Bayesian Statistics, Monte Carlo Methods

April 29, 2017

- Main objective of statistical theory: Derive from observations of a random phenomenon an inference about the probability distribution underlying this phenomenon.
- Parametric modelling: The observations x are the realization of a random variable X of probability density function $f(x|\theta)$.
- The function $f(x|\theta)$ considered as a function of θ for a fixed realization of the observation $X = x$ is called the likelihood function.
- The likelihood function is $l(\theta|x) = f(x|\theta)$ to emphasize that the observations are fixed.

- One alternative approach consists of incorporating as much as possible the complexity of a phenomenon, and thus aims at estimating the distribution underlying the phenomenon under minimal assumptions, generally using functional estimation (density, regression function, etc.).
- The parametric approach takes into account that a finite number of observations can efficiently estimate only a finite number of parameters.
- In any case, model checking/assessment or model choice should be considered.

Sufficiency principle

- When $X \sim f(x|\theta)$, a function T of X (also called a statistic) is said to be sufficient if the distribution of X conditional upon $T(X)$ is independent of θ,

$$P(x|\theta, T) = P(x|T),$$

or, equivalently:

$$P(\theta|T, x) = P(\theta|T), \text{ or } P(\theta, x|T) = P(\theta|T)P(x|T)$$

- Fisher Neyman Factorization

$$f(x|\theta) = h(x)g(T(x)|\theta).$$

- Example: Let $X = (X_1, \cdots, X_n)$ i.i.d. from $N \sim (\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$.

$$f(x|\theta) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_i - \mu)^2}{2\sigma^2})$$

- If $X_1, \cdots, X_n$ are independent and normally distributed with expected value θ (a parameter) and known finite variance σ^2, then $T(X) = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.
-

$$f(x|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{((x_i-\overline{x})-(\theta-\overline{x}))^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\overline{x})^2 + \sum_{i=1}^{n}(\theta-\overline{x})^2 - 2\sum_{i=1}^{n}(x_i-\overline{x})(\theta-\overline{x})\right)\right)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\theta-\overline{x})^2\right)\right)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\right) \exp\left(-\frac{n}{2\sigma^2}(\theta-\overline{x})^2\right)$$

- The joint density of the sample takes the form required by the Fisher-Neyman factorization theorem, by letting

$$h(x) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)$$

$$g(T(x)|\theta) = \exp\left(-\frac{n}{2\sigma^2}(\theta - \overline{x})^2\right)$$

- In general,

$$f(x|\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} - \frac{-\mu \sum_{i=1}^n x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2})$$

  $f(x|\theta)$ only depends on x through $(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)$, so $T(x) = (\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)$ is a set of sufficient statistics.
- Note that $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ is also a set of sufficient statistics because $\sum_{i=1}^n x_i^2 = s^2 - n\bar{x}^2$

- Consider the independent binomial rvs $X_1 \sim B(n_1, p)$, $X_2 \sim B(n_2, p)$, $X_3 \sim B(n_3, p)$, where n_1, n_2 and n_3 are known. Then

$$p(x_1, x_2, x_3|p) = C_{n_1}^{x_1} C_{n_2}^{x_2} C_{n_3}^{x_3} p^{x_1+x_2+x_3} (1-p)^{n_1+n_2+n_3-x_1-x_2-x_3}$$

  The statistics $T_1(x_1, x_2, x_3) = x_1 + x_2 + x_3$, or $T_2(x_1, x_2, x_3) = \frac{x_1+x_2+x_3}{n_1+n_2+n_3}$ are sufficient because $p(x_1, x_2, x_3|p)$ only depends on (x_1, x_2, x_3) through T_1 or T_2. $\frac{x_1}{n_1} + \frac{x_2}{n_2} + \frac{x_3}{n_3}$ is not sufficient.

- Let $X = (X_1, \cdots, X_n)$ be i.i.d. from $U(0, \theta)$ of density $f(x_i|\theta) = \theta^{-1}\mathbf{1}_{[0,\theta]}(x_i)$. Then

$$I(\theta|x) = f(x_1, x_2, \cdots, x_n|\theta) = \Pi_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n}\mathbf{1}_{[\max\{x_i\},\infty)}(\theta)$$

  The statistic $T(x) = \max(x_i)$ is sufficient.

- Let $X = (X_1, \cdots, X_n)$ i.i.d. from $P(\theta)$ of distribution $f(x_i|\theta) = e^{-\theta}\frac{\theta^{x_i}}{x_i!}$. Then

$$I(\theta|x) = \Pi_{i=1}^{n} f(x_i|\theta) = \frac{\theta^{\sum_{i=1}^{n} x_i}}{\Pi_{i=1}^{n} x_i!} e^{-n\theta}$$

  The statistic $T(x) = \sum_{i=1}^{n} x_i$ is sufficient.

- **Sufficiency principle**: Two observations x and y such that $T(x) = T(y)$ must lead to the same inference on θ.
- Consider the model $X_i \sim N(\mu, 1)$ and we want to estimate μ based on n data. In this case the sufficient statistic is $T(x_{1:n}) = \sum_{i=1}^{n} x_i$.
- Consider the estimate $\hat{\mu} = \frac{1}{n} T(x_{1:n})$, then this estimate satisfies the n sufficiency principle because if I have another dataset $x'_{1:n}$ such that $T(x_{1:n}) = T(x'_{1:n})$ then I obtain $\hat{\mu}_2 = \frac{1}{n} T(x'_{1:n}) = \frac{1}{n} T(x_{1:n}) = \hat{\mu}_1$.
- The estimate $\hat{\mu}_1 = x_1$ does not satisfies the sufficiency principle for $n > 1$ because even if I have another dataset $x'_{1:n}$ such that $T(x_{1:n}) = T(x'_{1:n})$, then $\hat{\mu}_2 = x'_1 \neq \hat{\mu}_2 = x_1$ if $x_1 \neq x'_1$.

Likelihood Principle

- **Likelihood Principle**. The information brought by an observation x about θ is entirely contained in the likelihood function $l(\theta|x) = f(x|\theta)$. Moreover, two likelihood functions contain the same information about θ if they are proportional to each other; i.e. $l_1(\theta|x) = c(x)l_2(\theta|x)$
- The maximum likelihood procedure does satisfy the likelihood principle

$$argmax_\theta l_1(\theta|x) = argmax_\theta l_2(\theta|x),$$

  if $l_1(\theta|x) = c(x)l_2(\theta|x)$.
- Classical approaches do not necessarily satisfy the likelihood principle.

- Testing Fairness. Suppose we want to test θ, the unknown probability of heads for possibly biased coin. Suppose

$$H_0 : \theta = 0.5, \text{v.s.} H_1 : \theta > 0.5.$$

- Scenario 1: Number of flips $n = 12$ predetermined and number of heads $X \sim B(n, \theta)$; that is if we collect $x = 9$ heads

$$P_\theta(X = x) = f(x|\theta) = C_n^x \theta^x (1-\theta)^{n-x} = C_{12}^9 \theta^9 (1-\theta)^3 = 220.\theta^9 (1-\theta)^3$$

- For a frequentist, the p-value of the test is $P_\theta(X \geq 9|H_0) = 0.073$ and H_0 is not rejected at level $\alpha = 0.05$.

- Scenario 2: Number of tails $\alpha = 3$ is predetermined, i.e. the flipping is continued until 3 tails are observed. Then $X \sim NB(3, 1-\theta)$ and assuming we collected $x = 9$ heads, then

$$P_\theta(X = x) = f(x|\theta) = C_{\alpha+x-1}^{\alpha-1}\theta^x(1-\theta)^\alpha = 55\theta^9(1-\theta)^3.$$

  For a frequentist, the p-value of the test is $P_\theta(X \geq 9|H_0) = 0.037$ and H_0 is rejected at level $\alpha = 0.05$.
- The likelihood principle is here violated because in both cases

$$f(x|\theta) \propto \theta^9(1-\theta)^3.$$

- Consider X_1, X_2 i.i.d. $N(\theta, 1)$. The likelihood function is

$$l(\theta|x_1, x_2) = f(x_1, x_2|\theta) \propto \exp(-(\frac{x_1 + x_2}{2} - \theta)^2)$$

- Now consider the alternative distribution

$$g(x_1, x_2|\theta) = \pi^{-3/2} \frac{\exp(-(\frac{x_1+x_2}{2} - \theta)^2)}{1 + (x_1 - x_2)^2} \propto l(\theta|x_1, x_2)$$

- If computing p-values, then one will obtain different results for $f(x_1, x_2|\theta)$ and $g(x_1, x_2|\theta)$ because of they have different tails and the likelihood principle will be violated.
- The likelihood principle does not bother about data you have not observed!

Stopping Rule Principle

- A direct implication of the likelihood principle is the stopping rule principle in sequential analysis.
- Consider a sequence of experiments that leads at time i to the observation $X_i \sim f(x_i|\theta)$ and we stops collecting data if at time n we have $(X_1, \cdots, X_n) \in A_n$; e.g. $A_n = \{X_1, \cdots, X_n : X_n > B\}$. In this case

$$l(\theta|x_1, \cdots, x_n) \propto \Pi_{i=1}^{n} f(x_i|\theta) \mathbf{1}_{A_n}(x_1, \cdots, x_n).$$

- Stopping rule principle: If a sequence of experiments is directed by a stopping rule which indicates when the experiments should stop, inference about θ must depend on the stopping rule only through the sample.

More on p-values

- Consider the case where $X_i \sim N(\theta, 1)$ and the hypothesis to be tested is $H_0 : \theta = 0$.
- The classical Neyman-Pearson test procedure at level 5% is to reject the hypothesis if $\frac{1}{n}|\sum_{i=1}^{n} x_i| > 1.96$ on the basis that

$$Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \theta| \geq \frac{1.96}{\sqrt{n}}|H_0) = Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i| \geq \frac{1.96}{\sqrt{n}}|H_0) = 0.05$$

- That is the decision is based on the event $\frac{1}{n}|\sum_{i=1}^{n} x_i| \geq 1.96$ rather than on the observations themselves (conditioning by this value is impossible using frequentist theory).
- The frequency argument is that in 5% of the cases when H_0 is true, it rejects wrongly the null hypothesis.

- The stopping rule principle is definitely incompatible with frequentist modelling.
- Consider $X_i \sim N(\theta, 1)$ and the hypothesis to be tested is $H_0 : \theta = 0$ and we stop collecting data at the first time n such that

$$|\frac{1}{n}\sum_{i=1}^{n} X_i| > \frac{1.96}{\sqrt{n}}.$$

- The resulting sample will always reject $H_0 : \theta = 0$ at the level 5%.

Maximum Likelihood Estimation

- The likelihood principle is fairly vague since it does not lead to the selection of a particular procedure.
- Maximum likelihood estimation is one way to implement the sufficiency and likelihood principles

$$\hat{\theta} = \arg\sup_{\theta} l(\theta|x)$$

- Proof:

$$\arg\sup_{\theta} l(\theta|x) = \arg\sup_{\theta} h(x)g(T(x)|\theta) = \arg\sup_{\theta} g(T(x)|\theta).$$

$$l_1(\theta|x) = c(x)l_2(\theta|x) \Rightarrow \arg\sup l_1(\theta|x) = \arg\sup l_2(\theta|x)$$

- Be careful: Maximum likelihood estimation is just one way to implement the likelihood principle.
- Maximization can be difficult or several equivalent global maxima. However, consistent and efficient in most cases. (asymptotic properties).
- ML estimates can vary widely for small variations of the observations (for small sample sizes).
  Example: Let $X = (X_1, \cdots, X_n)$ be i.i.d. from $U(0, \theta)$ of density $f(x_i|\theta) = \theta^{-1}\mathbf{1}_{[0,\theta]}(x_i)$. Then

$$l(\theta|x) = f(x_1, x_2, \cdots, x_n|\theta) = \Pi_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n}\mathbf{1}_{[\max x_i, \infty)}(\theta)$$

  $\Rightarrow \hat{\theta} = \max(x_i)$.
- Tests require frequentists justifications.

Alternative Approaches

- Many approaches have been proposed: penalized likelihood (e.g. Akaike Information Criterion) or stochastic complexity theory.
- Many of these approaches have a Bayesian flavor.
- A Bayesian model is made of a parametric statistical model $(X, f(x|\theta))$ and a prior distribution on the parameters $(\Theta, \pi(\theta))$.
- The unknown parameters are now considered RANDOM.
- Many statisticians do not like this although they accept the probabilistic modeling on the observations.

Frequentist

- The popular view of probability is the so-called frequentist approach:
- whereby the probability P of an uncertain event A, P(A), is defined by the frequency of that event based on previous observations.
- For example, in the UK 50.9% of all babies born are girls; suppose then that we are interested in the event A: ‘a randomly selected baby is a girl’.
- According to the frequentist approach P(A)=0.509.

Bayesianism

- The frequentist approach for defining the probability of an uncertain event is fine providing that we have been able to record accurate information about many past instances of the event. However, if no such historical database exists, then we have to consider a different approach.
- Bayesian probability is a formalism that allows us to reason about beliefs under conditions of uncertainty. If we have observed that a particular event has happened, such as Britain coming 10th in the medal table at the 2004 Olympics, then there is no uncertainty about it.
- However, suppose **a** is the statement "Britain sweeps the boards at 2016 Brazil Olympics, winning 36 Gold Medals!"
- Since this is a statement about a future event, nobody can state with any certainty whether or not it is true. Different people may have different beliefs in the statement depending on their specific knowledge of factors that might effect its likelihood.

- For example, Henry may have a strong belief in the statement **a** based on his knowledge of the current team and past achievements.
- Marcel, on the other hand, may have a much weaker belief in the statement based on some inside knowledge about the status of British sport; for example, he might know that British sportsmen failed in bids to qualify for the Euro 2008 in soccer, win the Rugby world cup and win the Formula 1 world championship-all in one weekend!
- Thus, in general, a person's subjective belief in a statement **a** will depend on some body of knowledge K. We write this as $P(a|K)$. Henry's belief in **a** is different from Marcel's because they are using different K's. However, even if they were using the same K they might still have different beliefs in **a**.
- The expression $P(\mathbf{a}|K)$ thus represents a belief measure. Sometimes, for simplicity, when K remains constant we just write $P(\mathbf{a})$, but you must be aware that this is a simplification.

Why are we here?

- Fundamental difference in the experimental approach.
- Normally reject (or fail to reject H_0) based on an arbitrarily chosen P value (conventionally$<$0.05) [In other words we choose our willingness to accept a Type I error]
- This tells us nothing about the probability of H_1.
- The frequentist conclusion is restricted to the data at hand, it doesn't take into account previous, valuable information.

Simple Examples

Suppose that we have two bags each containing black and white balls.

- One bag contains three times as many white balls as blacks. The other bag contains three times as many black balls as white.
- Suppose we choose one of these bags at random. For this bag we select five balls at random, replacing each ball after it has been selected. The result is that we find 4 white balls and one black.
- Which bag are the balls from? What is the probability that we were using the bag with mainly white balls?

- For events A and B, the Bayes rule is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} = \frac{P(B|A)P(A)}{P(B)}$$

- Be careful to subtle exchanging of $P(A|B)$ for $P(B|A)$.

- Solution. Let A be the random variable "bag chosen" then $A = \{a_1, a_2\}$ where a_1 represents "bag with mostly white balls" and a_2 represents "bag with mostly black balls". We know that $P(a_1) = P(a_2) = 1/2$ since we choose the bag at random.
- Let B be the event "4 white balls and one black ball chosen from 5 selections". Then we have to calculate $P(a_1|B)$.
- Now, for the bag with mostly white balls the probability of a ball being white is $3/4$ and the probability of a ball being black is $1/4$. Thus, we can use the Binomial Theorem, to compute $P(B|a_1)$ as:

$$P(B|a_1) = C_5^1(1/4)^1(3/4)^4 = 405/1024$$

$$P(B|a_2) = C_5^1(1/4)^4(3/4)^1 = 15/1024$$

- Hence,

$$P(a_1|B) = \frac{405/1024}{405/1024 + 15/1024} = 0.964.$$

Simple Examples

- Prosecutor's Fallacy. A zealous prosecutor has collected an evidence and has an expert testify that the probability of finding this evidence if the accused were innocent is one-in-a-million. The prosecutor concludes that the probability of the accused being innocent is one-in-a-million. This is WRONG.

- Assume no other evidence is available and the population is of 10 million people.
- Defining A = “The accused is guilty” then $P(A) = 10^{-7}$.
- Defining B =“Finding this evidence”, then $P(B|A) = 1 \& P(B|\bar{A}) = 10^{-6}$.
- Bayes formula yields

$$\frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} = \frac{10^{-7}}{10^{-7} + 10^{-6} \times (1 - 10^{-7}} \approx 0.1$$

Simple Examples

- Coming back from a trip, you feel sick and your GP thinks you might have contracted a rare disease (0.01% of the population has the disease).
- A test is available but not perfect.
  If a tested patient has the disease, 100% of the time the test will be positive.
  If a tested patient does not have the diseases, 95% of the the time the test will be negative (5% false positive).
- Your test is positive, should you really care?

- Let A be the event that the patient has the disease and B be the event that the test returns a positive result

$$P(A|B) = \frac{1 \times 0.0001}{1 \times 0.0001 + 0.05 \times 0.9999} \approx 0.002$$

- Such a test would be a complete waste of money for you or the National Health System.

Based on a Bayesian model, we can define:

- The joint distribution of (θ, X)

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

- The marginal distribution of X

$$\pi(x) = \int \pi(\theta) f(x|\theta) d\theta.$$

For a realization $X = x$, $\pi(x)$ is called marginal likelihood or evidence.

Introduction to Bayesian statistics

- Bayes' formula:

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

  It represents all the information on θ than can be extracted from x.
- The prior $\pi(\theta)$ is the probability density function of the parameter and represents what was taught before seeing the data.
- The likelihood $f(x|\theta)$ is the probability of the data given the parameter and represents the data now available.
- The posterior $\pi(\theta|x)$ represents what is taught given the prior information and the data.
- It relates the conditional density of a parameter with its unconditional density.

Ingredients of Bayesian Inference

- Bayesian statistics do satisfy automatically the sufficiency principle, and the likelihood principle.
- **Sufficiency principle**: If $f(x|\theta) = h(x)g(T(x)|\theta)$ then

$$\pi(\theta|x) = \frac{h(x)g(T(x)|\theta)\pi(\theta)}{\int h(x)g(T(x)|\theta)\pi(\theta)d\theta} = \frac{g(T(x)|\theta)\pi(\theta)}{\int g(T(x)|\theta)\pi(\theta)d\theta} = \pi(\theta|T(x))$$

- **Likelihood principle**: Assume we have $f_1(x|\theta) = c(x)f_2(x|\theta)$ then

$$\pi(\theta|x) = \frac{f_1(x|\theta)\pi(\theta)}{\int f_1(x|\theta)\pi(\theta)d\theta} = \frac{c(x)f_2(x|\theta)\pi(\theta)}{\int c(x)f_2(x|\theta)\pi(\theta)d\theta} = \frac{f_2(x|\theta)\pi(\theta)}{\int f_2(x|\theta)\pi(\theta)d\theta}$$

Ingredients of Bayesian Inference

- Explosion of Bayesian statistics over the past 15 years: approximately 30% of papers in top statistical reviews are about Bayesian statistics.
- The Bayesian approach is very well-adapted to many application areas: bioinformatics, genetics, epidemiology, econometrics, machine learning, nuclear magnetic resonance etc.
- It allows one to incorporate in a principled way any prior information available on a given problem.
- Straightforward to handle missing data, outliers, censored data etc.
- Can fit very realistic but complicated models.

- Why have Bayesian statistics enjoyed such an increasing popularity over the last 15 years?
  Implementation difficult and requires computational methods.
- For complex models, Bayesian methods require computing very high dimensional integrals.
- Deterministic methods are inefficient
  Curse of dimensionality.
- Monte Carlo methods are the only possible way to address such problems.
  Standard Monte Carlo methods are inefficient.

- (Bayes, 1764): A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at θ. A second ball O is then rolled n times under the same assumptions and X denotes the number of times the ball O stopped on the left of W. Given X, what inference can we make on θ?

- $X|\theta \sim B(n,\theta)$ binomial distribution and select $\theta \sim U[0,1]$ and

$$P(X = x|\theta) = f(x|\theta) = C_n^x \theta^x (1-\theta)^{n-x}$$

- Using Bayesian rule, we have

$$\pi(\theta|x) = \frac{C_n^x \theta^x (1-\theta)^{n-x}}{\int_0^1 C_n^x \theta^x (1-\theta)^{n-x} d\theta}$$

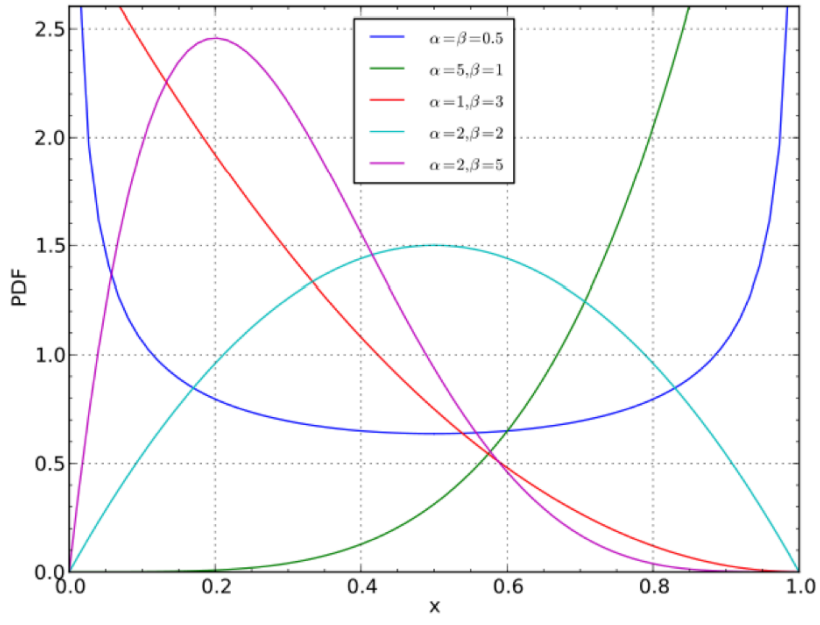
- $$\pi(x) = \int_0^1 P(X = x|\theta)\pi(\theta)d\theta = \frac{1}{n+1}, \text{ for } x = 0, \cdots, n.$$

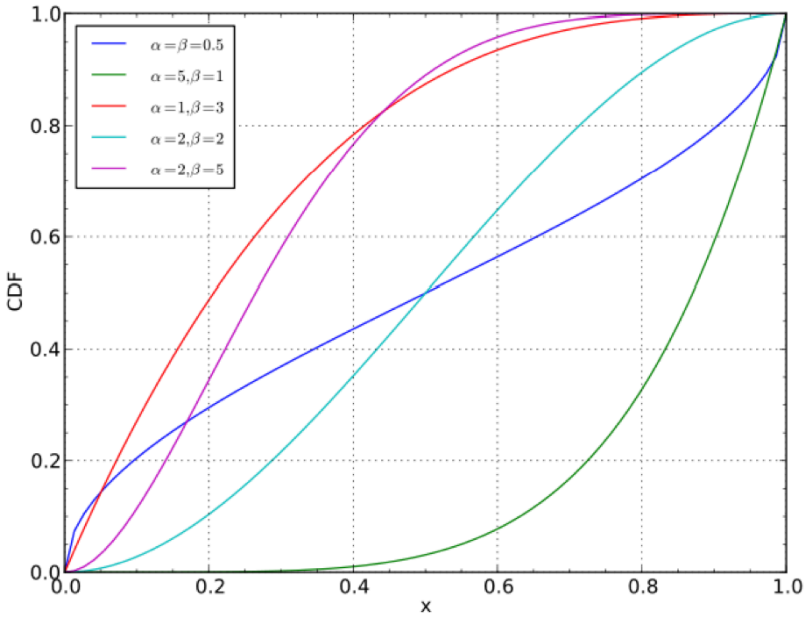
- It follows that
  $\pi(\theta|x) = (n+1)C_n^x \theta^x (1-\theta)^{n-x} \sim Beta(x+1, n+1-x)$.

Beta Distribution

- Uniform distribution is a special case of Beta distribution. Beta(1,1).
- The pdf for Beta distribution with parameters (α, β):

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$
$$= \frac{1}{\mathbf{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$


PDF
x
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
$\alpha=\beta=0.5$
$\alpha=5,\beta=1$
$\alpha=1,\beta=3$
$\alpha=2,\beta=2$
$\alpha=2,\beta=5$


CDF
x
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
$\alpha=\beta=0.5$
$\alpha=5,\beta=1$
$\alpha=1,\beta=3$
$\alpha=2,\beta=2$
$\alpha=2,\beta=5$

$$\int_0^1 f(x)dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} = 1$$

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\Gamma(x+1) = x\Gamma(x) = x! \text{ (if } x \text{ is an integer)}$$

$$\begin{aligned}
\mu = \mathsf{E}[X] &= \int_0^1 x f(x;\alpha,\beta)\, dx \\
&= \int_0^1 x\, \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathbf{B}(\alpha,\beta)}\, dx \\
&= \frac{\alpha}{\alpha+\beta} \\
&= \frac{1}{1+\frac{\beta}{\alpha}}
\end{aligned}$$

- Prediction. Given $X = x$, you roll the ball once more and $P(Y = 1|\theta) = \theta$,then

$$\begin{aligned} P(Y = 1|x) &= \int_0^1 P(Y = 1|\theta, x)\pi(\theta|x)d\theta \\ &= \int_0^1 \theta\pi(\theta|x) = E[\theta|x] = \frac{x+1}{n+2} \end{aligned}$$

- Application. Laplace developed independently such a model. From 1745 to 1770, 241,945 girls and 251,527 boys were born in Paris. Let θ be the probability that any birth is female, then $n = 251,527 + 241,945$.

$$P(\theta \geq 0.5 | x = 241,945) \approx 1.15 \times 10^{-42}.$$

- Remark: This is completely different from a p-value. We do not integrate over observations we have never seen.

Conjugate Prior

- In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

Gaussian Example

- Consider $X_1|\theta \sim N(\theta, \sigma^2)$, $\theta \sim N(m_0, \sigma_0^2)$

$$\begin{aligned}
\pi(\theta|x_1) &\propto f(x_1|\theta)\pi(\theta) \propto \exp(-\frac{(x_1-\theta)^2}{2\sigma^2} - \frac{(\theta-m_0)^2}{2\sigma_0^2}) \\
&\propto \exp(-\frac{\theta^2}{2}(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}) + \theta(\frac{x_1}{\sigma^2}+\frac{m_0}{\sigma^2})) \\
&\propto \exp(-\frac{1}{2\sigma_1^2}(\theta-m_1)^2) \\
&\Rightarrow \theta|x_1 \sim N(m_1, \sigma_1^2)
\end{aligned}$$

with

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \Rightarrow \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}$$

$$m_1 = \sigma_1^2(\frac{x_1}{\sigma^2} + \frac{m_0}{\sigma_0^2})$$

- To predict the distribution of a new observation $X|\theta \sim N(\theta, \sigma^2)$ in light of x_1 we use the predictive distribution

$$f(x|x_1) = \int f(x|\theta)\pi(\theta|x_1)d\theta$$

- Now assume that you observe a realization x_2 of $X_2|\theta \sim N(\theta, \sigma^2)$. Then you are interested now in

$$\begin{aligned}\pi(\theta|x_1, x_2) &\propto f(x_2|\theta) f(x_1|\theta)\pi(\theta) \\ &\propto f(x_2|\theta)\pi(\theta|x_1) \\ &\propto f(x_1|\theta)\pi(\theta|x_2).\end{aligned}$$

- Updating the prior one observation at a time, or all observations together, does not matter.
- The sequential approach can be useful for massive dataset. In this case at time n

$$\pi(\theta|x_1, \cdots, x_n) \propto f(x_n|\theta)\pi(\theta|x_1, \cdots, x_{n-1});$$

i.e. 'the prior at time n is the posterior at time $n-1$'.

Simple Gaussian example: Bayes vs ML

- ML estimate of θ at time n is simply

$$\theta_{ML} = \arg\sup \Pi_{i=1}^{n} f(x_i|\theta) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Posterior of θ at time n is

$$\theta|x_1, \cdots, x_n \sim N(m_n, \sigma_n^2),$$

where

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \Rightarrow \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \sim_{n\to\infty} \frac{\sigma^2}{n}$$

$$m_n = \sigma_n^2\left(\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{m_0}{\sigma_0^2}\right) \sim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Asymptotically in n the prior is washed out by the data and $E[\theta|x_1, \cdots, x_n] = m_n \approx \theta_{ML}$.

Bayes vs ML

- However, keep in mind that information provided by a Bayesian approach is much richer.
- You can compute for example posterior probabilities $Pr(\theta \in A|x_1, \cdots, x_n)$ or $var(\theta|x_1, \cdots, x_n)$
- ML can be reassuring because of consistency and efficiency. For finite sample sizes, do you really care? For time series models for example, there is no such thing.

Multinomial Example

- Assume that we have a variable X taking values on a finite set $\{a_1, \cdots, a_n\}$ and we have a series of independent observations of this distribution, $(x_1, x_2, \cdots, x_m)$ and we want to estimate the value $\theta_i = P(a_i), i = 1, \cdots, n$.
- Let N_i be the number of cases in the sample in which we have obtained the value a_i $(i = 1, \cdots, n)$.
- The MLE of θ_i is $\hat{\theta} = \frac{N_i}{m}$.
- The problems with small samples are completely analogous.

Dirichlet Prior

- We can also follow the Bayesian approach, but the prior distribution is the Dirichlet distribution, a generalization of the Beta distribution for more than 2 cases: $(\theta_1, \cdots, \theta_n)$.
- The expression of $D(\alpha_1, \cdots, \alpha_n)$ is:

$$f(x_1, \cdots, x_{K-1}; \alpha_1, \cdots, \alpha_K) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

$$x_1, \cdots, x_{K-1} > 0$$
$$x_K = 1 - x_1 - \cdots - x_{K-1}$$

- The normalizing constant is the multinomial Beta function, which can be expressed in terms of the gamma function:

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K).$$

- $$E(x_1, x_2, \cdots, x_K) = (\frac{\alpha_1}{\sum_i \alpha_i}, \cdots, \frac{\alpha_K}{\sum_i \alpha_i})$$
- If we have a set of data with counts $(N_1, \cdots, N_n)$, then the posterior distribution is also Dirichlet with parameters

$$D(\alpha_1 + N_1, \cdots, \alpha_n + N_n).$$

- The Bayesian estimation of probabilities are:

$$(\frac{\alpha_1 + N_1}{s + m}, \cdots, \frac{\alpha_n + N_n}{s + m}),$$

where $m = \sum_{i=1}^{n} N_i, s = \sum_{i=1}^{n} \alpha_i$.

- Imagine that we have an urn with balls of different colors: red(R), blue(B) and green(G); but on an unknown quantity.
- Assume that we picked up balls with replacement, with the following sequence: (B,B,R,R,B).
- If we assume a Dirichlet prior distribution with parameters: D(1,1,1), then the estimated frequencies for red,blue and green : (3/8, 4/8, 1/8)
- Observe, as green has a positive probability, even if never appears in the sequence.

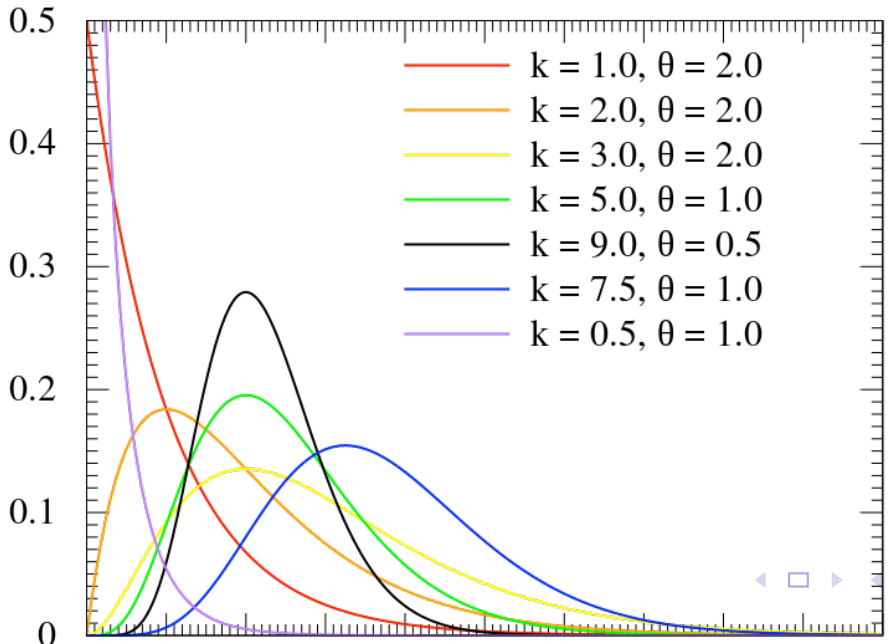
Gamma distribution

$$X \sim \Gamma(\alpha, \beta) \equiv \text{Gamma}(\alpha, \beta)$$

The corresponding probability density function in the shape-rate parametrization is

$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for } x > 0 \text{ and } \alpha, \beta > 0,$$

where $\Gamma(\alpha)$ is a complete gamma function. $\mathbf{E}[X] = \frac{\alpha}{\beta}$



A Simple Poisson Model

- Assume you have some counting observations $X_i \sim P(\theta)$; i.e.

$$f(x_i|\theta) = \frac{\theta^k}{x_i!} e^{-\theta}$$

- Assume we adopt a Gamma prior for θ; i.e. $\theta \sim Gamma(\alpha, \beta)$.

$$\pi(\theta) = Gamma(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$$

- We have

$$\pi(\theta|x_1, \cdots, x_n) = Gamma(\theta; \alpha + \sum_{i=1}^{n} x_i, \beta + n)$$

Testing hypotheses in a Bayesian framework

- Consider the problem where we have $\pi(\theta) = U[0,1]$ and

$$Pr(X = x|\theta) = C_n^x \theta^x (1-\theta)^{n-x},$$

then $\pi(\theta|x) = Beta(x+1, n+1-x)$

- If we want to test $H_0 : \theta \geq 1/2$ vs $H_1 : \theta < 1/2$, then in a Bayesian approach, you can simply compute

$$\pi(H_0|x) = 1 - \pi(H_1|x) = \int_{1/2}^{1} \pi(\theta|x) d\theta$$

- Golden rule of Bayesians: You shall not integrate with respect to observations (except for design...)
  $\Rightarrow$ Contrary to frequentists, your test is never based on observations you donât observe.

Bayes factors

- More generally, one wants to compare two hypothesis: $H_0 : \theta \sim \pi_0$ versus $H_1 : \theta \sim \pi_1$, then the prior is

$$\pi(\theta) = \pi(H_0)\pi_0(\theta) + \pi(H_1)\pi_1(\theta)$$

  where $\pi(H_0) + \pi(H_1) = 1$.
- To compare H_0 versus H_1, we typically compute the Bayes factor which partially eliminated the influence of the prior modelling (i.e. $\pi(H_i)$)

$$\begin{aligned} B^\pi &= \frac{\pi(x|H_1)}{\pi(x|H_0)} = \frac{\int f(x|\theta)\pi_1(\theta)d\theta}{\int f(x|\theta)\pi_0(\theta)d\theta} \\ &= \frac{\pi(H_1|x)\pi(H_0)}{\pi(H_0|x)\pi(H_1)} \end{aligned}$$

- Bayes factors are not limited to the comparison of models with the same parameter space.
- Assume you have some data and two statistical models. Under H_0, $\theta_0 \in \Theta_0$, the prior is $\pi_0(\theta_0)$ and the likelihood is $f_0(x|\theta_0)$, under $H_1, \theta_1 \in \Theta_1$, the prior is $\pi_1(\theta_1)$ and the likelihood is $f_1(x|\theta_1)$, then
$$B^{\pi} = \frac{\pi(x|H_1)}{\pi(x|H_0)} = \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_0(x|\theta_2)\pi_0(\theta_0)d\theta_0}$$
- One can have $\Theta_0 = R$ and $\Theta_1 = R^{1000}$.

Jeffreys' scale of evidence says that

- if $\log_{10}(B^{\pi})$ varies between 0 and 0.5, the evidence against H_0 is poor,
- if it is between 0.5 and 1, it is substantial,
- if it is between 1 and 2, it is strong, and
- if it is above 2, it is decisive.
- Bayes factor tell you where one should prefer H_0 to H_1: it does NOT tell you whether model H_1 any of these models are sensible!

Example: The celebrated coin example

- Assume you have a coin, you toss it 10 times and gets x = 10 heads. Is it biased?
- Let θ be the proba of having an head then we can test $H_0 : \theta = 1/2$
- The p-value $Pr(X \geq 10|H_0) = 2^{-10}$ and the hypothesis is rejected.
- In a Bayesian framework, we test H_0 versus $H_1 : \theta \sim U(1/2, 1]$ using

$$B^{\pi} = \frac{\frac{1}{2}\int_{1/2}^{1} \theta^x (1-\theta)^{10-x} d\theta}{(1/2)^x (1-1/2)^{10-x}} = \frac{1/2 \int_{1/2}^{1} \theta^{10} d\theta}{(1/2)^{10}} \simeq 50$$

Bayesian Model Selection

	*gene*1	*gene*2	$\cdots$	*gene*100	$\cdots$
*Patient*1	2	2	$\cdots$	1	$\cdots$
*Patient*2	1	1	$\cdots$	2	$\cdots$
*Patient*3	1	2	$\cdots$	1	$\cdots$
*Patient*4	3	1	$\cdots$	2	$\cdots$
*Control*1	2	2	$\cdots$	2	$\cdots$
*Control*2	3	1	$\cdots$	1	$\cdots$
*Control*3	1	2	$\cdots$	2	$\cdots$
*Control*4	3	1	$\cdots$	1	$\cdots$

Bayesian Model Selection

More generally,

	*Covariate*1	*Covariate*2	$\cdots$	*Covariate*100	$\cdots$
*A*1	2	2	$\cdots$	1	$\cdots$
*A*2	1	1	$\cdots$	2	$\cdots$
*A*3	1	2	$\cdots$	1	$\cdots$
*A*4	3	1	$\cdots$	2	$\cdots$
*B*1	2	2	$\cdots$	2	$\cdots$
*B*2	3	1	$\cdots$	1	$\cdots$
*B*3	1	2	$\cdots$	2	$\cdots$
*B*4	3	1	$\cdots$	1	$\cdots$

- H_1: A and B come from two independent distributions.
- H_2: A and B are from the same distribution.
- $P(H_1|data) > P(H_2|data)$??

- For covariate 1, suppose for everyone in A group, p_1 for 1, p_2 for 2, p_3 for 3. $p_1 + p_2 + p_3 = 1$.
- Then the probability for the four A's data is

$$P(\text{A}'\text{s data} \mid p_1, p_2, p_3) = (p_1)^2 p_2 p_3$$

- For everyone in B, p_1', p_2', p_3' for 1,2,3. $p_1' + p_2' + p_3' = 1$.
- Then the probability for the four B's data is

$$P(\text{B}'\text{s data} \mid p_1', p_2', p_3') = p_1' p_2' (p_3')^2$$

- $P(H_1|data) = P(data|H_1)P(H_1)/P(data)$
- $P(H_2|data) = P(data|H_2)P(H_2)/P(data)$
- $P(data|H_1) = P(A|H_1)P(B|H_1)$

$$
\begin{aligned}
P(A's\ data,\ p_1, p_2, p_3) &= P(A's\ data | p_1, p_2, p_3) P(p_1, p_2, p_3) \\
&= (p_1)^2 p_2 p_3 D(a_1, a_2, a_3) \\
&= \frac{1}{B(a)} \Pi_{i=1}^{3} p_i^{n_i + a_i - 1}, n_1 = 2, n_2 = 1, n_3 = 1.
\end{aligned}
$$

Integrating p out,

$$
\begin{aligned}
\int_p P(A's\ data,\ p_1, p_2, p_3) dp &= \int_p \frac{1}{B(a)} \Pi_{i=1}^{3} p_i^{n_i + a_i - 1} dp \\
&= \Pi_{i=1}^{3} \frac{\Gamma(n_i + a_i)}{\Gamma(a_i)} \frac{\Gamma(\sum_{i=1}^{3} a_i)}{\Gamma(\sum_{i=1}^{3} (n_i + a_i))}
\end{aligned}
$$

Similarly,

$$\begin{aligned}\int_p P(B's \ data \ p'_1, p'_2, p'_3)dp' &= \int_p' \frac{1}{B(a')}\Pi_{i=1}^3 p_i'^{n'_i+a'_i-1}dp \\ &= \Pi_{i=1}^3 \frac{\Gamma(n'_i+a'_i)}{\Gamma(a'_i)}\frac{\Gamma(\sum_{i=1}^3 a'_i)}{\Gamma(\sum_{i=1}^3(n'_i+a'_i))}\end{aligned}$$

If we let $a_1 = a_2 = a_3 = 1/3, n_1 = 2, n_2 = 1, n_3 = 1$,

$$\begin{aligned} P(A's\ data) &= \Pi_{i=1}^{3} \frac{\Gamma(n_i + a_i)}{\Gamma(a_i)} \frac{\Gamma(\sum_{i=1}^{3} a_i)}{\Gamma(\sum_{i=1}^{3}(n_i + a_i))} \\ &= \frac{\Gamma(2 + 1/3)}{\Gamma(1/3)} \frac{\Gamma(1 + 1/3)}{\Gamma(1/3)} \frac{\Gamma(1 + 1/3)}{\Gamma(1/3)} \frac{\Gamma(1)}{\Gamma(5)} \end{aligned}$$

In R: a=gamma(7/3)*gamma(4/3)^2*gamma(1)/gamma(1/3)^3/gamma(5)
=0.002057613
log(a)=-6.186209

If we let $a'_1 = a'_2 = a'_3 = 1/3, n'_1 = 1, n'_2 = 1, n'_3 = 2$,

$$\begin{aligned} P(B's\ data) &= \Pi_{i=1}^{3} \frac{\Gamma(n'_i + a'_i)}{\Gamma(a'_i)} \frac{\Gamma(\sum_{i=1}^{3} a'_i)}{\Gamma(\sum_{i=1}^{3}(n'_i + a'_i))} \\ &= \frac{\Gamma(1+1/3)}{\Gamma(1/3)} \frac{\Gamma(1+1/3)}{\Gamma(1/3)} \frac{\Gamma(2+1/3)}{\Gamma(1/3)} \frac{\Gamma(1)}{\Gamma(5)} \end{aligned}$$

In R: a=lgamma(4/3)*2+lgamma(7/3)+lgamma(1)-lgamma(1/3)*3-lgamma(5)=-6.186209

$$P(A's\ and\ B's\ data) = P(A's\ data)P(B's\ data) = exp(-6.186209 * 2)$$

This is the marginal likelihood under the hypothesis (H_1) that A's data and B's data come from two independent distribution. i.e. $p! = p'$.
So

$$\begin{aligned} P(A's\ and\ B's\ data|H_1) &= P(A's\ data|H_1)P(B's\ data|H_1) \\ &= exp(-6.186209 * 2) \end{aligned}$$

The second hypothesis (H_2) is that A's data and B's data come from the same distributions.

$$P(A's \text{ and } B's \text{ data}|H_2) = \int_p P(\text{pooled data}, \ p_1, p_2, p_3)dp$$

$$= \int_p \frac{1}{B(a)} \Pi_{i=1}^3 p_i^{n_i+a_i-1} dp$$

$$= \Pi_{i=1}^3 \frac{\Gamma(n_i + a_i)}{\Gamma(a_i)} \frac{\Gamma(\sum_{i=1}^3 a_i)}{\Gamma(\sum_{i=1}^3 (n_i + a_i))}, n_1 = 3, n_2 = 2, n_3 = 3.$$

If we let $a_1 = a_2 = a_3 = 1/3, n_1 = 3, n_2 = 2, n_3 = 3$,

$$\begin{aligned} P(\textit{pooled data}) &= \Pi_{i=1}^{3} \frac{\Gamma(n_i + a_i)}{\Gamma(a_i)} \frac{\Gamma(\sum_{i=1}^{3} a_i)}{\Gamma(\sum_{i=1}^{3}(n_i + a_i))} \\ &= \frac{\Gamma(3 + 1/3)}{\Gamma(1/3)} \frac{\Gamma(2 + 1/3)}{\Gamma(1/3)} \frac{\Gamma(3 + 1/3)}{\Gamma(1/3)} \frac{\Gamma(1)}{\Gamma(9)} \end{aligned}$$

In R: a=lgamma(10/3)*2+lgamma(7/3)+lgamma(1)-lgamma(1/3)*3-lgamma(9)=-11.3428.

$P(A's\ and\ B's\ data|H_1) = exp(-6.186209 * 2)$
$P(A's\ and\ B's\ data|H_2) = exp(-11.3428)$
Prior: $P(H_1) = P(H_2) = 0.5$
Posterior: $P(H_1|data) \propto \exp(-6.186209 * 2)$
$P(H_2|data) \propto \exp(-11.3428)$
$P(H_1|data) + P(H_2|data) = 1.$

Bayesian Model Selection

	*Covariate*1	*Covariate2*	...	*Covariate*100	...
*A*1	2	2	...	1	...
*A*2	1	1	...	2	...
*A*3	1	2	...	1	...
*A*4	3	1	...	2	...
*B*1	2	2	...	2	...
*B*2	3	1	...	1	...
*B*3	1	2	...	2	...
*B*4	3	1	...	1	...

Define indicators $I = (I_1, I_2, \cdots, I_{100})$.
I_i means covariate i from H_1, $I_i = 2$ means covariate i from H_2.

$$P(data|I) = \Pi_{i=1}^{N_{covariate}} P(covariate_i|I_i)$$

$$P(covariate_i|I_i = 1) = P(covariate_i|H_1)$$

$$P(covariate_i|I_i = 2) = P(covariate_i|H_2)$$

$$P(I) = \Pi_{i=1}^{N_{covariate}} P(I_i)$$

$$P(I|data) \propto P(I)P(data|I) = \Pi_{i=1}^{N_{covariate}} P(I_i)P(covariate_i|I_i)$$

But are covariates independent?

- H_3: Covariate 2 and Covariate 100 are dependent, but A's and B's are from different distributions.
- H_4: Covariate 2 and Covariate 100 are dependent, but A's and B's are from the same distributions.

- Under H_3, for covariate 2 and covariate 100, suppose for everyone in A, p_1, p_2, p_3, p_4 for (1,1),(1,2), (2,1),(2,2), $p_1 + p_2 + p_3 + p_4 = 1$.
- The probability for A's data is

$$P(A's\ data|p_1, p_2, p_3, p_4) = (p_2)^2(p_3)^2$$

- For everyone in B, p'_1, p'_2, p'_3, p'_4 for (1,1),(1,2), (2,1),(2,2), $p'_1 + p'_2 + p'_3 + p'_4 = 1$.
- The probability for B's data is

$$P(B's\ data|p_1, p_2, p_3, p_4) = (p'_1)^2(p'_4)^2$$

We use prior on p and p'. $p \sim D(a_1, a_2, a_3, a_4)$,$p' \sim D(a'_1, a'_2, a'_3, a'_4)$.

$$\begin{aligned}\int_p P(A's \text{ data } p_1, p_2, p_3, p_4)dp &= \int_p \frac{1}{B(a)} \Pi_{i=1}^4 p_i^{n_i+a_i-1} dp \\ &= \Pi_{i=1}^4 \frac{\Gamma(n_i + a_i)}{\Gamma(a_i)} \frac{\Gamma(\sum_{i=1}^4 a_i)}{\Gamma(\sum_{i=1}^4 (n_i + a_i))}, \\ & n_1 = 0, n_2 = 2, n_3 = 2, n_4 = 0\end{aligned}$$

Similarly, we got the expression for B's data.

- $a_1 = a_2 = a_3 = a_4 = 1/4, n_1 = 0, n_2 = 2, n_3 = 2, n_4 = 0$
  $a'_1 = a'_2 = a'_3 = a'_4 = 1/4, n'_1 = 2, n'_2 = 0, n'_3 = 0, n'_4 = 2$
- For H_3,
  a=lgamma(1/4)+lgamma(2+1/4)+lgamma(2+1/4)+lgamma(1/4)
  +lgamma(1)-lgamma(1/4)*4-lgamma(5)=-5.504355.
  b=lgamma(2+1/4)*2+lgamma(1/4)+lgamma(0+1/4)+lgamma(2+1/4)
  +lgamma(1)-lgamma(1/4)*4-lgamma(5)=-5.504355.
  $P(data|H_3) = exp(-2 * 5.504355) = exp(-11.00871)$.
- For H_4, a=lgamma(2+1/4)+lgamma(2+1/4)+lgamma(2+1/4)
  +lgamma(1)-lgamma(1/4)*4-lgamma(9)=-15.25721.
  $P(data|H_4) = exp(-15.25721)$.

$P(data|H_1) = exp(-7.506836 * 2) = exp(-15.01367)$
$P(data|H_2) = exp(-6.84186 * 2) = exp(-13.68372)$
$P(data|H_3) = exp(-11.00871)$
$P(data|H_4) = exp(-15.25721)$
$P(H_1) = P(H_2) = P(H_3) = P(H_4)$.
$P(H_3|data)$ is the largest!

Total number of models

- $N_{covariate} = 100$, the number of possible I is 4^{100}.
- The total number of models is 4^{100}.
- Sometimes, the number of covariates could be larger!!!

Implementations problems for Bayesian inference

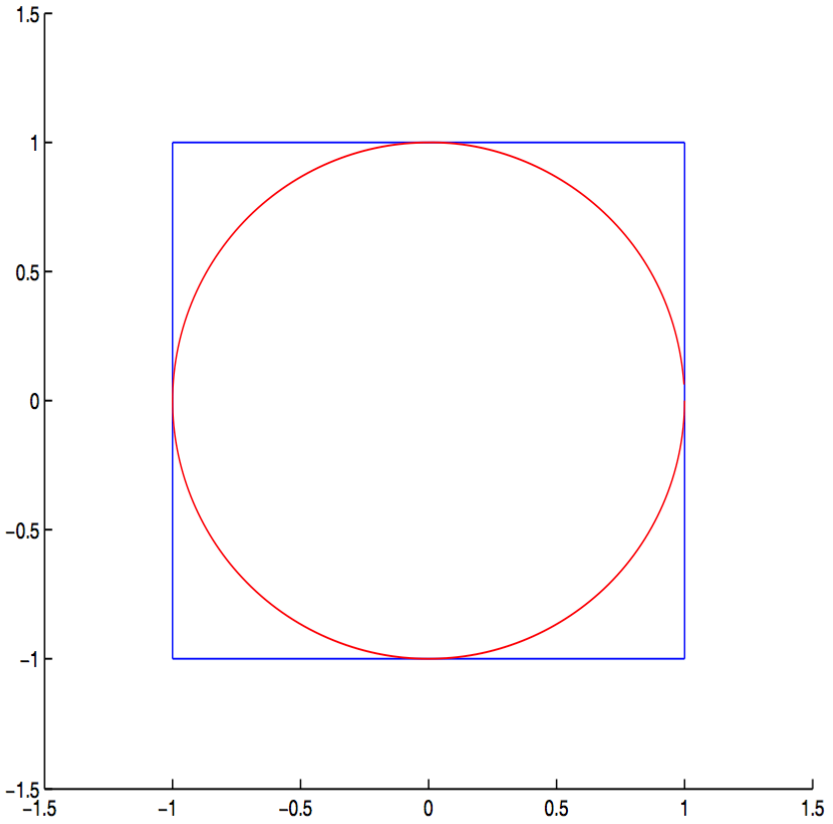
- Bayesian inference is conceptually simple (once the model is set) but how do you perform Bayesian inference for complex models??? It requires computing high dimensional integrals.
- In practice, Bayesian inference is not only used to determine whether coins are biased and for Gaussian models.
- Monte Carlo methods have appeared in the 90's in statistics and have truly revolutionized the whole field.

Monte Carlo Markov Chains

- MCMC method are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution.
- The state of the chain after a large number of steps is then used as a sample from the desired distribution.

Introduction to Monte Carlo: A simple example

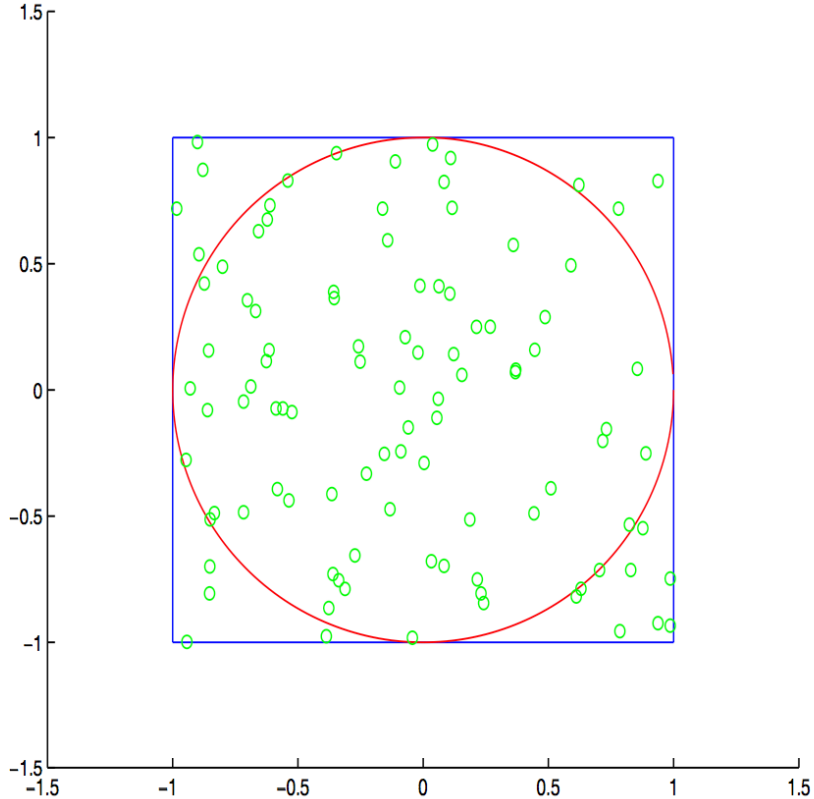
Consider the 2×2 square, say $S \in R^2$, with inscribed disc A of radius 1.



- An “idealised” rain falls uniformly on the square S, i.e. the probability for a drop to fall in a region A is proportional to the area of A.
- Let D be the random variable defined on $\Theta = S$ representing the location of a drop and A a region of the square, then

$$P(D \in A) = \frac{\int_A dxdy}{\int_S dxdy},$$

  where x and y are the Cartesian coordinates.
- Assume we observe N such independent drops, say $\{D_i, i = 1, \cdots, N\}$.


1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5

- Intuitively, imagining that you have never followed any statistics course, a sensible technique to estimate the probability $P(D \in A)$ of falling in a given region $A \subset S$ (and think for example of $A = D$) would consist of using

$$P(D \in A) \simeq \frac{\text{number of drops that fell in A}}{N}.$$

- We want a statistical justification to it.

Probability of this event as an expectation

- Let us denote the indicator function of a set A as follows,

$$\mathcal{I}_A(x,y) = \begin{cases} 1 & \text{if point } d=(x,y) \in A, \\ 0 & \text{otherwise.} \end{cases}$$

- We have

$$P(D \in A) = \frac{\int_S \mathcal{I}_A(x,y)dxdy}{\int_S dxdy} = \frac{\int_S \mathcal{I}_A(x,y)dxdy}{4} = \int_S \frac{1}{4}\mathcal{I}_A(x,y)dxdy$$

- Since

$$\begin{aligned} \int_{S=A\cup S\setminus A} \mathcal{I}_A(x,y)dxdy &= \int_A \mathcal{I}_A(x,y)dxdy + \int_{S\setminus A} \mathcal{I}_A(x,y)dxdy \\ &= \int_A 1dxdy + \int_{S\setminus A} 0dxdy \end{aligned}$$

- 1/4 is the probability density associated to P, i.e. the density of the uniform distribution on S denoted by U_S.
- Let us define the r.v. $V(D) := \mathcal{I}_A(D) := \mathcal{I}_A(X, Y)$,where X, Y are the rvs representing the Cartesian coordinates of a uniformly distributed point on S, denoted by $U_S(D \sim U_S)$, where a drop falls. With this notation, we understand that

$$P(D \in A) = \int_S \frac{1}{4} \mathcal{I}_A(x, y) dxdy = E_{U_S}(V).$$

- Introduce $\{V_i := V(D_i), i = 1, \cdots, N\}$ the r.v.s associated to the drops $\{D_i, i = 1, \cdots, N\}$ and consider the sum:

$$S_N = \frac{\sum_{i=1}^{N} V_i}{N} = \frac{\text{number of drops that fell in } A}{N}$$

- This expression shows that our suggested approximation of $P(D \in A)$ is the empirical average of i.i.d. r.v.s $\{V_i, i = 1, \cdots, N\}$.
- Assuming that the rain lasts forever (i.e. $N \to +\infty$) then the law of large numbers (since $E_{U_s}(|V|) < +\infty$ here) yields

$$\lim_{N \to \infty} S_N = E_{U_s}, \text{ almost surely}$$

  where we have already proved that $P(D \in A) = E_{U_s}(V)$.
- When N is sufficiently large, this mathematically justifies our intuitive method.

- As we have $P(D \in A) = \int_A 1/4 dxdy = \pi/4$, then S_N is an (unbiased) estimator of $\pi/4$.
- It is a r.v.,i.e. $S_N = \pi/4 + E_N$, where E_N is an error term.
- To characterise the precision of our estimator, we can use

$$var(E_N) = var(S_N) = \frac{1}{N^2} \sum_{i=1}^{N} var(V_i) = \frac{1}{N} var(V_1),$$

  as the V_i are independent.
- This means that

$$\sqrt{var(S_N)} = \sqrt{E(S_N - E(S_N))^2} = \sqrt{E[(S_N - P(D \in A))^2]},$$

  which implies that the mean square error between S_N and $P(D \in A)$ decreases as $1/\sqrt{N}$.

Generalization

- Consider the case where $\Theta = R^{n_\theta}$ for any n_θ, and in particular $n_\theta >> 1$. Replace the S and A above with a hypercube S_{n_x} and an inscribed hyperball A_{n_θ} in Θ.
- If we could observe a hyperrain, the same estimator could be built; the only thing we need to calculate $\mathcal{I}_{A^{n_\theta}}(D)$ pointwise. Arguments that lead earlier to the formal validation of the Monte Carlo approach remain identical here.
- In particular the rate of convergence of the estimator in the mean square sense is again in $\sqrt{1/N}$ and independent of the dimension n_x.
- Monte Carlo methods are extremely attractive when n_x is large.

Generalization

- Assume $N >> 1$ i.i.d. samples $\theta(i) \sim \pi(i = 1, \cdots, N)$ are available to us (since it is unlikely that rain can generate samples from any distribution π, we will address the problem of sample generation later).
- Now consider any set $A \subset \Theta$ and assume that we are interested in $\pi(A) = P(\theta \in A)$ for $\theta \sim \pi$. We naturally choose the following estimator

$$\pi(A) \simeq \frac{\text{number of samples in } A}{\text{total number of samples}},$$

which by the law of large numbers is a consistent estimator of $\pi(A)$ since

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathcal{I}_A(\theta^{(i)}) = E_\pi(\mathcal{I}_A(\theta)) = \pi(A)$$

- Now we generalise this idea to tackle the generic problem of estimating

$$E_\pi(f(\theta)) = \int_\Theta f(\theta)\pi(\theta)d\theta,$$

  where $f : \Theta \to R^{n_f}$ and π is a probability distribution on $\Theta \subset R^{n_x}$.
- We will assume that $E_\pi(|f(\theta)|) < +\infty$ but that it is difficult to obtain an analytical expression for $E_\pi(f(\theta))$.
- Here π is any probability distribution and not necessary the prior.

Generalization

- A way of generalising this in order to evaluate $E_\pi(f(\theta))$ consists of considering the unbiased estimator:

$$S_N(f) = \frac{1}{N}\sum_i f(\theta^{(i)}),$$

- From the law of large numbers $S_N(f)$ will converge and

$$\lim_{N\to\infty} \frac{1}{N}\sum_i f(\theta^{(i)}) = E_\pi f(\theta), a.s.$$

- A good measure of the approximation is the variance of $S_N(f)$,

$$var_\pi[S(f)] = var_\pi \frac{1}{N}\sum_i f(\theta^{(i)}) = \frac{var(f(\theta^{(i)}))}{N}$$

Now the central limit theorem applies if $var_\pi[f(\theta)] < \infty$ and tells us that

$$\sqrt{N}(S_N(f) - E_\pi(f(\theta)) \to_d N(0, var_\pi(f(\theta)), as N \to \infty$$

Generalization

The conclusions drawn in the rain example are still valid here.

- The rate of convergence is immune to the dimension of Θ.
- It is easy to take complex integration domains into account.
- It is easily implementable and general. The requirements are
  to be able to evaluate $f(\theta)$ for any $\theta \in \Theta$,
  to be able to produce samples distributed according to π.

- Let us introduce the Dirac-delta function δ_{θ_0} for $\theta_0 \in \Theta$ defined for any $f : \Theta \to R^{n_f}$ as follows:

$$\int_\Theta f(\theta)\delta_{\theta_0}(\theta)d\theta = f(\theta_0)$$

- Note that this implies in particular that for $A \subset \Theta$,

$$\int_\Theta \mathcal{I}_A(\theta)\delta_{\theta_0}(\theta)d\theta = \int_A \delta_{\theta_0}(\theta)d\theta = \mathcal{I}_A(\theta_0)$$

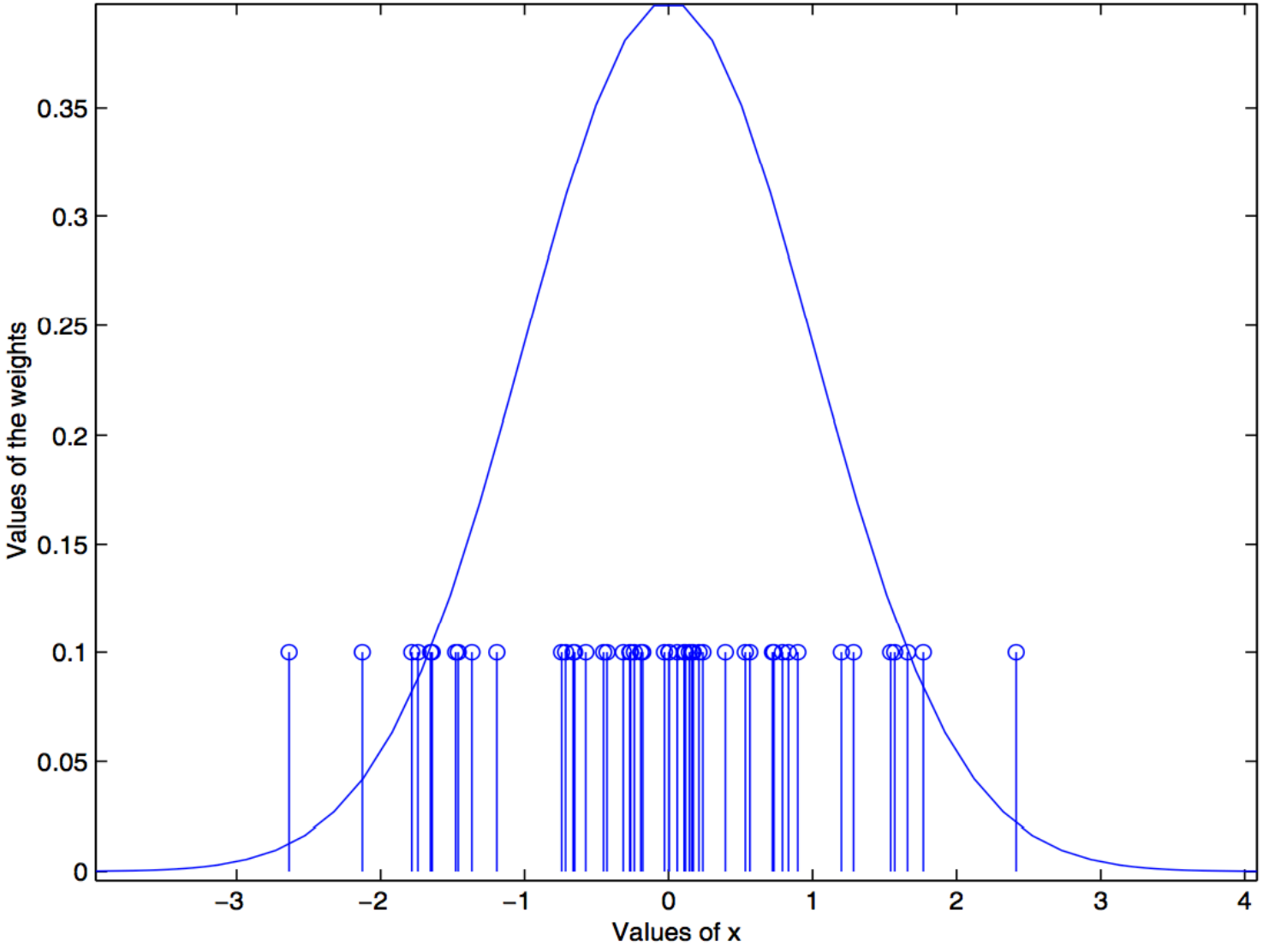
- Now, for $\theta^{(i)} \sim \pi, i = 1, 2 \cdots, N$, we can introduce the following mixture of Dirac-delta functions

$$\hat{\pi}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta^{(i)}}(\theta),$$

which is the empirical measure, and consider for any $A \subset \Theta$

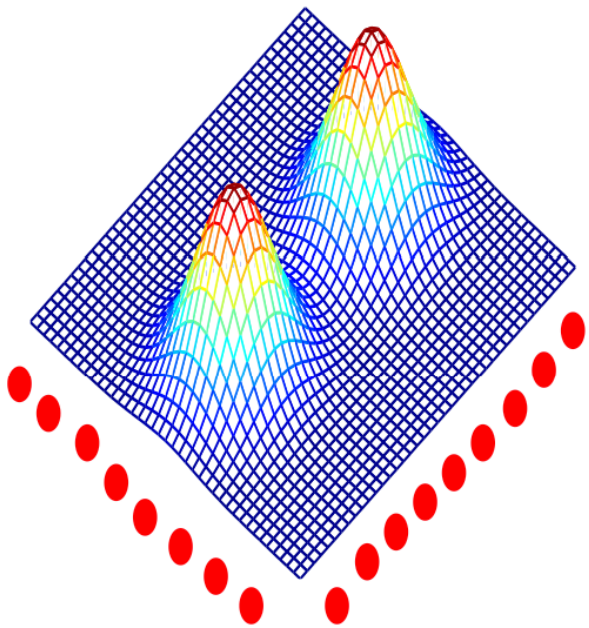
$$\hat{\pi}_N(A) = \int_A \hat{\pi}_N(\theta)d\theta = \sum_{i=1}^{N}\int_A \frac{1}{N}\delta_{\theta^{(i)}}(\theta)d\theta = \frac{1}{N}\sum_{i=1}^{N}\mathcal{I}_A(\theta) = S_N(A)$$

- The concentration of points in a given region of the space represents π.
- This approach is in contrast with what is usually done in parametric statistics, i.e. start with samples and then introduce a distribution with an algebraic representation for the underlying population.
- Note that here each sample $\theta^{(i)}$ has a weight of $\frac{1}{N}$, but that it is also possible to consider weighted sample representations of π.

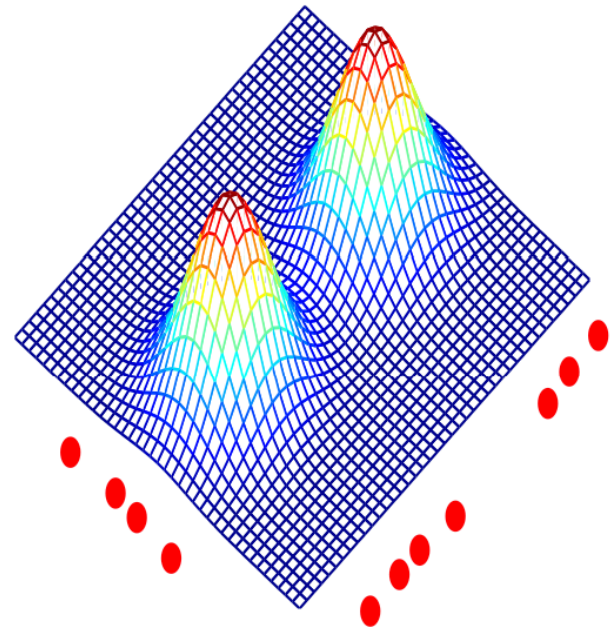


Sample representation of a Gaussian distribution

Deterministic Integration



Monte Carlo Integration



- Now consider the problem of estimating $E_\pi(f)$. We simply replace π with its sample representation $\hat{\pi}_N$ and obtain

$$E_\pi(f) \simeq \int_\Theta f(\theta) \sum_{i=1}^{N} \frac{1}{N} \delta_{\theta^{(i)}}(\theta) d\theta$$

$$= \sum_{i=1}^{N} \int_\Theta f(\theta) \frac{1}{N} \delta_{\theta^{(i)}}(\theta) d\theta = \frac{1}{N} \sum_{i=1}^{N} f(\theta^{(i)})$$

  which is precisely $S_N(f)$, the Monte Carlo estimator suggested earlier.
- Clearly based on $\hat{\pi}_N$, we can easily estimate $E_\pi(f)$ for any f.
- For example

$$var_\pi(f) = E_\pi(f^2) - E_\pi^2(f) \simeq \frac{1}{N} \sum_{i=1}^{N} f^2(\theta^{(i)}) - (\frac{1}{N} \sum_{i=1}^{N} f(\theta^{(i)}))^2$$

- If you could sample easily from an arbitrary probability distribution, then you could easily estimate all the quantities you are interested in.
- Problem: How do you sample from an arbitrary probability distribution???

- Rejection Sampling and Inverse transform are two general methods but limited to problems of moderate dimensions.
- “Problem”: We try to sample all the components of a potentially high-dimensional parameter simultaneously.
- There are two ways to implement incremental strategies.
  - Iteratively: Markov chain Monte Carlo.
  - Sequentially: Sequential Monte Carlo.

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

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

  and we will write

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

- Given a target π, design a transition kernel P such that asymptotically as $n \to \infty$,

$$\frac{1}{N} \sum_{i=1}^{N} \psi(X_n) \to \int \psi(x)\pi(x)dx \quad \text{and/or } X_n \sim \pi$$

- It should be easy to simulate the Markov chain even if π is complex.

Example

- Consider the autoregression for $|\alpha| < 1$.

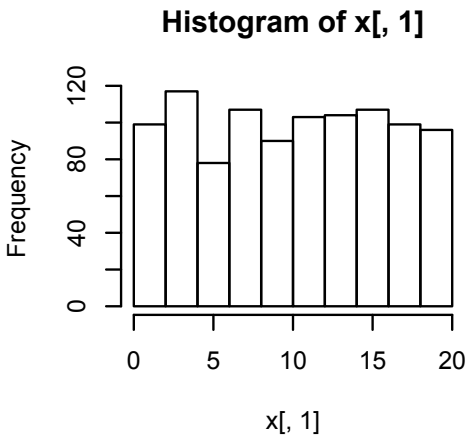
$$X_n = \alpha X_{n-1} + V_n, \text{ where } V_n \sim N(0, \sigma^2).$$

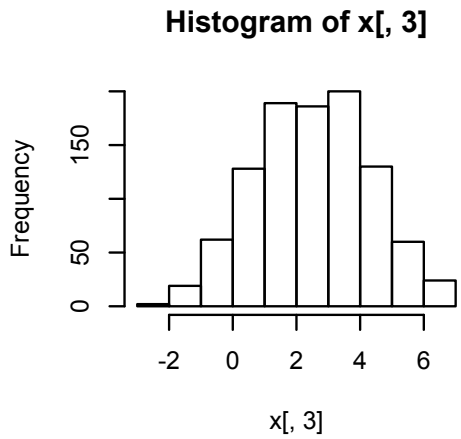
- The limiting distribution is

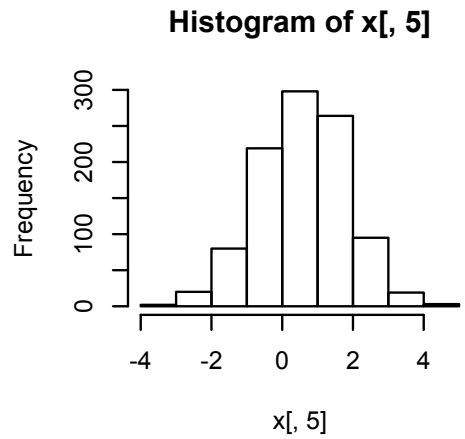
$$\pi(x) = N(0, \frac{\sigma^2}{1 - \alpha^2}).$$

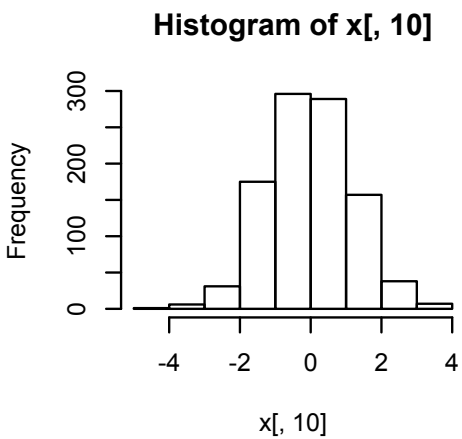
- To sample from π, we could just sample the Markov chain and asymptotically we would have $Xn \sim \pi$.
- Obviously, in this case this is useless because we can sample from π directly.

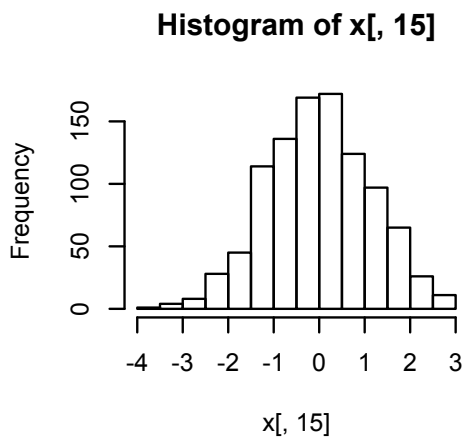
- Graphically, consider 1000 independent Markov chains run in parallel.
- We assume that the initial distribution of these Markov chains is U[0,20]. So initially, the Markov chains samples are not distributed according to π.

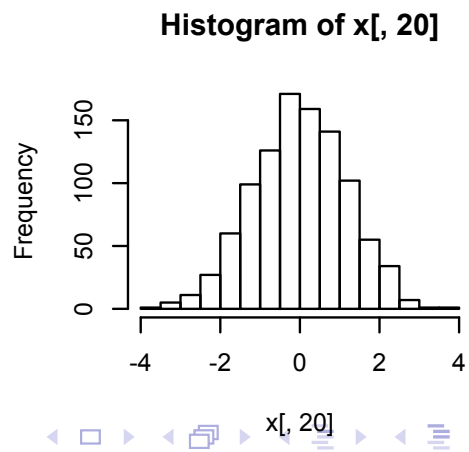




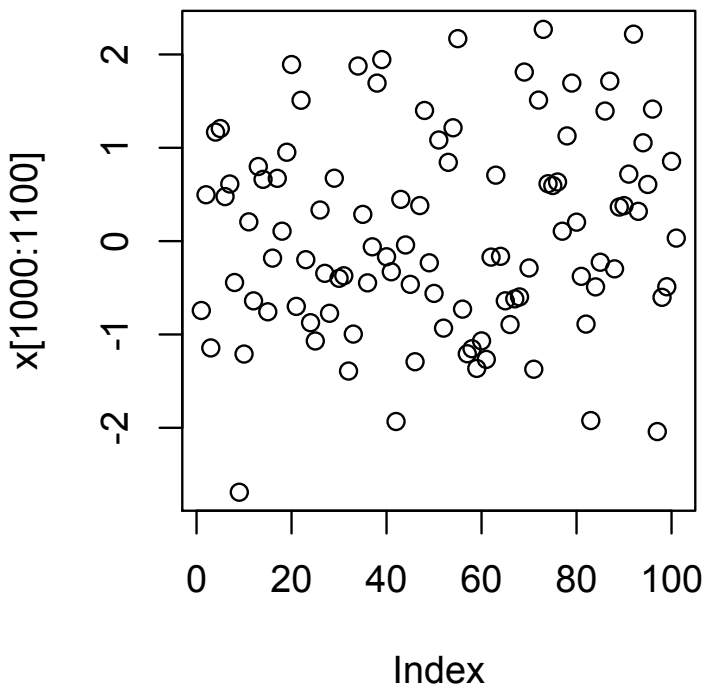


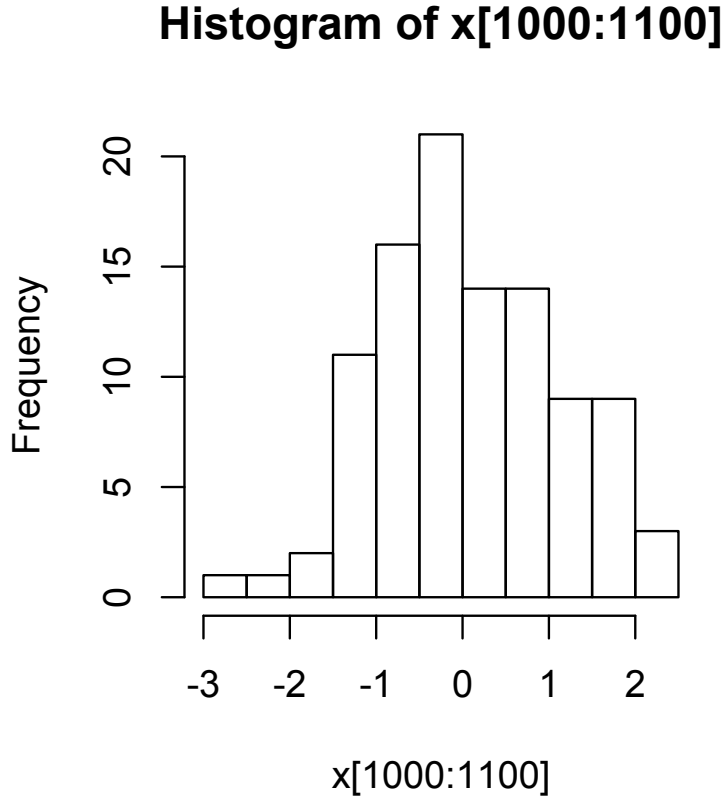






- The target normal distribution seems to “attract” the distribution of the samples and even to be a fixed point of the algorithm.
- This is is what we wanted to achieve, i.e. it seems that we have produced 1000 independent samples from the normal distribution.
- In fact one can show that in many (all?) situations of interest it is not necessary to run N Markov chains in parallel in order to obtain 1000 samples, but that one can consider a unique Markov chain, and build the histogram from this single Markov chain by forming histograms from one trajectory.


x[1000:1100]
2
1
0
-1
-2
0
20
40
60
80
100
Index


Histogram of x[1000:1100]
Frequency
0
5
10
15
20
-3
-2
-1
0
1
2
x[1000:1100]

MCMC

- The estimate of the target distribution, through the series of histograms, improves with the number of iterations.
- Assume that we have stored $\{X_n, 1 \leq n \leq N\}$ for N large and wish to estimate $\int \psi(x)\pi(x)dx$.
- In the light of the numerical experiments, one can suggest the estimator $\frac{1}{N}\sum_{n=1}^{N} \psi(X_n)$, which is exactly the estimator that we would use if $\{X_n, 1 \leq n \leq N\}$ were independent.
- In fact, it can be proved, under relatively mild conditions, that such an estimator is consistent despite the fact that the samples are NOT independent! Under additional conditions, a CLT also holds with a rate of CV in $\sqrt{1/N}$.

To summarize, we are interested in Markov chains with transition kernel P which have the following three important properties observed above:

- The desired distribution π is a "fixed point" of the algorithm or, in more appropriate terms, an invariant distribution of the Markov chain, i.e. $\int_X \pi(x)P(x,y)dx = \pi(y)$
- The successive distributions of the Markov chains are "attracted" by π, or converge towards π.
- The estimator $\frac{1}{N}\sum_{n=1}^{N} \psi(X_n)$ converges towards $E_\pi(\psi(X))$ and asymptotically $X_n \sim \pi$.

Applications of MCMC

- Simulation:

$$(x, y) \sim f(x, y) = cC_n^x y^{x+\alpha-1}(1-y)^{n-x+\beta-1},$$

  where $x = 0, 1, 2, \cdots, n, 0 \leq y \leq 1, \alpha, \beta$ are known.
- Integration: computing in high dimensions.
- Bayesian Inference: Posterior distributions, posterior means...

- Given $\pi(x)$, there is an infinite number of kernels $P(x, y)$ which admits $\pi(x)$ 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 towards π but very difficult to obtain rates of convergence.

The Gibbs Sampler

- Consider the target distribution $\pi(\theta)$ such that $\theta = (\theta^1, \theta^2)$. Then the 2 component Gibbs sampler proceeds as follows.
  Initialization:
  Select deterministically or randomly $\theta_0 = (\theta_0^1, \theta_0^2)$.
  Iteration i, $i \geq 1$:
  Sample $\theta_i^1 \sim \pi(\theta^1 | \theta_{i-1}^2)$;
  Sample $\theta_i^2 \sim \pi(\theta^2 | \theta_{i-1}^1)$;
- Sampling from these conditional is often feasible even when sampling from the joint is impossible.

- Clearly $\{(\theta_i^1, \theta_i^2)\}$ is a Markov chain and its transition kernel is

$$P((\theta^1, \theta^2), (\tilde{\theta}^1, \tilde{\theta}^2)) = \pi(\tilde{\theta}_1|\theta_2)\pi(\tilde{\theta}_2|\tilde{\theta}_1).$$

- Then $\int\int \pi(\theta_1, \theta_2)P((\theta_i^1, \theta_i^2), (\tilde{\theta}_i^1, \tilde{\theta}_i^2))d\theta^1 d\theta^2$ satisfies

$$\begin{aligned}
&\int\int \pi(\theta_1, \theta_2)\pi(\tilde{\theta}_1|\theta_2)\pi(\tilde{\theta}_2|\tilde{\theta}_1)d\theta^1 d\theta^2 \\
&= \int \pi(\theta_2)\pi(\tilde{\theta}_1|\theta_2)\pi(\tilde{\theta}_2|\tilde{\theta}_1)d\theta^2 \\
&= \int \pi(\tilde{\theta}_1, \theta_2)\pi(\tilde{\theta}_2|\tilde{\theta}_1)d\theta^2 \\
&= \pi(\tilde{\theta}_1)\pi(\tilde{\theta}_2|\tilde{\theta}_1) = \pi(\tilde{\theta}^1, \tilde{\theta}^2)
\end{aligned}$$

Irreducibility

- This does not ensure that the Gibbs sampler does converge towards the invariant distribution!
- Additionally it is required to ensure irreducibility: loosely speaking the Markov chain can move to any set A such that $\pi(A) > 0$ for (almost) any starting point.
- This ensures that

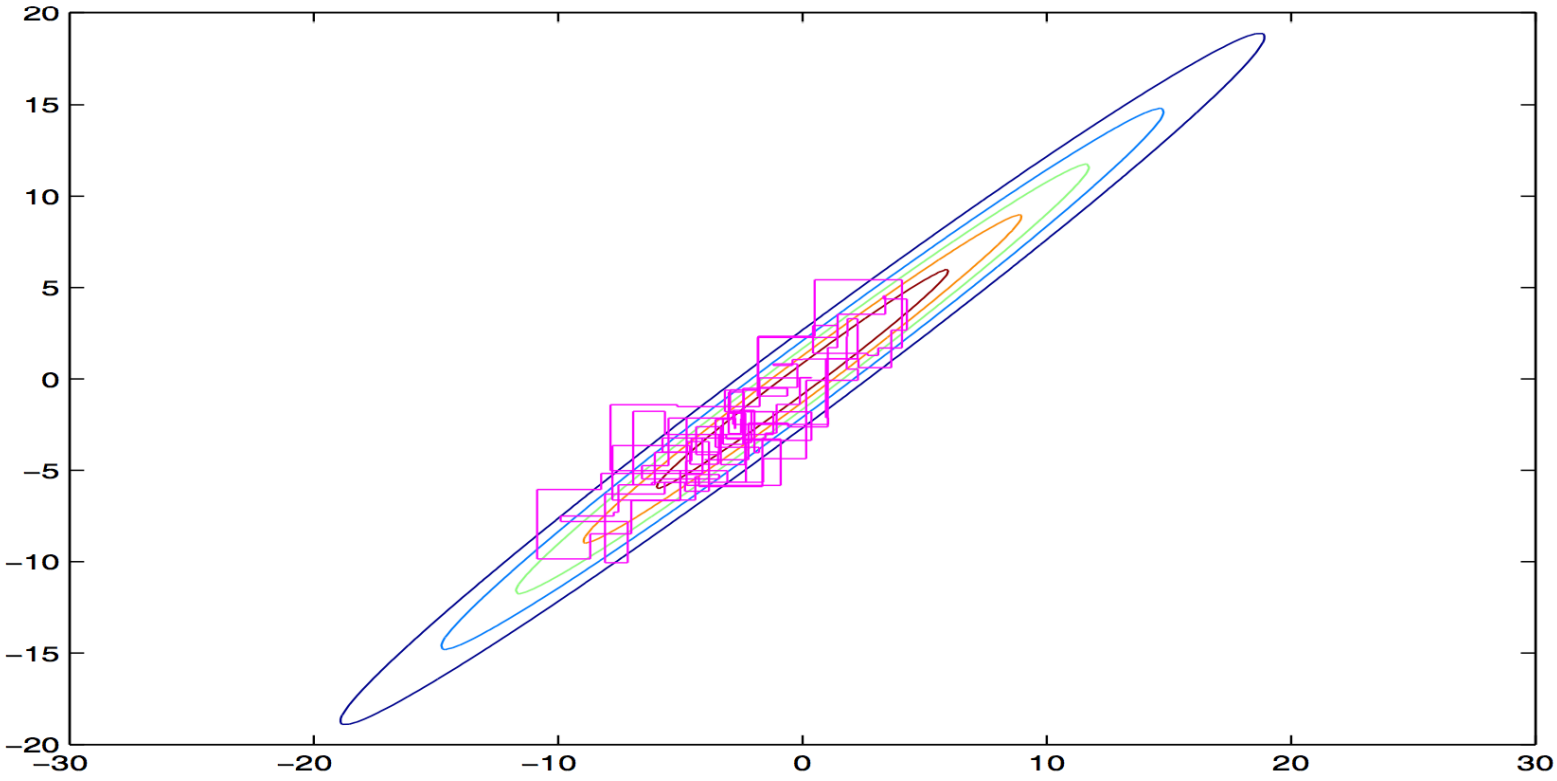
$$\frac{1}{N}\sum_{n=1}^{N}\psi(\theta_n^1,\theta_n^2) \to \int \psi(\theta^1,\theta^2)\pi(\theta^1,\theta^2)d\theta^1 d\theta^2$$

but NOT that asymptotically $(\theta_n^1,\theta_n^2) \sim \pi$.

Aperiodicity

- Consider a simple example where $X = \{1, 2\}$ and $P(1, 2) = P(2, 1) = 1$. Clearly the invariant distribution is given by $\pi(1) = \pi(2) = 0.5$.
- However, we know that if the chain starts in $X_0 = 1$, then $X_{2n} = 1$ and $X_{2n+1} = 0$ for any n.
- We have $\frac{1}{N}\psi(X_n) \to \int \psi(x)\pi(x)dx$, but clearly X_n is NOT distributed according to π.
- You need to make sure that you do NOT explore the space in a periodic way to ensure that $X_n \sim \pi$ asymptotically.

Even when irreducibility and aperiodicity are ensured, the Gibbs sampler can still converge very slowly.



- If $\theta = (\theta_1, \cdots, \theta_p)$ where $p > 2$, the Gibbs sampling strategy still applies.
- Initialization:
  Select deterministically or randomly $\theta_0 = (\theta_0^1, \cdots, \theta_0^p)$.
  Iteration i, $i \geq 1$:
  For $k = 1 : p$
  Sample $\theta_i^k \sim \pi(\theta^k | \theta_i^{-k})$;
  where $\theta^{-k} = (\theta_i^1, \cdots, \theta_i^{k-1}, \theta_{i-1}^{k+1}, \cdots, \theta_{i-1}^p)$.

Random Scan Gibbs sampler

- Initialization:
  Select deterministically or randomly $\theta_0 = (\theta_0^1, \cdots, \theta_0^p)$.
  Iteration i, $i \geq 1$:
  Sample $K \sim U\{1, \cdots, p\}$.
  Set $\theta_i^{-K} = \theta_{i-1}^{-K}$
  Sample $\theta_i^K \sim \pi(\theta^K | \theta_i^{-K})$;
  where $\theta^{-K} = (\theta_i^1, \cdots, \theta_i^{K-1}, \theta_i^{K+1}, \cdots, \theta_i^p)$.

Practical Recommendations

- Try to have as few “blocks” as possible.
- Put the most correlated variables in the same block.
- If necessary, reparametrize 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.

Example 1

- Generate the random variable $(x, y) \sim f(x, y) = (1/28)(2x + 3y + 2), 0 < x < 2, 0 < y < 2$.
- The conditional distribution for x was:

$$f(x|y) = \frac{f(x, y)}{f(y)} = \frac{2x + 3y + 2}{6y + 8}$$

- The conditional distribution for y was:

$$f(y|x) = \frac{f(x, y)}{f(x)} = \frac{2x + 3y + 2}{4x + 10}$$

- Thus, a Gibbs sampler for sampling x and y in this problem would follow these steps:
  1. Set $j = 0$ and establish starting values. Here, let's set $x^0 = -5$ and $y^0 = -5$.
  2. Sample x^{j+1} from $f(x|y = y^j)$.
  3. Sample y^{j+1} from $f(y|x = x^{j+1})$.
  4. Increment $j = j + 1$ and return to step 2 until $j = 2000$.

- We use the inverse tranform method to generate $f(y|x), f(x|y)$.
- To generate $f(y|x)$, we compute

$$u = \int_0^z \frac{2x + 3y + 2}{4x + 10} dy$$

- We get
  $z = \sqrt{(2/3)u(4x + 10) + ((1/3)(2x + 2))^2} - (1/3)(2x + 2)$.
- Similarly, we get $x = \sqrt{u(6y + 8) + (1.5y + 1)^2} - (1.5y + 1)$

Example 2

- Generate the r.v.s

$$(x, y) \sim f(x, y) = cC_n^x y^{x+\alpha-1}(1-y)^{n-x+\beta-1},$$

where $x = 0, 1, 2, \cdots, n, 0 \leq y \leq 1, \alpha, \beta$ are known.

- One can see that

$$f(x|y) = \frac{f(x,y)}{f(y)} \propto C_n^x y^x (1-y)^{n-x} \sim \textit{Binomial}(n, y)$$

$$f(y|x) = \frac{f(x,y)}{f(x)} \propto y^{x+\alpha-1}(1-y)^{n-x+\beta-1} \sim \textit{Beta}(x+\alpha, n-x+\beta)$$

Gibbs sampling Algorithm:

- Initial Setting: $y^0 \sim U[0,1]$ or an arbitrary value in [0,1], $x^0 \sim Binomial(n, y^0)$
- For $t = 0, \cdots, n$, sample a value (x^{t+1}, y^{t+1}) from

$$y^{t+1} \sim Beta(x^t + \alpha, n - x^t + \beta)$$

$$x^{t+1} \sim Bin(n, y^t)$$

- Return (x^t, y^t).

Example: Bivariate normal distribution

- Multivariate normal distribution:

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$

- Bivariate normal distribution with mean $(0, 0)$, variance $(1, 1)$, and correlation parameter ρ. The pdf is:

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

$$p(x \mid y) = \frac{p(x,y)}{p(y)}$$

$$= \frac{\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-a_1)^2}{\sigma_1^2} - 2\rho\frac{(x-a_1)(y-a_2)}{\sigma_1\sigma_2} + \frac{(y-a_2)^2}{\sigma_2^2}\right]}}{\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-a_2)^2}{2\sigma_2^2}}}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} e^{-\frac{1}{2\sigma_1^2(1-\rho^2)}\left[x-\left(a_1+\rho\frac{\sigma_1}{\sigma_2}(y-a_2)\right)\right]^2}$$

- $$X|Y = y \sim N(\rho y, 1-\rho^2) \sim \rho y + \sqrt{1-\rho^2}N(0,1)$$
  $$Y|X = x \sim N(\rho x, 1-\rho^2) \sim \rho x + \sqrt{1-\rho^2}N(0,1)$$

- Generate (X_0, Y_0). Set $i = 1$
- Run until convergence: Generate

$$X_i \sim \rho y + \sqrt{1 - \rho^2} N(0, 1)$$

$$Y_i \sim \rho x + \sqrt{1 - \rho^2} N(0, 1)$$

Limitations of the 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.
- Even if it is possible to implement the Gibbs sampler, the algorithm might be very inefficient because the variables are very correlated or sampling from the full conditionals is extremely expensive/inefficient.

Metropolis-Hastings algorithm

- The Metropolis-Hastings algorithm is an alternative algorithm to sample from probability distribution $\pi(\theta)$ known up to a normalizing constant.
- This can be interepreted as the basis of all MCMC algorithm: It provides a generic way to build a Markov kernel admitting $\pi(\theta)$ as an invariant/stationary distribution.
- The Metropolis algorithm was named the “Top algorithm of the 20th century” by computer scientists, mathematicians, physicists.

Description of the algorithm

- Introduce a proposal distribution/kernel $q(\theta, \theta')$, i.e. $\int q(\theta, \theta') d\theta' = 1$ for any θ.
- The basic idea of the MH algorithm is to propose a new candidate θ' based on the current state of the Markov chain θ.
- We only accept this algorithm with respect to a probability $\alpha(\theta, \theta')$ which ensures that the invariant distribution of the transition kernel is the target distribution $\pi(\theta)$.

Metropolis-Hastings algorithm

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

$$\alpha(\theta^{(i-1)}, \theta^*) = \min(1, \frac{\pi(\theta^*)q(\theta^*, \theta^{(i-1)})}{\pi(\theta^{(i-1)})q(\theta^{(i-1)}, \theta^*)})$$

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

Metropolis-Hastings algorithm

- It is not necessary to know the normalizing constant of $\pi(\theta)$ to implement the algorithm.
- This algorithm is extremely general: $q(\theta, \theta')$ can be any proposal distribution. So in practice, we can select it so that it is easy to sample from it.
- There is much more freedom than in the Gibbs sampler where the proposal distributions are fixed.

Metropolis algorithm

- The original Metropolis algorithm (1953) corresponds to the following choice for $q(\theta, \theta')$, $\theta' = \theta + Z$, where $Z \sim f$; i.e. this is a so-called random walk proposal.
- The distribution $f(z)$ is the distribution of the random walks increments Z and

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

- If $f(\theta' - \theta) = f(\theta - \theta')$, e.g. $Z \sim N(0, 1)$, then

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

Example 1: Generating exponential distribution

- $p(x) \propto e^{-x}$
- We use Metropolis algorithm to generate the distribution.
- Algorithm: for (i in 2:1000){
  currentx = x[i-1]
  proposedx = currentx + rnorm(1,mean=0,sd=1)
  A = target(proposedx)/target(currentx)
  if(runif(1)<A){
  x[i] = proposedx # accept move with probabily min(1,A)
  } else{
  x[i] = currentx # otherwise "reject" move, and stay where we are }
  }

Metropolis algorithm

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

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

  i.e. this is a so-called independent proposal.
- In this case, the acceptance probability is given by

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

  where π^* and q^* are unnormalized versions of π and q.
- The ratio $\pi^*(\theta)/q^*(\theta)$ appearing in the Accept/Reject and Importance Sampling methods also reappears here.

Properties of the Metropolis-Hastings algorithm

- To establish that the MH chain 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.

Stationary distribution of the Metropolis-Hastings algorithm

Discrete case:

- Define the Markov chain $\{X_n\}$:
  When $X_n = i$, a random variable X such that $P(X = j) = q(i, j)$ is generated.
  If $X = j$, then X_{n+1} is set equal to j with probability $\alpha(i, j)$ and is set equal to i with probability $1 - \alpha(i, j)$.
- The transition probability matrix

$$P_{i,j} = q(i, j)\alpha(i, j), j \neq i$$

$$P_{i,i} = q(i, i) + \sum_{k \neq i} q(i, k)(1 - \alpha(i, k))$$

Stationary distribution of the Metropolis-Hastings algorithm

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

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

- Remark: This is a lose notation for

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

- Clearly, we have

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

- 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: For all θ, θ',

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

  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.
- Indeed the reversibility condition implies that

$$\begin{aligned}\int \pi(\theta)K(\theta,\theta')d\theta &= \int \pi(\theta')K(\theta',\theta)d\theta \\ &= \pi(\theta')\int K(\theta',\theta)d\theta = \pi(\theta')\end{aligned}$$

Proof that the MH kernel is reversible

- By definition of the kernel, we have

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

- Then,

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

- We have obviously

$$(1-\int \alpha(\theta,u)q(\theta,u)du)\delta_\theta(\theta')\pi(\theta) = (1-\int \alpha(\theta',u)q(\theta',u)du)\delta'_\theta(\theta)\pi(\theta$$

- It follows that

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

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

- To ensure irreducibility, a sufficient but not necessary condition is that
$$\pi(\theta') > 0 \Rightarrow q(\theta, \theta') > 0.$$
- Aperiodicity is automatically ensured as there is always a strictly positive probability to reject the candidate.
- Theoretically, the MH algorithm converges under very weak assumptions to the target distribution π. In practice, this convergence can be so slow that the algorithm is useless.

Example 2:Estimating an allele frequency

- Data AA, Aa and aa have frequencies $p \times p, 2 \times p \times (1-p)$ and $(1-p) \times (1-p)$
- Prior: $p \sim U[0,1]$.
- Likelihood: $(p^2)^{n_{AA}}(2p(1-p))^{n_{Aa}}((1-p)^2)^{n_{aa}}$
- Posterior $\propto$ Prior$\times$Likelihood

Example 3: Linear regression

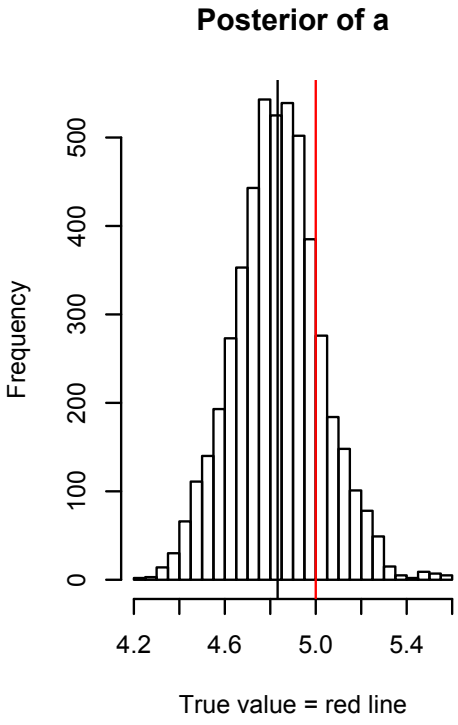
- $$Y = aX + b + \epsilon, \epsilon \sim N(0, \delta^2)$$

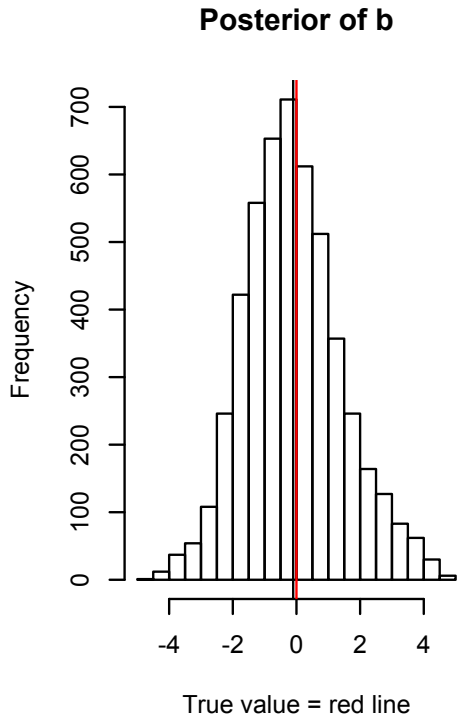
  Given X, Y, we estimate a, b and δ^2.
- For frequentists, ususally solve the optimization problem:

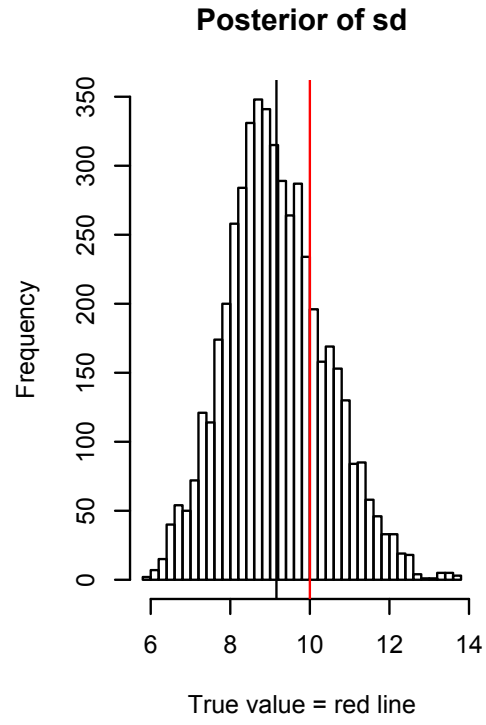
$$\min \| Y - aX \|_2^2$$

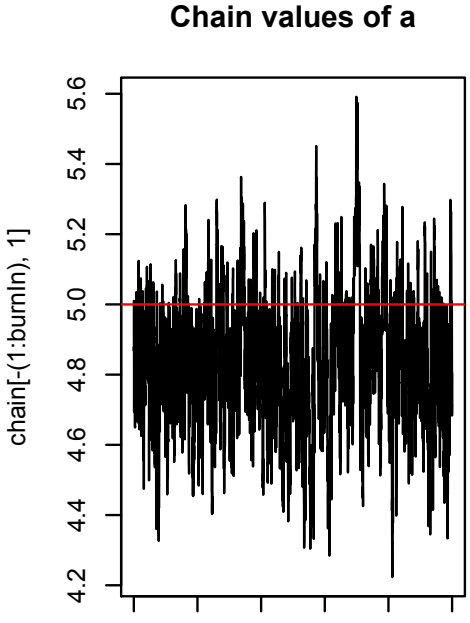
- With Bayesian, we need to define the likelihood function, prior distribution.

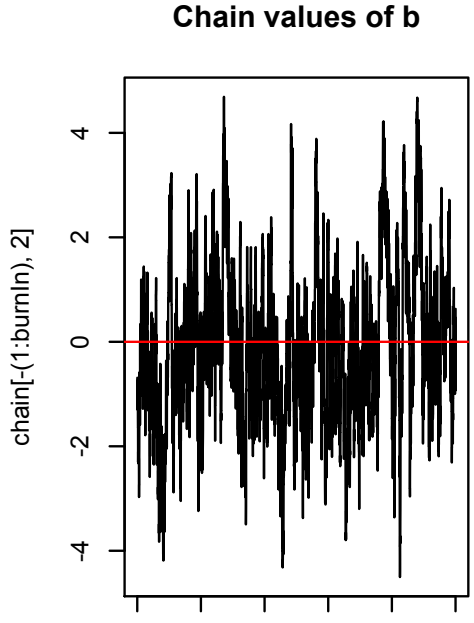
- For the likelihood function, $\epsilon \sim N(0, \delta^2)$. We may use the normal distribution. Here, we suggest use log likelihood, because likelihoods, where a lot of small probabilities are multiplied, can get ridiculously small pretty fast. At some stage, computer programs are getting into numerical rounding or underflow problems then. When you program something with likelihoods, always use logarithms!!!
- For the prior, we use uniform distributions and normal distributions for all three parameters.
- Posterior$\propto$ Likelihood$\times$ Prior

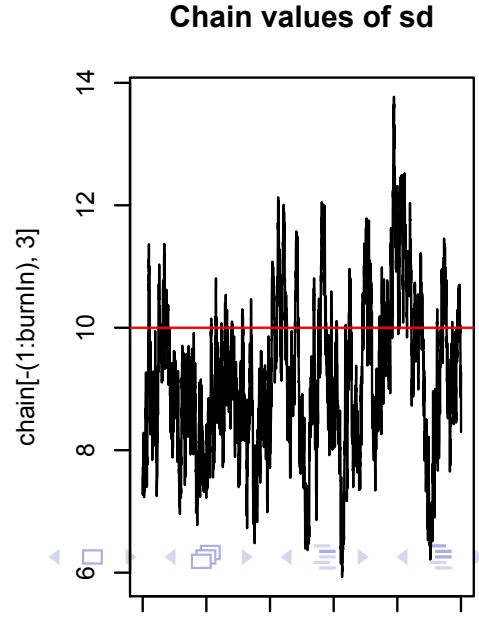

Posterior of a
Frequency
0
100
200
300
400
500
4.2
4.6
5.0
5.4
True value = red line


Posterior of b
Frequency
0
100
200
300
400
500
600
700
-4
-2
0
2
4
True value = red line


Posterior of sd
Frequency
0
50
100
150
200
250
300
350
6
8
10
12
14
True value = red line


Chain values of a
chain[-(1:burnIn), 1]
4.2
4.4
4.6
4.8
5.0
5.2
5.4
5.6


Chain values of b
chain[-(1:burnIn), 2]
-4
-2
0
2
4


Chain values of sd
chain[-(1:burnIn), 3]
6
8
10
12
14

How to select the proposal distribution

- If you are using independent proposals then you would like to have $q(\theta) \simeq \pi(\theta)$.
- In practice, similarly to Rejection sampling, you need to ensure that to obtain good performance.

$$\frac{\pi(\theta)}{q(\theta)} \leq c$$

- If you don't ensure this condition, the algorithm might give you the impression it works well... but it does NOT.

Example

Consider the case where

$$\pi(\theta) \propto e^{-\frac{\theta^2}{2}}$$

-
- We implement the MH algorithm for

$$q_1(\theta) \propto e^{-\frac{\theta^2}{2\times(0.2)^2}},$$

so $\pi(\theta)/q_1(\theta) \to \infty$ as $\theta \to \infty$

- For

$$q_2(\theta) \propto e^{-\frac{\theta^2}{2\times(5)^2}},$$

so $\pi(\theta)/q_2(\theta) \leq C < \infty$ as $\theta \to \infty$

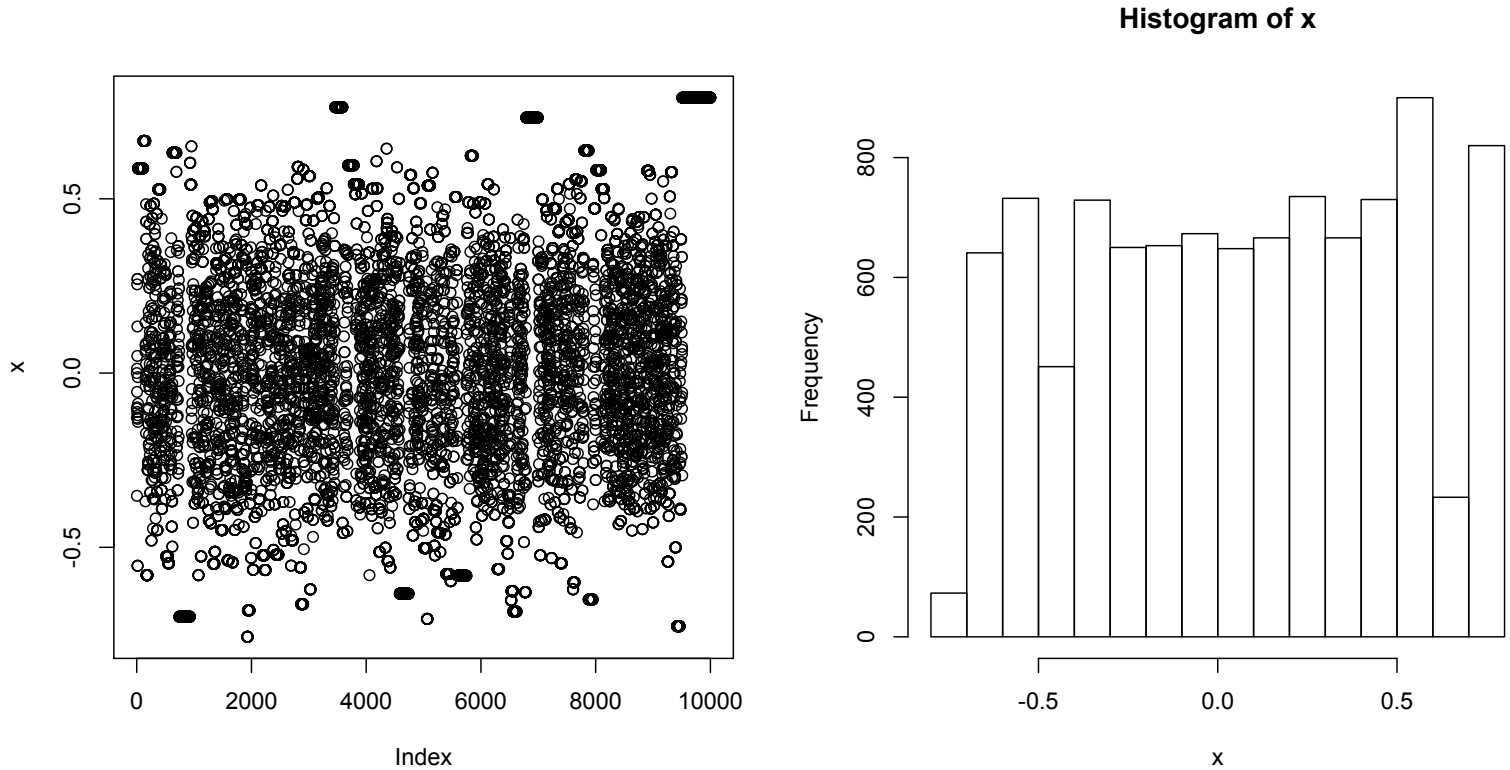


Figure : MCMC for q_1.

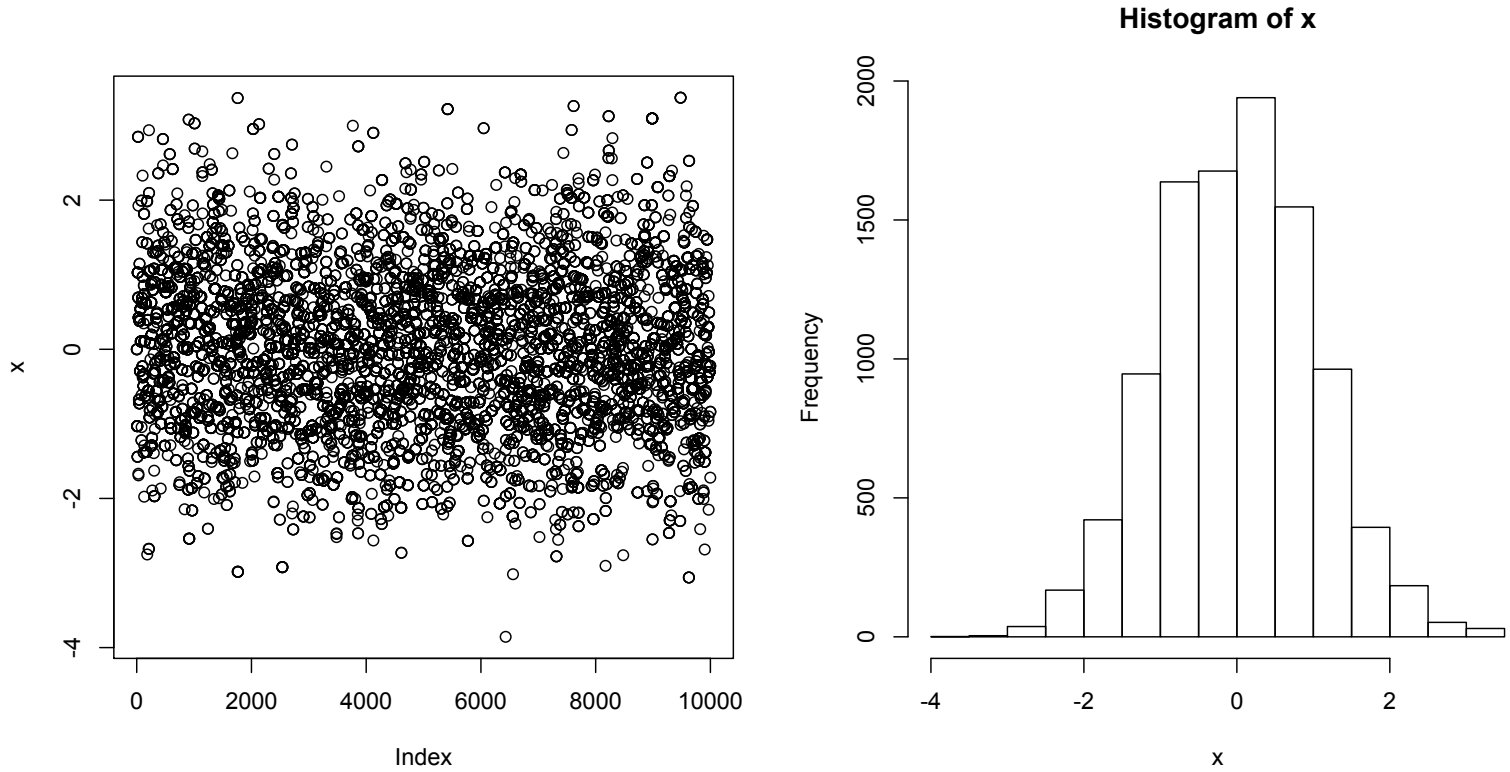


Figure : MCMC for q_2.

How to select the proposal?

- Consider now a random walk move. In this case, there is no clear guideline how to select the proposal distribution.
- When the variance of the random walk increments (if it exists) is very small then the acceptance rate can be expected to be around 0.5-0.7.
- You would like to scale the random walk moves such that it is possible to move reasonably fast in regions of positive probability masses under π.

Example

Consider the case where

$$\pi(\theta) \propto e^{-\frac{\theta^2}{2}}$$

-
- We implement the MH algorithm for

$$q_1(\theta) \propto e^{-\frac{(\theta'-\theta)^2}{2\times(0.2)^2}},$$

  so $\pi(\theta)/q_1(\theta) \to \infty$ as $\theta \to \infty$
- For

$$q_2(\theta) \propto e^{-\frac{(\theta'-\theta)^2}{2\times(5)^2}},$$

  so $\pi(\theta)/q_2(\theta) \leq C < \infty$ as $\theta \to \infty$

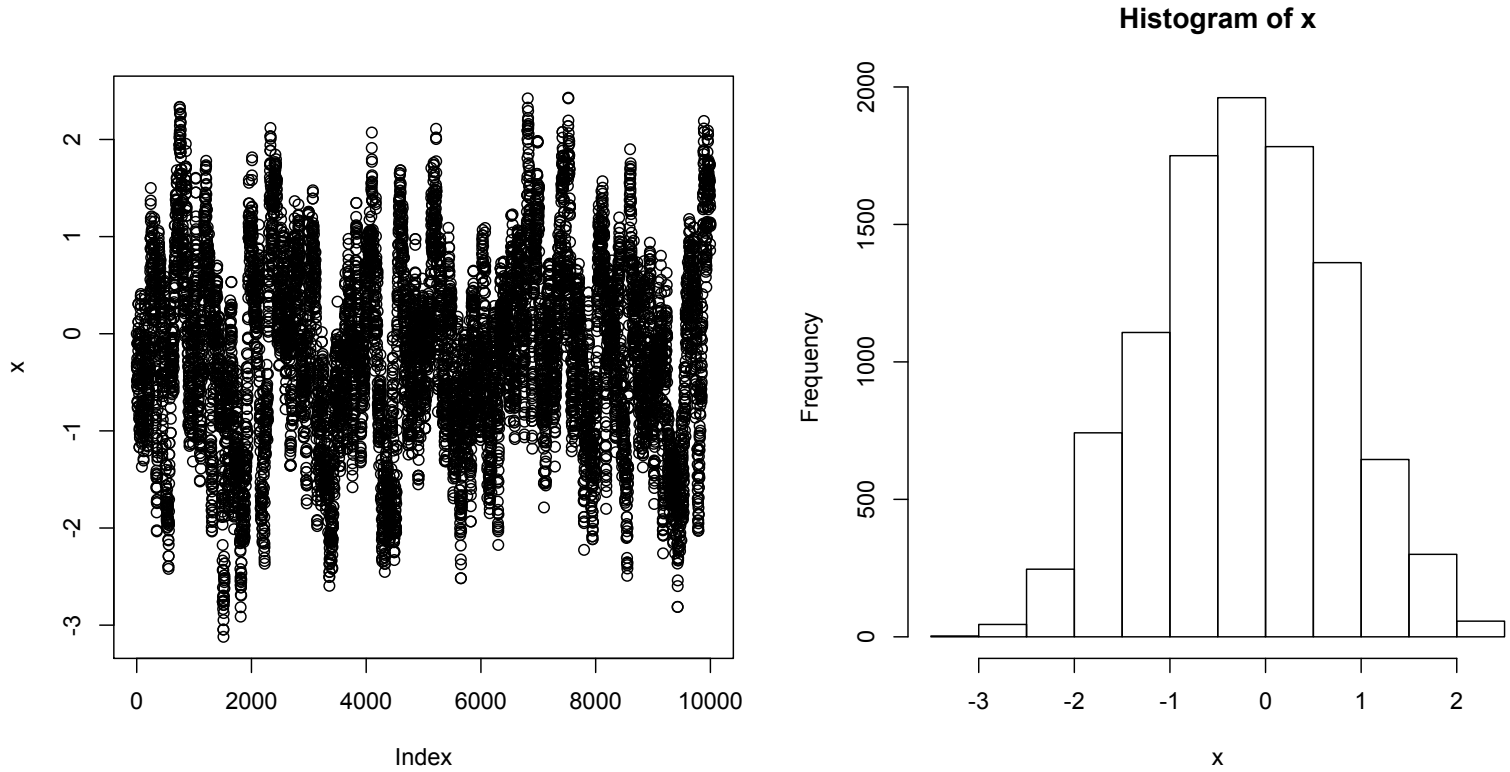


Figure : MCMC for q_1.

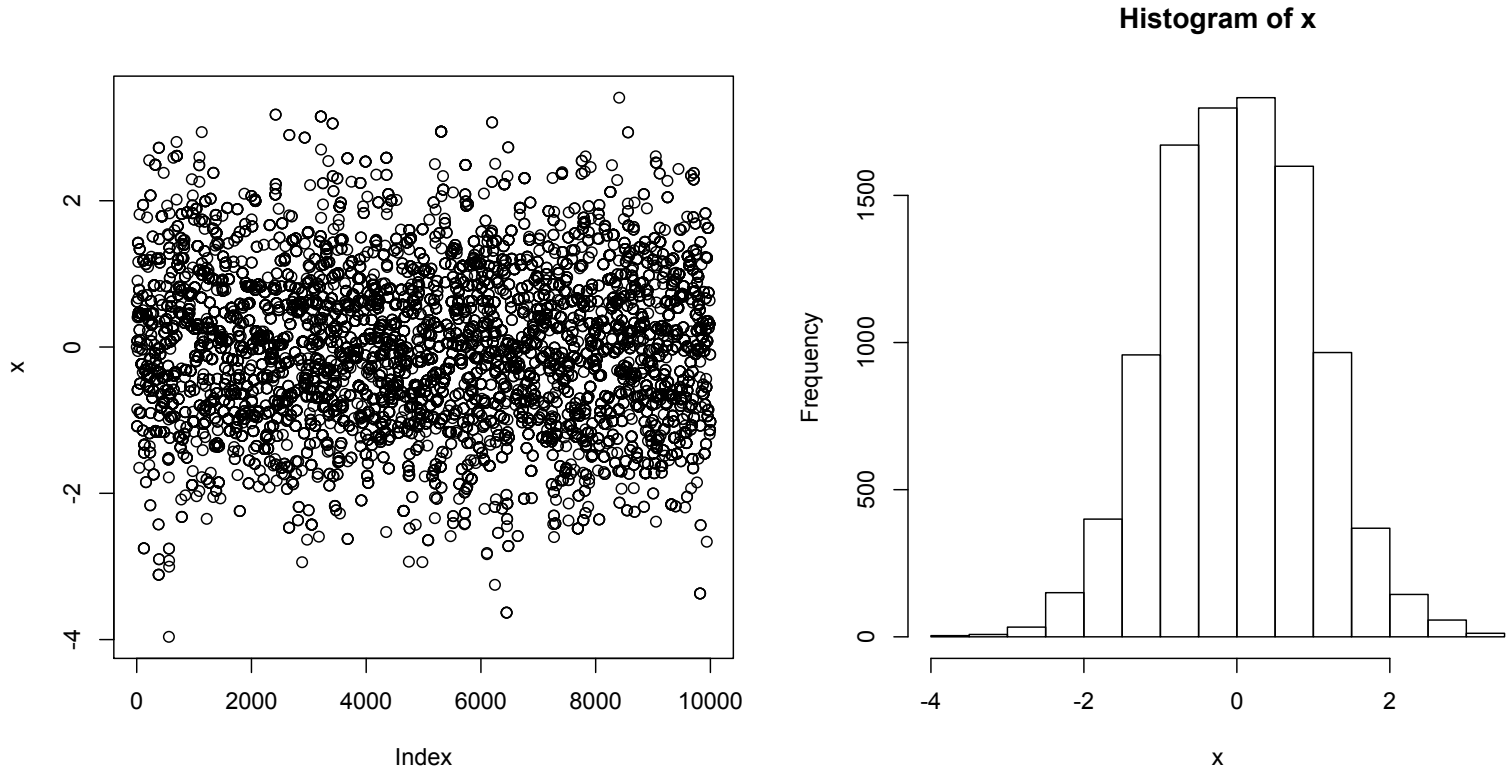


Figure : MCMC for q_2.

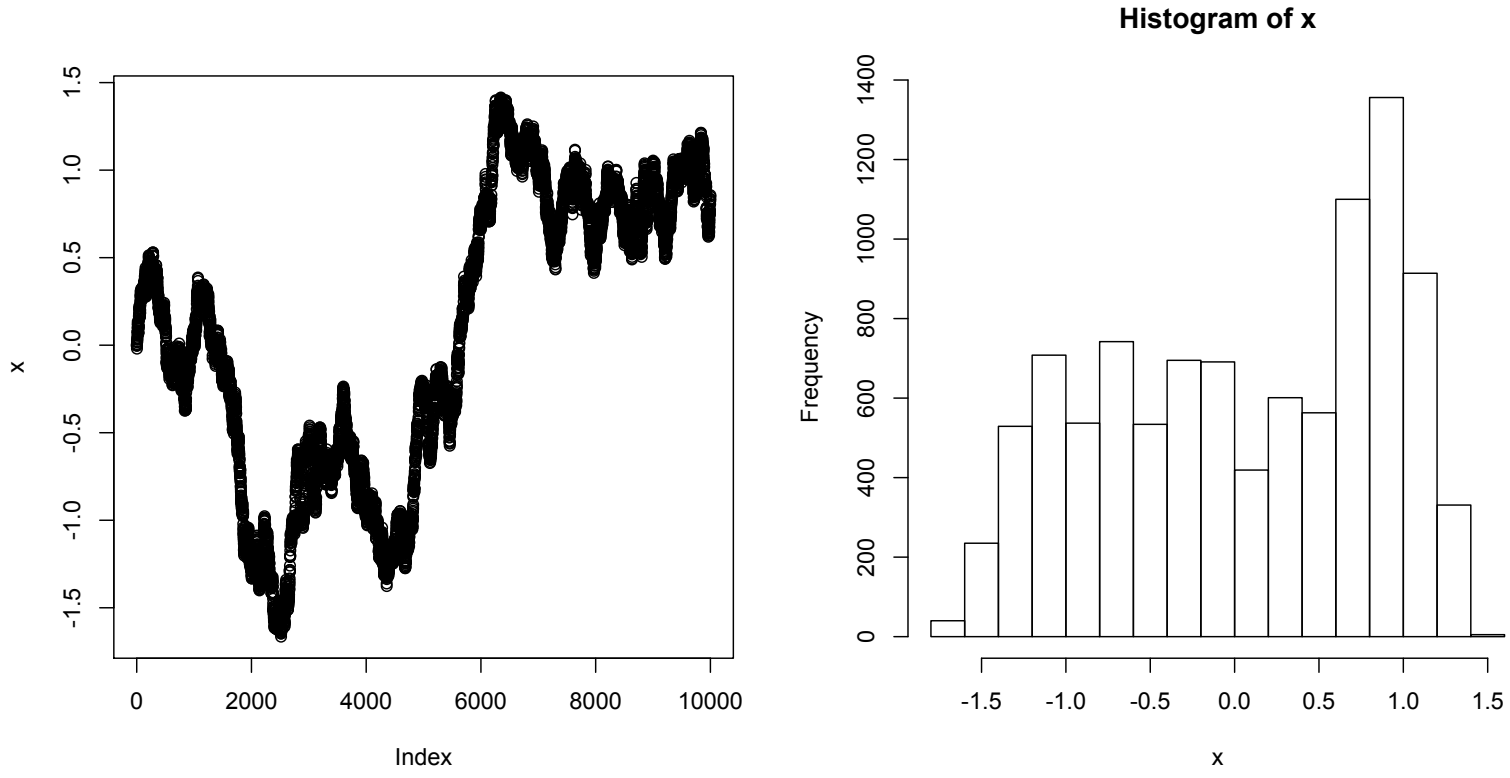


Figure : MCMC for $q_3.q_3(\theta) \propto e^{-\frac{(\theta'-\theta)^2}{2\times(0.02)^2}}$

Example 4:mixture of two 2-dimensional normal densities

- $$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{x}-\mu)\right)$$
- $$f = 0.3 f_1 + 0.7 f_2$$

Burn-in

- Since convergence usually occurs regardless of our starting point, we can usually pick any feasible (for example, picking starting draws that are in the parameter space) starting point. However, the time it takes for the chain to converge varies depending on the starting point.
- As a matter of practice, most people throw out a certain number of the first draws, known as the burn-in. This is to make our draws closer to the stationary distribution and less dependent on the starting point.
- However, it is unclear how much we should burn-in since our draws are all slightly dependent and we don't know exactly when convergence occurs.

Acceptance Rates

- It is important to monitor the acceptance rate (the fraction of candidate draws that are accepted) of your Metropolis-Hastings algorithm.
- If your acceptance rate is too high, the chain is probably not mixing well (not moving around the parameter space quickly enough).
- If your acceptance rate is too low, your algorithm is too inefficient (rejecting too many candidate draws).
- What is too high and too low depends on your specific algorithm, but generally
  random walk: somewhere between 0.25 and 0.50 is recommended
  independent: something close to 1 is preferred