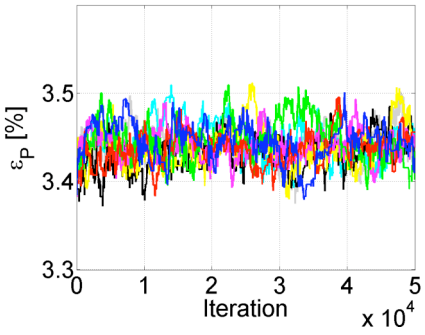
Estimated Parameters:

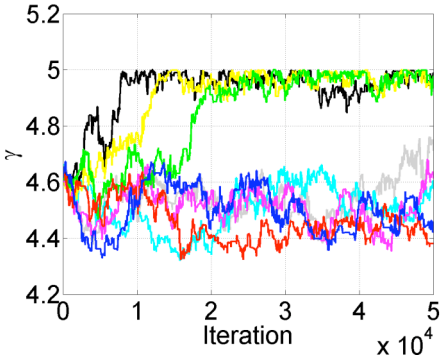
$$q = [a, \varepsilon_P, h, \gamma, EI]$$

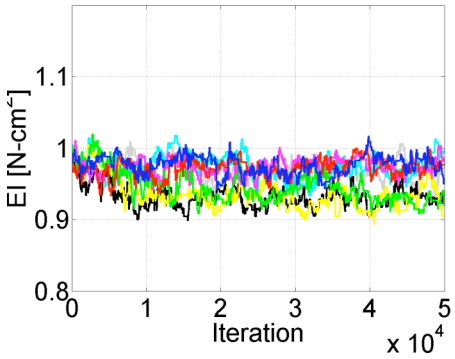
Markov Chain Monte Carlo: Example

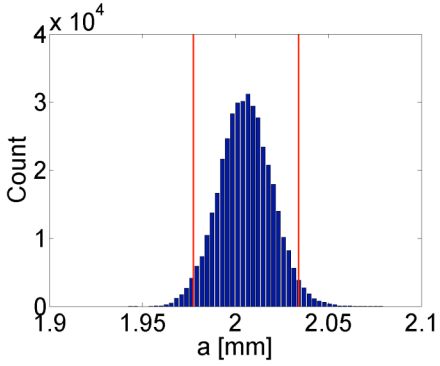
Example: SMA-driven bending actuator -- talk with John Crews

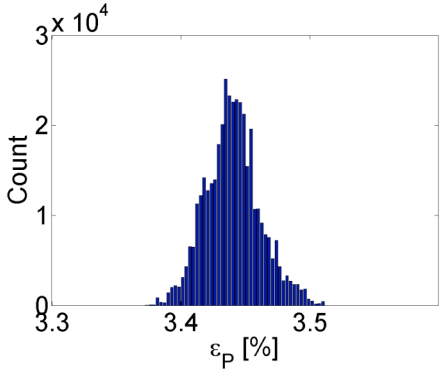


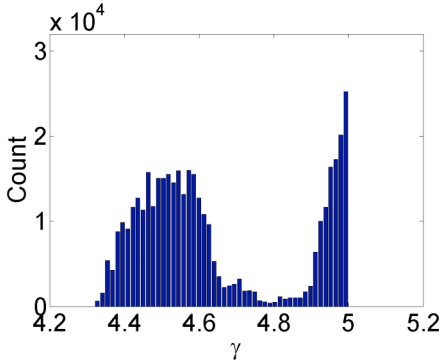


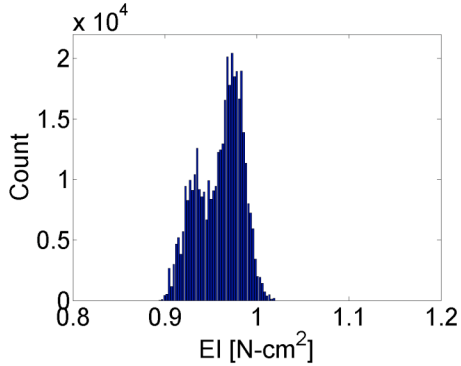






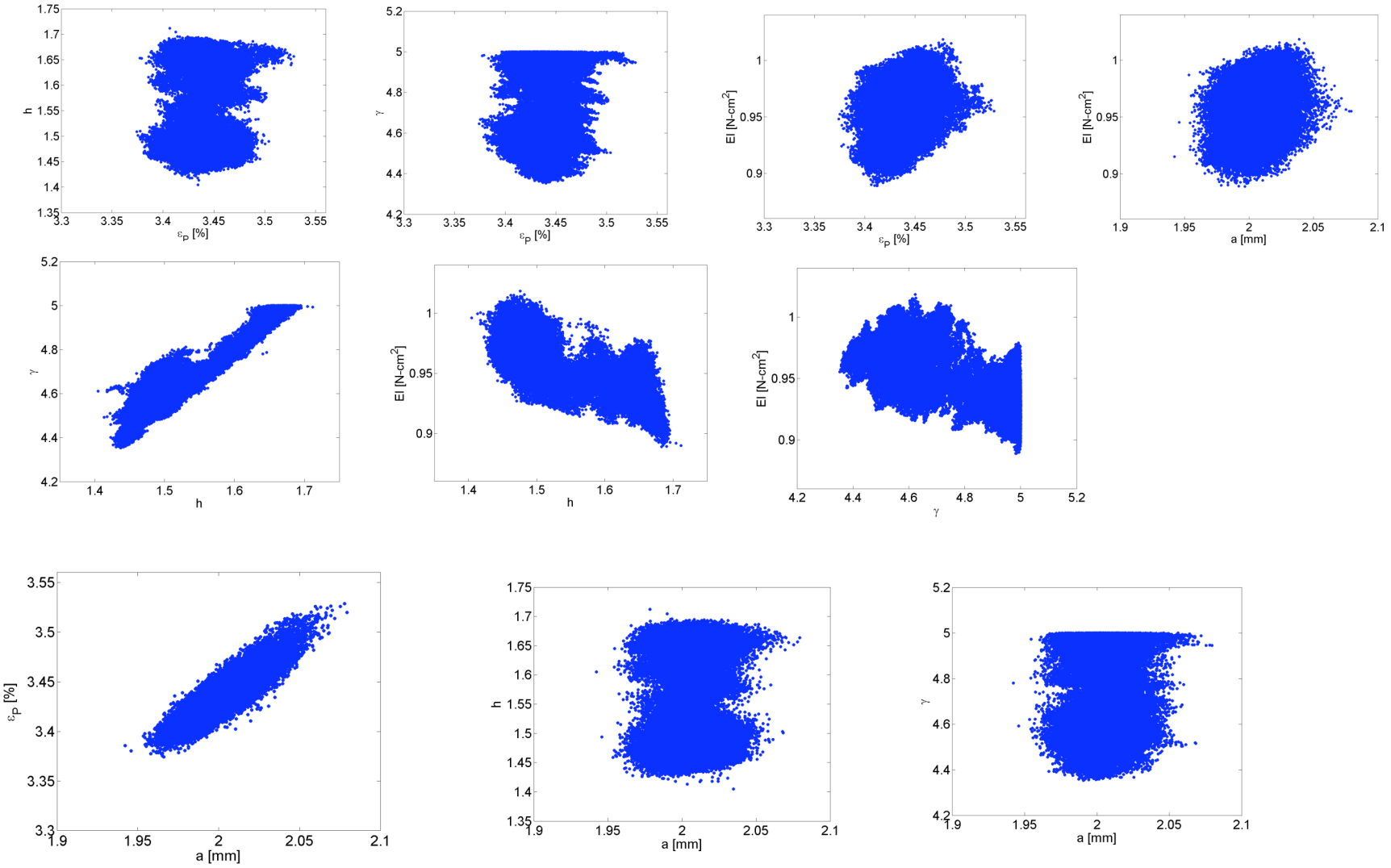






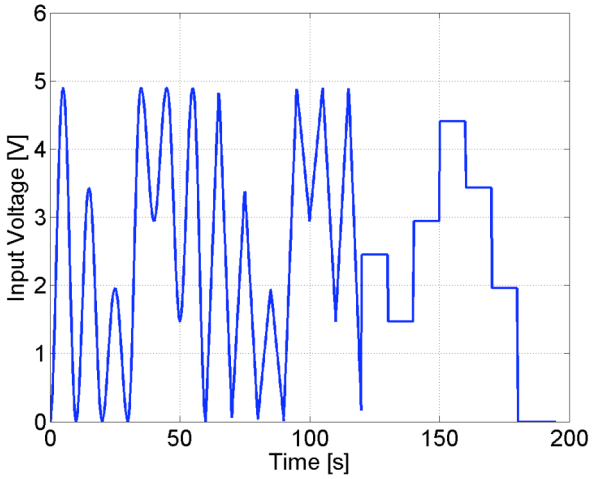
Markov Chain Monte Carlo: Example

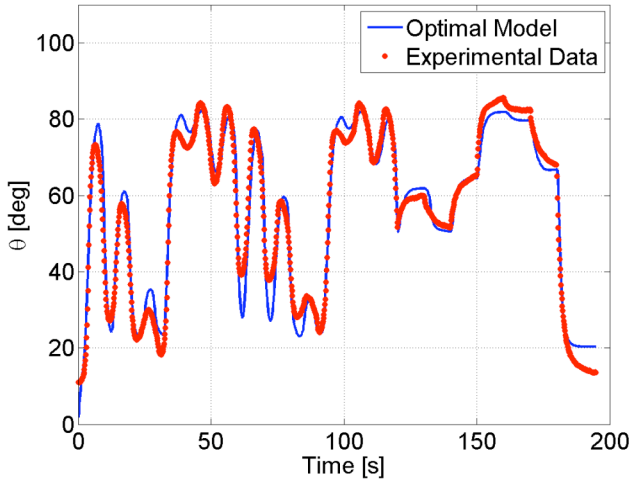
Example: SMA-driven bending actuator -- talk with John Crews



Markov Chain Monte Carlo: Example

Example: SMA-driven bending actuator -- talk with John Crews





Transition Kernel and Detailed Balance Condition

Transition Kernel: Recall that

$$\begin{aligned} p_{i-1,i} &= P(X_{n+1} = x_i | X_n = x_{i-1}) \\ &= P(X_{n+1} = \theta^i | X_n = \theta^{i-1}) \end{aligned}$$

is the probability of moving from state θ^{i-1} to state θ^i. Here

$$\begin{aligned} p_{i-1,i} &= P\left(\text{proposing } \theta^i\right) P\left(\text{accepting } \theta^i\right) \\ &= J(\theta^i | \theta^{i-1}) \alpha(\theta^i, \theta^{i-1}) \\ &= J(\theta^i | \theta^{i-1}) \min\left(1, \frac{p(\theta^i|y) J(\theta^{i-1}|\theta^i)}{p(\theta^{i-1}|y) J(\theta^i|\theta^{i-1})}\right) \end{aligned}$$

Detailed Balance Condition:

$$\begin{aligned} &\pi_i p_{ij} = \pi_j p_{ji} \\ &\Rightarrow \pi_{i-1} p_{i-1,i} = \pi_i p_{i,i-1} \\ &\Rightarrow p(\theta^{i-1}|y) p_{i-1,i} = p(\theta^i|y) p_{i,i-1} \end{aligned}$$

Transition Kernel and Detailed Balance Condition

Detailed Balance Condition: Here

$$p(\theta^{i-1}|y)p_{i-1,i} = p(\theta^{i-1}|y)J(\theta^i|\theta^{i-1})\min\left(1, \frac{p(\theta^i|y)J(\theta^{i-1}|\theta^i)}{p(\theta^{i-1}|y)J(\theta^i|\theta^{i-1})}\right)$$

$$= p(\theta^i|y)J(\theta^{i-1}|\theta^i)\min\left(1, \frac{p(\theta^{i-1}|y)J(\theta^i|\theta^{i-1})}{p(\theta^i|y)J(\theta^{i-1}|\theta^i)}\right)$$

$$= p(\theta^i|y)p_{i,i-1}$$

Note:

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

Transition Kernel: Definition

$$p_{ij} = J(\theta^j|\theta^i)\min\left(1, \frac{p(\theta^j|y)J(\theta^i|\theta^j)}{p(\theta^i|y)J(\theta^j|\theta^i)}\right), \; i \neq j$$

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}$$

Sampling Error Variance

Strategy: Treat error variance σ^2 as parameter to be estimated.

Recall: Assumption that errors

$$\varepsilon_j = y_j - y(t_j; q)$$

are normally distributed yields

$$p\left((q,y)|\sigma^2\right) \propto \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\varepsilon_j^2} = \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}SS_q}$$

Goal: Determine posterior distribution

$$p\left(\sigma^2|(q,y)\right) \propto p\left((q,y)|\sigma^2\right) p(\sigma^2)$$

Strategy:

- Choose prior so that posterior is from same family --- termed *conjugate prior*.
- For normal distribution with unknown variance, conjugate prior is inverse Gamma distribution $\Gamma^{-1}(\alpha,\beta)$ which is equivalent to inverse χ^2-distribution, Inv-$\chi^2(n,s^2)$.

Sampling Error Variance

Definition:

- Inv-$\chi^2(n, s^2)$

  $$p(x) \propto x^{-(n/2+1)} e^{-(ns^2)/2x}$$

- $\Gamma^{-1}(\alpha, \beta)$

  $$p(x) \propto x^{-(\alpha+1)} e^{-\beta/x}$$

Strategy: Take $\sigma^2 \sim \text{Inv} - \chi^2(n_0, \sigma_0^2)$ to get

$$\begin{aligned} p\left(\sigma^2|(q,y)\right) &\propto p(\sigma^2) p\left((q,y)|\sigma^2\right) \\ &\propto (\sigma^2)^{-(n_0/2+1)} e^{-n_0\sigma_0^2/2\sigma^2} \cdot (\sigma^2)^{-n/2} e^{-SS_q/2\sigma^2} \\ &= (\sigma^2)^{-[(n_0+n)/2+1]} e^{-\frac{1}{2}(n_0\sigma_0^2+SS_q)/\sigma^2} \end{aligned}$$

Thus

$$\sigma^2|(q,y) \sim \Gamma^{-1}\left(\frac{n_0+n}{2}, \frac{n_0\sigma_0^2+SS_q}{2}\right)$$

Note:

$$X \sim \Gamma^{-1}(a,b) \Leftrightarrow Y = \frac{1}{X} \sim \Gamma(a, 1/b)$$

Note:

- n_0 taken small;
  e.g., $n_0 = 1$ or $n_0 = .01$
- Take $\sigma_0^2 = s^2 = \frac{R^T R}{n-p}$

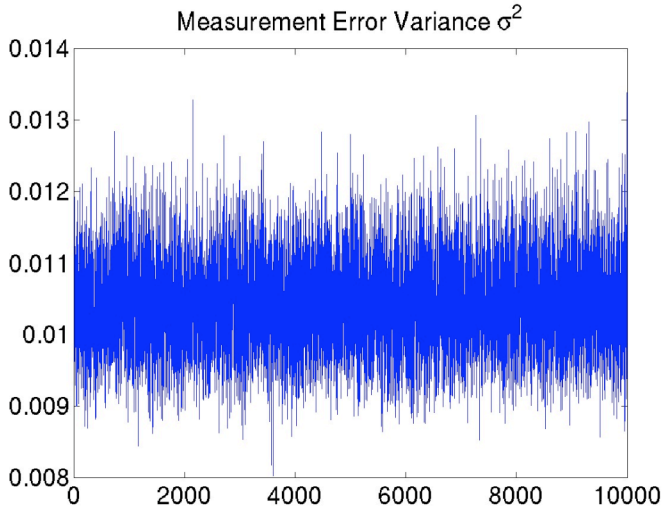
Sampling Error Variance

Example: Consider the spring model

$$m\ddot{y} + c\dot{y} + ky = 0$$

$$y(0) = 2\ ,\ \dot{y}(0) = 0$$

where we take $\theta = q = (c, k)$.



Related Topics

Note: This is an active research area and there are a number of related topics

- Burn in and convergence
- Adaptive algorithms
- Population Monte Carlo methods
- Sequential Monte Carlo methods and particle filters
- Gaussian mixture models
- Development of metamodels, surrogates and emulators to improve implementation speeds

References:

- A. Solonen, “Monte Carlo Methods in Parameter Estimation of Nonlinear Models,” Masters Thesis, 2006.
- H. Haario, E. Saksman, J. Tamminen, “An adaptive Metropolis algorithm,” *Bernoulli,* 7(2), pp. 223-242, 2001.
- C. Andrieu and J. Thomas, “A tutorial on adaptive MCMC,” *Statistics and Computing*, 18, pp. 343-373, 2008.
- M. Vihola, “Robust adaptive Metropolis algorithm with coerced acceptance rate,” arXiv:1011.4381v2.