Stochastic Models in Space and Time

Owen Ward

November 19, 2015

Abstract

Summary notes based on ST3453, as taught by Jason Wyse, Michaelmas Term 2014.

Contents

1 Examples of Stochastic Processes and some basics

1.1 Some Definitions

Many processes in economics, biology, physics and other areas can be modelled using Markov models. To put it very simply, Markov models use data from the past to expect what will happen in the future.

Definition 1.1. *If we have some process X_n, we say that X_n has the* ***markov property*** *if, given the current state X_n, then all other previous information about past events is irrelevant for predicting the next state X_{n+1}.*

If we take some simple game, where each time you either win €1 with probability $p = 0.4$ or lose €1 with probability $1 - p = q = 0.6$, and $X_n = i$, then

$$Pr\{X_{n+1} = i + 1 \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i\} = p = 0.4,$$

so it does not matter how much money you currently have, or whether you won previous games.

Definition 1.2. *X_n is a discrete time (if time is indexed by the positive integers)* ***markov chain*** *with transition probabilities $p(i, j)$ if*

$$Pr\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, x_0 = i_0\} = Pr\{X_{n+1} = j \mid X_n = i\} = p(i, j).$$

Definition 1.3. *$p(i, j)$ is said to be* ***temporally homogeneous*** *if $p(i, j) = Pr\{x_{n+1} = j \mid X_n = i\}$ does not depend on n.*

The transitions probabilities basically determine everything about the process.

Definition 1.4. *We can express all these transition probabilities succinctly in a* ***transition matrix****, where $p(i, j)$ is the i, jth entry in the matrix.*

Definition 1.5. *If we have a transition matrix where the sum of the entries in every row is 1, then we call it a* ***stochastic matrix*** *and similarly, if the sum of all rows and columns is 1, we call it a* ***doubly stochastic matrix****.*

1.2 An Example

Example The Genome sequence of living organisms is a string of 4 characters, A(adenine), C(cytosine), G(guanine) and T(thymine). In DNA terminology these are bases and there are complimentary base pairs, A with T and C with G. Evolution of organisms occurs because of mutations in these base pairs. Modern evolutionary models are based on Markov Processes in continuous time. At any site on the genome we have a stochastic variable $X(t)$ taking one of the values $\{1, 2, 3, 4\} \leftrightarrow \{A, C, T, G\}$. Then

$$Pr\{X(t + s) = j \mid X(s) = i\} = Pr\{X(t) = j \mid X(0) = i\} = p_{i,j}(t)$$

so the probabilities have the Markov property, and we can form the transition matrix

$$P = \begin{pmatrix} Pr\{A \mid A, t\} & Pr\{C \mid A, t\} & Pr\{G \mid A, t\} & Pr\{T \mid A, t\} \\ \vdots & \ddots & \vdots & \\ Pr\{A \mid T, t\} & Pr\{C \mid T, t\} & Pr\{G \mid T, t\} & Pr\{T \mid T, t\} \end{pmatrix}.$$

Using some further assumptions on the structure of the probabilities we can get the Dukes-Cantor model for base substitution, with

$$p_{i,j}(t) = \frac{1}{4}\left(1 + 3e^{-4\alpha t}\right),$$
$$p_{j,i}(t) = \frac{1}{4}\left(1 - e^{-4\alpha t}\right),$$

for the parameter α which can be estimated from the data.

2 The Markov Property and Markov Chains

2.1 The Markov Process

Definition 2.1. *If we have a stochastic process $\{X_n,\ n=0,1,2,\ldots\}$ that can take on a finite number of values, denoted by the non-negative integers. The process is in state i at time n if $X_n = i$. Since the time index is discrete, we can say X_n is a discrete time process. It is also a finite state process. If we assume there is a fixed probability that the process will be in state j at time $n+1$, given it is in state i at time n, then*

$$Pr\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = p(i,j),$$

*for all states $i_0, \ldots, i_{n-1}, i, j$ and for all $n \geq 0$, then X_n is a **markov process***

2.2 The Markov Property

If X_n is a Markov Chain then it has the Markov property. This means the conditional distribution of any future state X_{n+1} given all past states $X_0, \ldots, X_n$ depends only on X_n, and independent of all states before X_n, i.e

$$Pr\{X_{n+1} \mid X_n, \ldots, X_0\} = Pr\{X_{n+1} \mid X_n\}.$$

2.3 Transition Probabilities

The one step transition probabilities $p(i,j)$ give the probability that the chain X_n goes from state i to state j in one step. As $p(i,j)$ are probabilities, it is clear that $p(i,j) \geq 0, \forall 1 \leq i,j \leq k$ and since the chain either stays where it is or changes to a different state $\sum\limits_{j=1}^{k} p(i,j) = 1$.

Example Let $\{X_t, t \geq 1\}$ be independently identically distributed (iid) (so $\forall\ t, E[X_t] = \mu, \mu \in \mathbb{R}$ and $var(X_t) = \sigma^2,\ \sigma \in \mathbb{R},\ Cov(X_t, X_k) = 0$). Suppose

$$Pr\{X_t = l\} = a_l \ l = 0, \pm 1, \ldots,$$

and that $S_n = 0,\ S_n = \sum\limits_{t=1}^{n} X_t$. To show S_n has the markov property,

$$Pr\{S_{n+1} = j | S_n = i, \ldots, S_0 = 0\}$$

$$= Pr\{S_n + X_{n+1} = j | S_n = i, \ldots, S_0 = 0\}$$

$$= Pr\{X_{n+1} = j - i | S_n = i, \ldots, S_0 = 0\}$$

$$= Pr\{X_{n+1} = j - i\} = a_{j-1}.$$

So S_n does indeed satisfy the Markov property. This process S_n is known as a random walk.
A simple random walk is a process $\{S_n, n \geq 0\},\ S_0 = 0$ where $S_n = \sum\limits_{t=1}^{n} X_t$ with X_t (iid) and

$$Pr\{X_t = 1\} = p,\ Pr\{X_t = -1\} = 1 - p = q, \quad \text{for } 0 < p < 1.$$

It can also be shown that $\|S_n\|$ the distance of a random walk from the origin is also a Markov process:

$$Pr\{S_n = i \mid |S_n| = i, \ldots, |S_1| = i_1\}.$$

We let $i_0 = 1$ and j be the last timepoint where the process crossed zero.

$$j = \max\{k : 0 \leq k \leq n, i_k = 0\}.$$

Then $S_j = 0$, so

$$Pr\{S_n = i \mid |S_n| = i, \ldots, |S_1| = i_1\} = Pr\{S_n = i \mid |S_n| = i, \ldots, |S_j| = 0\}.$$

There are two possible values for this sequence, $S_j + 1, \ldots, s_n$ for which $|S_{j+1} = i_{j+1}, \ldots, |S_n| = i$. since the process doesn't cross zero in this time period these are $i_{j+1}, \ldots, i$ or $-i_{j+1}, \ldots, -i$. If we assume $i_{j+1} > 0$ we look at the first of these sequences. In these $n - j$ steps there are i more up steps than down steps. Letting ds be the number of down steps, then we have

$$(ds + i) + ds = n - j \text{ so } ds = \frac{n - j - i}{2}.$$

So the probability of this sequence will be $p^{\frac{n-j-i}{2}+i}q^{\frac{n-j-i}{2}} = p^{\frac{n-j+i}{2}}q^{\frac{n-j-i}{2}}$ and similarly the probability of the second sequence will be $p^{\frac{n-j-i}{2}}q^{\frac{n-j+i}{2}}$ so

$$Pr\{S_n = i \mid |S_n| = i, \ldots, |S_{j+1}| = i_{j+1}\}$$

$$= \frac{p^{\frac{n-j+i}{2}}q^{\frac{n-j-i}{2}}}{p^{\frac{n-j+i}{2}}q^{\frac{n-j-i}{2}} + p^{\frac{n-j-i}{2}}q^{\frac{n-j+i}{2}}} = \frac{p^i}{p^i + q^i},$$

and also

$$Pr\{S_n = -i \mid |S_n| = i, \ldots, |S_{j+1}| = j+1\} = \frac{q^i}{p^i + q^i},$$

and then, on conditioning on whether S_n is $-i$ or i, we get

$$Pr\{|S_{n+1}| \mid |S_n| = i, \ldots, |S_1| = i_1\}$$

$$= Pr\{S_{n+1} = i+1 \mid S_n = i\}Pr\{S_n = i \mid |S_n| = i, \ldots, |S_1| = i_1\}$$

$$+Pr\{S_{n+1} = -(i+1) \mid S_n = -i\}Pr\{S_n = -i \mid |S_n| = i, \ldots, |S_1| = i_1\}$$

$$= p\frac{p^i}{p^i + q^i} + q\frac{q^i}{p^i + q^i} = \frac{p^{i+1} + q^{i+1}}{p^i + q^i}.$$

So $\{|S_n|, n \geq 1\}$ is a Markov Chain, with transition probabilities

$$p(i, i+1) = \frac{p^{i+1} + q^{i+1}}{p^i + q^i}$$

$$p(i, i-1) = \frac{p^i(1-p) + q^i(1-q)}{p^i + q^i}$$

$$p(0, 1) = 1.$$

2.4 Multistep Transition Probabilites

The probability $p(i,j) = Pr\{X_{n+1} = j \mid X_n = i\}$ gives the probability of going from state i to state j in one step. If we wish to go from i to j in m steps,

$$Pr\{X_{n+m} = j \mid X_n = i\} = p^m(i,j).$$

If we look at the two step case, then we get

$$p^2(i,j) = Pr\{X_{n+2} = j \mid X_n = i\} = \sum_{l=1}^{k} p(i,l)p(l,j)$$

If we think of this in terms of the transition matrix P, then it can be seen that $p^2(i,j)$ is the dot product of the i^{th} row of P with the j^{th} column of P, which is the $(i,j)^{th}$ entry of P^2.

2.5 The Chapman-Kolmogrov Equation

This equation is very useful for understanding multi step transition probabilities. It states that

$$p^{m+n}(i,j) = \sum_{l=1}^{k} p^m(i,l)p^n(l,j).$$

Proof. To prove this we break it down according to the states at time m.

$$Pr\{X_{m+n} = j \mid X_0 = i\} = \sum_{l=1}^{k} Pr\{X_{m+n} = j, X_m = l \mid X_0 = i\}.$$

We then use conditional probability to compute the term in the sum

$$Pr\{X_{m+n} = j, X_m = l \mid X_0 = i\} = \frac{Pr\{X_{m+n} = j, X_m = l, x_0 = i\}}{Pr\{X_0 = i\}}$$

$$\frac{Pr\{X_{m+n} = j, X_m = l, X_0 = i\}}{Pr\{X_m = l, X_0 = i\}} \frac{Pr\{X_m = l, X_0 = i\}}{Pr\{X_0 = i\}}$$

$$Pr\{X_{m+n} = j \mid X_m = l, X_0 = i\}Pr\{X_m = l \mid X_0 = i\}.$$

By the Markov property, the first term is is $Pr\{X_{m+n} = j \mid X_m = l\}$ so that

$$\begin{aligned} Pr\{X_{m+n} = j, X_m = l \mid X_0 = i\} &= Pr\{X_{m+n} = j \mid X_m = l\}Pr\{X_m = l \mid X_0 = i\} \\ &= p^n(l,j)p^m(i,l), \end{aligned}$$

so we have

$$p^{m+n}(i,j) = \sum_{l=1}^{k} p^m(i,l)p^n(l,j).$$

□

Taking $n = 1$ in this equation we get

$$p^{m+1}(i,j) = \sum_{l=1}^{k} p^m(i,l)p(l,j),$$

which is the i^{th} row of the m step transition matrix multiplied by the j^{th} column of P. So the $m+1$ step transition matrix is given by P^{m+1}.

The m-step transition matrix is given by the one step matrix raised to the power of m.

Example Consider the general two state chain with transition matrix

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \text{ for } 0 \le a \le 1,\ 0 \le b \le 1.$$

What is P^n in general, and what is the limiting behaviour?

Writing $P = Q\Lambda Q^{-1}$, then $P^n = \left(Q\Lambda Q^{-1}\right)^n = Q\Lambda^n Q^{-1}$, for Λ a diagonal matrix and Q some matrix to be found. Computing the eigen decomposition, we find the eigenvalues are $\lambda_1 = 1, \lambda_2 = 1-a-b$, so $\Lambda = \begin{pmatrix} 1 & \\ & 1-a-b \end{pmatrix}$ and the eigenvectors can also be easily computed, giving $Q = \begin{pmatrix} y & z \\ y & -\frac{b}{a}z \end{pmatrix}$ say. Then

$$Q^{-1} = \frac{-a}{yz(a+b)} \begin{pmatrix} \frac{-b}{a}z & -z \\ -y & y \end{pmatrix},$$

and then using this to get P^n, we find that

$$P^n = \begin{pmatrix} \frac{b}{a+b} + \frac{a}{a+b}(1-a-b)^n & \frac{a}{a+b} - \frac{a}{a+b}(1-a-b)^n \\ \frac{b}{a+b} - \frac{b}{a+b}(1-a-b)^n & \frac{a}{a+b} + \frac{b}{a+b}(1-a-b)^n \end{pmatrix}.$$

We then consider as $n \to \infty$? If $|1-a-b| < 1$ then $(1-a-b)^n \to 0$ as $n \to \infty$. $|1-a-b| < 1$ if $0 < a+b < 2$ then

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}.$$

So

$$\left(\frac{b}{a+b}, \frac{a}{a+b}\right)$$

is the stationary distribution of the chain.

3 Properties of Markov Chains

This section examines properties that can be used to classify the behaviour of Markov chains.

3.1 Decomposability

Definition 3.1. *A set of states A is* ***closed*** *if $Pr\{X_{n+1} \in A \mid X_n = x\} = 1$ for all states $x \in A$. If we start in a closed state A, then we will always stay in A, and nothing outside of it matters.*

Definition 3.2. *A markov chain is* ***indecomposable*** *if its set of states doesn't contain two or more disjoint closed sets of states.*

Definition 3.3. *If there is a transition from i to j, i.e there is some m such that $p^m(i,j) > 0$ then we write $i \to j$. If additionally there is some n such that $p^n(j,i) > 0$, i.e can transition both ways then we say i* ***communicates*** *with j and write $i \leftrightarrow j$.*

If for every pair of states i and j, at least one of $i \to j$ or $j \to i$ is possible then the chain's set of states is indecomposable.

3.2 Periodicity

Periodicity can occur if the set of states decomposes into say two closed sets, for example if there are two disjoint sets B_1, B_2 such that

$$p^2(i, B_1) = 1 \ \forall i \in B_1$$

$$p^2(i, B_2) = 1 \ \forall i \in B_2.$$

If we consider a simple random walk, if the current state is an odd integer, then the next will be even, and the following will be odd. Then, if B_1 is the even integers and B_2 the odd, we have

$$p^2(i, B_1) = 1 \ \forall i \text{ even},$$

$$p^2(i, B_2) = 1 \ \forall i \text{ odd}.$$

To summarise periodic behaviour, let $d \geq 1$ be the largest integer such that the states can be decomposed into d disjoint subsets $B_1, \ldots, B_d$, each closed under the d step transition probability. Then the markov chain will cycle among the $B_1, \ldots, B_d$. If the starting state is B_1, the next state will be in B_2, and so on until the chain transitions from B_2 back to B_1.

3.3 Stability and Computing Stable Distributions

Stability can sometimes be used to make statements about the chain after a large number of movements. Is there a limiting distribution $\pi(A)$ such that

$$p^m(x, A) \to \pi(A) \text{ as } m \to \infty.$$

If the answer is yes, the chain is stable. If the chain is decomposable or periodic, we can't have stability. For decomposability, we can see that $p^m(x, A)$ as m increases will depend on which set of states we start in. Looking at the periodic example above it is clear why that won't work either.

If the chain is stable, no matter what state x we start from, the proportion of time the chain spends in the set of states A will be $\pi(A)$. Let

$$f(X_t) = \begin{cases} 1 & \text{if } X_t \in A \\ 0 & \text{otherwise} \end{cases},$$

then $\frac{1}{m} \sum\limits_{t=1}^{m} f(X_t)$ is the proportion of time spent in A.

$$\mathrm{E}[f(X_t)] = 1p^t(x, A) + 0\left(1 - p^t(x, A)\right) = p^t(x, A)$$

so the expected proportion of time spent in A is $\frac{1}{m} \sum\limits_{t=1}^{m} p^t(x, A)$. Since the chain is stable, $p^m(x, A) \to \pi(A)$ so that

$$\frac{1}{m} \sum_{t=1}^{m} p^t(x, A) \to \pi(A).$$

There is also a law of large numbers for this, namely

$$Pr\left\{\left|\frac{1}{m}\sum_{t=1}^{m} f(X_t) - \pi(A)\right| > \delta\right\} \to 0 \text{ as } m \text{ gets large.}$$

To compute a stable distribution, we know that

$$p^{m+1}(x, A) = \sum_{l=1}^{k} p^m(x, l)p(l, A),$$

by assumption, $p^{m+1}(x, A) \to \pi(A)$ as $m \to \infty$ and similarly, $p^m(x, l) \to \pi(l)$ as $m \to \infty$ so π must satisfy

$$\pi(A) = \sum_{l=1}^{k} \pi(l)p(l, A)$$

which is equivalent to solving

$$\Big(\pi(1), \dots, \pi(n)\Big) = \Big(\pi(1), \dots, \pi(n)\Big)P,$$

where $\pi = \Big(\pi(1), \dots, \pi(n)\Big)$ is the stable distribution vector, and P is the one step transition matrix, as defined previously.

Example If we wise to find the stationary, or stable distribution for the transition matrix

$$P = \begin{pmatrix} 1-a & a \\ a & 1-b \end{pmatrix},$$

then we have to solve the system of equations

$$\Big(\pi(1), \pi(2)\Big) = \Big(\pi(1), \pi(2)\Big)\begin{pmatrix} 1-a & a \\ a & 1-b \end{pmatrix},$$

which can easily be solved to give the stable distribution

$$\Big(\pi(1), \pi(2)\Big) = \Big(\frac{b}{a+b}, \frac{a}{a+b}\Big).$$

Theorem 3.4. *For an indecomposable, non periodic chain with transition probabilities $p(x, A)$ such that any two states x and y communicate, then the system of equations*

$$\pi(j) = \sum_{l=1}^{k} p(l, j)\pi(l) \;\; j = 1, \dots, k-1, \;\; \sum_{l=1}^{k} \pi(l) = 1$$

will give a set of k linearly independent equations with unique solution π.

3.4 Detailed Balance

Definition 3.5. *$\pi(\cdot)$ is said to satisfy **detailed balance** if*

$$\pi(x)p(x, y) = \pi(y)p(y, x).$$

This is stronger than $\pi P = \pi$ and in fact

$$\sum_{x=1}^{k} \pi(x)p(x, y) = \sum_{x=1}^{k} \pi(y)p(y, x)$$

$$= \pi(y)\sum_{x=1}^{k} p(y, x) = \pi(y).$$

To think about this, it means that everything 'going from' x to y at any time is completely balanced by everything 'going from' y to x, while a stationary distribution says that the total transferred between each will be the same after all transfers.

Example A graph is described by

1. A set of vertices V, which is finite.
2. An adjacency matrix $A(u,v)$ which is 1 if there is an edge between u and v, and 0 otherwise.

By convention, $A(v,v)=0$. The **degree** of a vertex u is equal to the number of neighbours it has, i.e,

$$d(u)=\sum_v A(u,v).$$

Now consider a random walk X_n on this graph, with transition probability given by

$$p(u,v)=\frac{A(u,v)}{d(u)},$$

so if $X_n=n$ then we jump randomly to one of its neighbours at the next time point. Now $d(u)p(u,v)=A(u,v)$ and since A is symmetric, and non directed,

$$\pi(u)p(u,v)=\pi(v)p(v,u).$$

Taking $\pi(u)=cd(u)$ for c some positive constant, then

$$\pi(u)p(u,v)=cd(u)p(u,v)=cA(u,v)=cA(v,u)=cd(v)p(v,u)=\pi(v)p(v,u)$$

and we have detailed balance. To get the stable distribution, $\pi(u)=cd(u)$ and so

$$\sum_{v\in V}\pi(v)=1\ \Rightarrow c=\frac{1}{\sum\limits_{v\in V}d(v)} \text{ and } \pi(u)=\frac{d(u)}{\sum\limits_{v\in V}d(v)}$$

4 The Poisson Process

The number of events occurring in an interval of time will often be a random variable, and can be modelled by a Poisson Process.

4.1 Assumptions of the Poisson Process

Assume we observe the process for a fixed period of time of length t, and the number of events occurring in this interval $(0, t]$ is a random variable X, which is discrete with its probability depending on how events occur. We assume

- In a sufficiently short length of time Δt then either 1 or 0 events can occur, no more,
- The probability of exactly one event occurring in the interval Δt is $\lambda \Delta t$, i.e the probability of an event occurring is proportional to the length of the interval,
- Non overlapping intervals of length Δt are independent Bernoulli trials,

These are the assumptions of a **Poisson Process** with parameter λ.

4.2 Probability Law

If we divide $(0, t]$ into $n = \frac{t}{\Delta t}$ non overlapping equal intervals, which by assumption are independent Bernoulli trials. Each of these has probability of an event occurring $p = \lambda \Delta t$, and the probability of no event is $q = 1 - \lambda \Delta t$. Then X, the number of events in the interval $(0, t]$ is binomial with $n, p = \lambda \Delta t = \frac{\lambda t}{n}$.

$$Pr\{X = k\} = \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k}$$

$$= \frac{n!}{k!(n-k)!} \frac{(\lambda t)^k}{n^k} \left(1 - \frac{\lambda t}{n}\right)^n \left(1 - \frac{\lambda t}{n}\right)^{-k}$$

$$= \frac{(\lambda t)^k}{k!} \left(1 - \frac{\lambda t}{n}\right)^n \left(1 - \frac{\lambda t}{n}\right)^{-k} \frac{n(n-1)\dots(n-k+1)}{n^k}.$$

Considering the limiting case as $\Delta t \to 0$ and $n \to \infty$,

$$\lim_{n \to \infty} \left(1 - \frac{\lambda t}{n}\right)^{-k} = 1$$

$$\lim_{n \to \infty} \left(1 - \frac{\lambda t}{n}\right)^n = e^{-\lambda t}$$

$$\frac{n(n-1)\dots(n-k+1)}{n^k} \to 1,$$

so we have

$$\lim_{\Delta t \to \infty} Pr\{X = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t},$$

which is the Poisson probability law. If we sum this over all possible values, we see

$$\sum_{k=0}^{\infty} Pr\{X = k\} = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} = e^{-\lambda t} e^{\lambda t} = 1.$$

4.3 Moments of the Poisson Distribution

For any $l \geq 1$ it can be shown that

$$E\{x(x-1)\dots(x-l+1)\} = (\lambda t)^l.$$

Proof. Note $x(x-1)\dots(x-l+1) = 0$ if $x \leq l-1$ so then

$$E\{x(x-1)\dots(x-l+1)\} = \sum_{k=l}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} k(k-1)\dots(k-l+1)$$

$$= \sum_{k=l}^{\infty} \frac{(\lambda t)^{k-l}}{(k-l)!} (\lambda t)^l e^{-\lambda t} = e^{-\lambda t} (\lambda t)^l \sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!}$$

$$= e^{-\lambda t} (\lambda t)^l e^{\lambda t} = (\lambda t)^l.$$

□

So we have $E\{X\} = \lambda t,\ E\{X(X-1)\} = (\lambda t)^2$ and

$$Var\{X\} = E\{X(X-1)\} - (E\{X\})^2 + E\{X\} = (\lambda t)^2 - (\lambda t)^2 + \lambda t = \lambda t.$$

Example Molecules in a gas occur at a rate of α per cubic metre. Assume they are distributed independently, so the number of molecules in a cubic metre of air is a Poisson random variable with rate parameter α. If we wanted to be $(100-\delta)\%$ confident of finding at least one molecule in a sample, what size should the sample be?

Let the sample size be S with X the number of molecules, which is Poisson distributed with rate αS. So we would require

$$Pr\{X \geq 1\} = 1 - Pr\{X = 0\} = 1 - e^{-\alpha S} \equiv 1 - \delta.$$

So

$$\begin{aligned} & e^{-\alpha S} \leq \delta \\ \Rightarrow\ & -\alpha S \leq \log \delta \\ \Rightarrow\ & S \geq \frac{-1}{\alpha} \log \delta \end{aligned}$$

is the amount of air that would be required.

4.4 Time to First Arrival

Suppose we begin to observe a process at time $t = 0$ and let T be the time to the first event. T is a continuous random variable with range $R_T = \{t : t > 0\}$. Let t be any fixed positive number and consider the event $\{T > t\}$, the time to the first event being greater than t, which occurs if there are 0 events in $(0, t]$, which has probability

$$Pr\{X = 0\} = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}.$$

$Pr\{T > t\} = 1 - F_T(t) = e^{-\lambda t}$ is the survival function while $Pr\{T \leq t\} = F_T(t)$ is the distribution function. Then

$$F_T(t) = 1 - e^{-\lambda t}\ t > 0,$$

which has density function

$$f_T(t) = \lambda e^{-\lambda t} = \frac{d}{dt} F_t(t),$$

the density of an exponential random variable. Therefore the time to the first event in a Poisson process is exponentially distributed with parameter λ. The expected value of an exponential random variable is

$$E\{t\} = \int_0^\infty t \lambda e^{-\lambda t} dt = -\frac{(\lambda t + 1)}{\lambda} e^{-\lambda t} \Big|_{t=0}^{t=\infty} = \frac{1}{\lambda},$$

and it has moment generating function

$$M_T(t) = E(e^{tT}) = \int_0^\infty e^{ts} \lambda e^{-\lambda s} ds = \int_0^\infty \lambda e^{-s(\lambda - t)} ds$$

$$= \frac{-\lambda e^{-s(\lambda-t)}}{\lambda - t} \Bigg|_{t=0}^{t=\infty} = \frac{\lambda}{\lambda - t}, \quad \text{for } \lambda > t.$$

So then

$$E\{x\} = \frac{d}{dt} M_T(t) \Big|_{t=0} \text{ and } E\{x^j\} = \frac{d^j}{dt^j} M_T(t) \Big|_{t=0}.$$

This can also be used to verify $Var(t) = \frac{1}{\lambda^2}$, so the standard deviation is the same as the mean.

Example Suppose students arrive at lectures at a rate of 2 per minute. What is the probability no students arrive in 3 minutes?
$\lambda = 2$ so then $Pr(x = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-6} = 0.0025.$

4.5 Memoryless Property

If t is an exponential random variable with parameter λ and a, b positive constants, then

$$Pr\{T > a+b|T > a\} = \frac{Pr\{T > a+b\}}{Pr\{T > a\}} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = Pr\{T > b\}.$$

So it has the memoryless property. It is the only continuous distribution with this property. There are some similarities to the geometric probability distribution. The geometric distribution is the number of trials to first success while the exponential distribution represents the time to first event in a Poisson process. If y is geometric with parameter p, then $Pr\{Y > n\} = (1-p)^n$. To derive the Poisson process, set $p = \lambda\delta t = \frac{\lambda t}{n}$, having sub-divided $(0, t]$ into n pieces of length λ. Then $\{Y > n\}$ and $\{T > t\}$ are equivalent events, with

$$Pr\{T > t\} = \lim_{n\to\infty} Pr\{Y > n\} = \lim_{n\to\infty} (1 - \frac{\lambda t}{n})^n = e^{-\lambda t},$$

i.e, the exponential distribution is the limit of the geometric distribution function.

4.6 Time to r^{th} event

Let T_r be the time to occurrence of the r^{th} event of a Poisson process, $r \geq 1$. This random variable is analogous to a negative binomial random variable. Let t be some fixed number and consider $\{T_r > t\}$, the time to the r^{th} event greater than t. $\{T_r > t\}$ is the same as $\{X \leq r-1\}$ where X is the number of events in $(0, t]$, since T_r can only exceed t if there are $r-1$ or fewer events in $(0, t]$. Then X is poisson with parameter λt so

$$Pr\{T_r > t\} = Pr\{X \leq r-1\} = \sum_{k=0}^{r-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t},$$

with the distribution function for T_r

$$F_{T_r}(t) = Pr\{T_r \leq t\} = 1 - Pr\{T_r > t\} = 1 - \sum_{k=0}^{r-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

T_r is an **Erlang** random variable with parameters r, λ. It has density function

$$f_{T_r}(t) = \frac{d}{dt} F_{T_r}(t) = \frac{d}{dt}\left(1 - e^{-\lambda t} - \lambda t e^{-\lambda t} - \ldots - \frac{(\lambda t)^{r-1}}{(r-1)!} e^{-\lambda t}\right)$$

$$= \lambda e^{-\lambda t} - \lambda e^{-\lambda t} + \lambda^2 t e^{-\lambda t} - \lambda^2 t e^{-\lambda t} + \ldots - \frac{\lambda^r t^{r-1}}{(r-2)!} e^{-\lambda t} + \frac{\lambda^r t^{r-1}}{(r-1)!} e^{-\lambda t} = \frac{\lambda^t t^{r-1}}{(r-1)!} e^{-\lambda t},\ t > 0$$

$$= \frac{\lambda^r t^{r-1}}{\Gamma(r)} e^{-\lambda t},$$

where $\Gamma(\alpha) = \int\limits_0^\infty t^{\alpha-1} e^{-t} dt$. This distribution is a particular case of the gamma distribution. The time to the r^{th} occurrence in a Poisson distribution is in fact Gamma distributed with shape parameter r and rate λ.

Example Telephone calls to a call centre are a Poisson process with $\lambda = 120$ per hour. Starting at 9 a.m, let T_{10} be the time to the tenth call. Then T_{10} is gamma distributed with shape $r = 10$ and rate $2/min$. The expected time of the 10^{th} call is $E\{T_{10}\} = \frac{10}{2} = 5$, so 9.05 a.m. The probability the tenth call occurs before 9.05 is

$$Pr\{T_{10} < 5\} = 1 - \sum_{k=0}^{9} \frac{(5(2))^k}{k!} e^{-5(2)} = 1 - \sum_{k=0}^{9} \frac{10^k}{k!} e^{-10} = .542.$$

Similarly, the probability the tenth call is received between 9.05 and 9.07 is

$$Pr\{5 < T_{10} \leq 7\} = \left(1 - \sum_{k=0}^{9} \frac{14^k}{k!} e^{-14}\right) - \left(1 - \sum_{k=0}^{9} \frac{10^k}{k!} e^{-10}\right) = .349.$$

4.7 Inter-arrival Times

It has been seen that:

- The distribution of the time to the next event is exponential,
- Times between events is exponential.
- The time to the r^{th} event is gamma distributed.

It has been assumed that λ is constant, i.e a time homogeneous Poisson process. Letting $X(t)$ denote the number of elements in $(0, t]$, then the poisson process has 'independent increments'. Let $T_1, \ldots,$ denote the arrival times of the process and define $T_0 = 0$, then we have $X(T_1) - X(T_0), X(T_2) - X(T_1), \ldots$ are independent of each other, as $X(t + s) - X(s),\ t \geq 0$ is a poisson process with rate λ and is independent of $X(v)$, for $0 \leq v < s$.

4.8 General Poisson Process

Let $X(t)$ be the number of events in an interval $(0, t]$. Then $X(t)$ with rate $\lambda(t)$ is a poisson process if

1. $X(0) = 0$.
2. $X(t)$ has independent increments.
3. $X(t) - X(s)$ for $s < t$ is a poisson process with mean $\int\limits_s^t \lambda(v)dv$.

If $\lambda(v) = \lambda$, a constant, then the mean is just $X(t) - X(s) = \int\limits_s^t \lambda(v)dv = \lambda(t - s)$, the process we have seen already. For a time homogeneous process we need to show the time between arrivals follows an exponential distribution. In general, where $\lambda(t)$ depends explicitly on t, this isn't the case.

Let T_1 be the time to the first arrival, then

$$Pr\{T_1 > t\} = Pr\{X(t) = 0\} \text{ which is poisson, } \mu = \int\limits_0^t \lambda(v)dv$$

$$= \left[\int\limits_0^t \lambda(v)dv\right]^0 \frac{\exp\left(-\int\limits_0^t \lambda(v)dv\right)}{0!} = \exp\left(-\int\limits_0^t \lambda(v)dv\right).$$

For the distribution of T_1, the cdf is

$$F_{T_1}(t) = Pr\{T_1 \leq t\} = 1 - Pr\{T_1 > t\}$$

$$= 1 - \exp\left(-\int\limits_0^t \lambda(v)dv\right),$$

and so

$$f_{T_1}(t) = \frac{d}{dt}F_{T_1}(t) = \frac{d}{dt}\left[\int\limits_0^t \lambda(v)dv\right]\exp\left(-\int\limits_0^t \lambda(v)dv\right)$$

$$= \lambda(t)\exp\left(-\int\limits_0^t \lambda(v)dv\right).$$

Then, calling $\mu(t) = \int\limits_0^t \lambda(v)dv$, it is clear that in general, $f_{T_1}(t) = \lambda(t)e^{-\mu(t)}$ will not be an exponential distribution. When $\lambda(t)$ depends explicitly on t this is a time inhomogeneous process. That a poisson process satisfies the markov property in general follows from the independent increments property, but the markov property in continuous time needs to be defined.

For continuous time, we observe the process at arbitrary points in time, say,

$$0 = s_0 < s_1 < \ldots < s_k < s < t < t_1 < \ldots < t_n,$$

with states $i_0, i_1, \ldots, i_n, i, j, j_1, \ldots, j_n$. The markov property holds if

$$Pr\{X(t) = j, X(t_1) = j_1, \ldots, X(t_n) = j_n | X(s_0) = i_0, \ldots, X(s_n) = i_n, X(s) = i\}$$

$$= Pr\{X(t) = j, X(t_1) = j_1, \dots, X(t_n) = j_n | X(s) = i\}.$$

For the Poisson process,

$$Pr\{X(t) = j|X(s) = i\} = \frac{Pr\{X(t) = j, X(s) = i\}}{Pr\{X(s) = i\}}$$

$$= \frac{Pr\{X(t) - X(s) = j - i\}Pr\{X(s) = i\}}{Pr\{X(s) = i\}} \text{ by independent increments}$$

$$= Pr\{X(t) - X(s) = j - i\} = \frac{\left(\int\limits_s^t \lambda(r)dr\right)^{j-1} \exp\left(-\int\limits_s^t \lambda(r)dr\right)}{(j-1)!},$$

verifying it does satisfy the Markov property. For continuous time processes, we use $Pr\{X(t) = j|X(s) = i\}$ by $P_{s,t}(i, j)$.

4.9 Compound Poisson Processes

A compound poisson process associates an independent and identically distributed variable Y_i with each arrival of the process. The Y_i are assumed independent of the poisson process describing the arrivals, and are independent of each other.

Example Claims arriving to a large insurance company follow a poisson process and the size of each claim (Y_i) can be assumed to be independent. The compound process will be a measure of total liability. Considering the sum of all Y_i up to some time t, there will be $X(t)$ events of the poisson process, $Y_1, \dots, Y_{X(t)}$. $S(t) = Y_1 + \dots + Y_{X(t)}$ where $S(t) = 0$ if $X(t) = 0$.

For $Y_1, \dots, Y_{X(t)}$ i.i.d and $S(t) = \sum\limits_{i=1}^{X(t)} Y_i$ then we have the following results

- If $E\{Y_i\} < \infty$ and $E\{X(t)\} < \infty$ then
$$E\{S(t)\} = E\{X(t)\}E\{Y\}.$$
- If $E\{Y_i^2\} < \infty$ and $E\{X(t)^2\} < \infty$ then
$$Var\{S(t)\} = E\{X(t)\}Var\{Y\} + Var\{X(t)\}E\{Y^2\}.$$

Proof. When $X(t) = n$ then $S(t) = Y_1 + \dots + Y_n$ and $E\{S(t)\} = nE\{Y\}$. Breaking this down according to the value of $X(t)$,

$$E\{S(t)\} = \sum_{n=0}^{\infty} E\{S(t)|X(t) = n\} Pr\{X(t) = n\}$$
$$= \sum_{n=0}^{\infty} nE\{Y\} Pr\{X(t) = n\}$$
$$= E\{Y\}\sum_{n=0}^{\infty} nPr\{X(t) = n\}$$
$$= E\{Y\}E\{X(t)\}.$$

For the second statement, we have $Var\{S(t)\} = Var\{Y_1 + \dots + Y_n\} = nVar\{Y\}$. Hence,

$$E\{S(t)^2\} = \sum_{n=0}^{\infty} E\{S(t)^2|X(t) = n\}Pr\{X(t) = n\}$$

$$= \sum_{n=0}^{\infty} \left[nVar\{Y\} + E\{S(t)|X(t) = n\}^2\right] Pr\{X(t) = n\}$$

$$= \sum_{n=0}^{\infty} \left[nVar\{Y\} + n^2E\{Y\}^2\right] Pr\{X(t) = n\}$$
$$= Var\{Y\}E\{X(t)\} + E\{Y\}^2E\{X(t)\}^2,$$

then,

$$Var\{S(t)\} = E\{S(t)^2\} - E\{S(t)\}^2$$
$$= Var\{Y\}E\{X(t)\} + E\{Y\}^2E\{X(t)^2\} - E\{Y\}^2E\{X(t)\}^2$$
$$= Var\{Y\}E\{X(t)\} + E\{Y\}^2Var\{X(t)\}.$$

□

5 Other Continuous Time Processes

5.1 Brownian Motion

Looking again at the symmetric random walk, which has equal probability of going up or down. Let S_n be the sum of all previous steps. If we consider smaller and smaller time intervals and smaller and smaller increments up and down, this becomes a continuous time process. Consider intervals of length δt, with steps of size δx. Let $X(t)$ be the value of the process at time t, with $n = \frac{t}{\delta t}$ time intervals. Then

$$X(t) = \delta x X_1 + \ldots + \delta x X_{\left[\frac{t}{\delta t}\right]} = \delta x \left[X_1 + \ldots + X_{\left[\frac{t}{\delta t}\right]}\right].$$

and consider the mean and variance of $X(t)$.

$$E\{X(t)\} = \delta x \left(\frac{t}{\delta t}\right) E\{X_i\} = 0 \text{ as } E\{X_i\} = 0$$

$$Var\{X(t)\} = (\delta x)^2 \left(\frac{t}{\delta t}\right) Var\{X_i\} = (\delta x)^2 \left(\frac{t}{\delta t}\right) \text{ as } E\{X_i^2\} = 1.$$

Then, taking the limits as δx and δt go to 0. Let $\delta x = c\sqrt{\delta t}$, with c some positive constant. Then $Var\{X(t)\} = c^2 t$. In the limit, this process is **Brownian Motion**. It has the following properties:

1. Since $X(t) = \delta x \left[X_1 + \ldots + X_{\left[\frac{t}{\delta t}\right]}\right]$ by the central limit theorem $X(t)$ follows a normal distribution with mean 0 and variance $c^2 t$.

2. As the distribution of the change in position of the random walk is independent over non overlapping time intervals $\{X(t),\ t \geq 0\}$ has independent increments.

3. The process also has stationary increments, since the change in the process value, $X(t) \sim N(0, c^2 t)$ over a time interval depends only on the length of the interval. For $c = 1$, this process is often called the Weiner process. The independent increments assumption implies that $X(t+s) - X(s)$ is independent of the process values before time s.

$$Pr\left\{X(t+s) \leq a | X(s) = x, X(u), 0 \leq u \leq s\right\}$$
$$= Pr\left\{X(t+s) - X(s) \leq a - x | X(s) = x, X(u), 0 \leq u \leq s\right\} \text{ by independent increments}$$
$$= Pr\{X(t+s) - X(s) \leq a - x\} = Pr\{X(t+s) = a | X(s) = a\},$$

   so this tells us that Brownian motion satisfies the Markov property.

Let $X(t)$ be standard Brownian motion, so $X(t) \sim N(0, t)$ and the density of $X(t)$ is

$$f_t(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2t}}.$$

Since it has stationary and independent increments, the joint distribution of $X(t_1), \ldots, X(t_n)$ is

$$f(x_1, \ldots, x_n) = f_{t_1}(x_1) f_{t_2 - t_1}(x_2 - x_1) \ldots f_{t_n - t_{n-1}}(x_n - x_{n-1}) \text{ for } t_1 \leq \ldots \leq t_n$$

which can be used to compute properties of Brownian motion. For example, the conditional distribution of $X(s)$ given $X(t) = B$, with $s < t$ is

$$f_{s|t}(x|B) = \frac{f_{s,t}(x, B)}{f_t(B)} \text{ where } f_t(x) = \frac{1}{\sqrt{2\pi t}} e^{\frac{-x^2}{2t}}$$

$$= \frac{f_s(x) f_{t-s}(B - x)}{f_t(B)} = \frac{\frac{1}{\sqrt{2\pi s}} e^{\frac{-x^2}{2s}} \frac{1}{\sqrt{2\pi(t-s)}} e^{\frac{-(B-x)^2}{2(t-s)}}}{\frac{1}{\sqrt{2\pi t}} e^{\frac{-B^2}{2t}}}$$

$$= \frac{1}{\sqrt{2\pi \frac{s(t-s)}{t}}} \exp\left\{\frac{-1}{2}\left[\frac{x^2}{s} + \frac{(B-x)^2}{t-s} - \frac{B^2}{t}\right]\right\}$$

$$= \frac{1}{\sqrt{2\pi \frac{s(t-s)}{t}}} \exp\left\{\frac{-1}{2}\left[\frac{x^2}{s} + \frac{B^2 - 2xB + x^2}{t-s} - \frac{B^2}{t}\right]\right\}.$$

$$= \frac{1}{\sqrt{2\pi \frac{s(t-s)}{t}}} \exp\left\{\frac{-1}{2}\left[\left(\frac{1}{s}+\frac{1}{t-s}\right)x^2 - \frac{2B}{t-s}x + B^2\left(\frac{1}{t-s}-\frac{1}{t}\right)\right]\right\}$$

$$= \frac{1}{\sqrt{2\pi \frac{s(t-s)}{t}}} \exp\left\{\frac{-1}{2}\left[\frac{t}{s(t-s)}x^2 - 2Bx\frac{s}{t} + \frac{s^2B^2}{t^2}\right]\right\}$$

$$= \frac{1}{\sqrt{2\pi \frac{s(t-s)}{t}}} \exp\left\{\frac{-1}{2\frac{s(t-s)}{t}}\left(x - \frac{sB}{t}\right)^2\right\},$$

the density of a Normal distribution with mean $\frac{Bs}{t}$ and variance $\frac{s(t-s)}{t}$, which says

$$E\{X(s)|X(t)=B\} = \frac{Bs}{t} \text{ and } Var\{X(s)|X(t)=B\} = \frac{s(t-s)}{t}$$

noting the variance does not depend on B. For $\alpha = \frac{s}{t}$ then for $0 < \alpha < 1$ it has mean $\alpha X(t)$ and variance $\alpha(1-\alpha)t$.

If we consider process values only between 0 and 1 and conditional on $X(1) = 0$, then the process is a Brownian bridge.

5.2 Gaussian Processes

A stochastic process is called a Gaussian process if $X(t_1), \ldots, X(t_n),\ t_1 < \ldots < t_n$ has a multivariate normal distribution for all $t_1, \ldots, t_n$, which is defined for a random vector $\vec{X} = (X(t_1), \ldots, X(t_n))$ by

$$f_{\vec{x}}(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{\frac{-1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right\}$$

with Σ the $n \times n$ covariance matrix and μ the mean vector.

Example If $X_1, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$ then

$$\Sigma = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$$

and $\vec{\mu} = (\mu, \ldots, \mu)$ and $|\Sigma| = (\sigma^2)^n$ and also $\Sigma^{-1} = diag(\frac{1}{\sigma^2}, \ldots, \frac{1}{\sigma^2})$. Then

$$(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu}) = \frac{1}{\sigma^2}(\vec{x}-\vec{\mu})^T I (\vec{x}-\vec{\mu})$$

$$= \frac{1}{\sigma^2}(\vec{x}-\vec{\mu})^T (\vec{x}-\vec{\mu})$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\Rightarrow \quad f_{\vec{x}}(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

The joint density of $X(t_1), \ldots, X(t_n)$ with Brownian motion was

$$f(x_1, \ldots, x_n) = f_{t_1}(x_1) f_{t_2 - t_1}(x_2 - x_1) \ldots f_{t_n - t_{n-1}}(x_n - x_{n-1}),$$

showing Brownian motion is a Gaussian process.

5.3 Brownian motion with drift

$\{X(t), t \geq 0\}$ is Brownian motion with drift coefficient μ if

1. $X(0) = 0$.
2. $\{X(t), t \geq 0\}$ has stationary and independent increments.
3. $X(t)$ is normally distributed with mean μt and variance t.

So it can be written as

$$X(t) = \mu t + W(t)$$

where $W(t)$ is standard Brownian motion.

6 Applications, Model Estimation through Markov Chain Monte Carlo

6.1 Likelihood and Maximum Likelihood

If a specific probability law or distribution is assumed for observed data, then a likelihood function can be formed. Maximum likelihood finds the parameter values which maximize the likelihood. Assume $X_1, \dots, X_n$ are a random sample of a random variable X, which we assume has density $f(x|\theta)$, where θ are the unknown parameter(s). If X is discrete then it is a probability mass function. Then the likelihood function is

$$\pi(x|\theta) = f(x_1|\theta) \dots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

which can be thought of as the probability of observing the given random sample with parameters θ.

Example Suppose the time to failure of a component is exponentially distributed. A sample of n failure times is $\vec{x} = (x_1, \dots, x_n)$ and then the likelihood function is

$$\pi(\vec{x}|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum\limits_{i=1}^{n} x_i}$$

Maximum likelihood involves maximising the likelihood function with respect to the unknown parameter θ. Normally easier to work with the log-likelihood.

$$\log \pi(\vec{x}|\theta) = \log \left(\prod_{i=1}^{n} f(x_i|\theta) \right) = \sum_{i=1}^{n} \log f(x_i|\theta),$$

and then take the gradient of this and set it equal to zero.

$$\nabla_\theta \log \pi(\vec{x}|\theta) = 0.$$

The value of θ which satisfies this, $\hat{\lambda}$ is the max likelihood estimator (M.L.E).

$$\log \pi(\vec{x}|\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d}{d\lambda} \log \pi(\vec{x}|\lambda) = \frac{n}{\lambda} - \sum x_i$$

$$\Rightarrow \frac{n}{\hat{\lambda}} - \sum x_i = 0 \text{ so } \frac{1}{\hat{\lambda}} = \frac{\sum x_i}{n} = \bar{x}.$$

Example Assume $X_1, \dots, X_n \sim Bernoulli(p)$.Find the MLE of p.
Let $\vec{X} = (x_1, \dots, x_n)$. $f(x|p) = p^x (1-p)^{1-x}$.

$$\pi(\vec{x}|p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum x_i} (1-p)^{\sum (1-x_i)}$$

$$\log \pi(\vec{x}|p) = \sum x_i \log p + (n - \sum x_i) \log(1-p)$$

$$\frac{d}{dp} = \frac{\sum x_i}{\hat{p}} - \frac{n - \sum x_i}{1 - \hat{p}} = 0$$

$$\frac{\hat{p}}{1-\hat{p}} = \frac{\sum x_i}{n - \sum x_i} \Rightarrow \frac{1-\hat{p}}{\hat{p}} = \frac{n - \sum x_i}{\sum x_i}$$

$$\Rightarrow \frac{1}{\hat{p}} - 1 = \frac{n}{\sum x_i} - 1 \Rightarrow \hat{p} = \frac{\sum x_i}{n} = \bar{x}.$$

6.2 Prior Distribution

In finding the maximum likelihood estimates in the previous section, only the observed sample values $x_1, \ldots, x_{\vec{n}}$ were used to construct estimates of $\vec{\theta}$. Maximum likelihood doesn't require any other information to estimate $\vec{\theta}$ other than the sample values. If we had some prior information, we could not incorporate it. However, we can use information available to inform a prior distribution for $\vec{\theta}$ and then use a Bayesian approach for estimation. The prior distribution of a parameter $\vec{\theta}$ is a probability function or density expressing our degree of belief about the value of $\vec{\theta}$ prior to observing a sample of a random variable X whose distribution function depends on $\vec{\theta}$. The prior distribution makes use of any information available beyond what's observed in a random sample.

Example Consider a brand new coin and we wish to estimate θ, the probability of a head. We know $\theta \in [0, 1]$. A prior for θ could be that it is uniform from 0 to 1, i.e that all values are equally likely. Alternatively, we may be justified in assuming a priori $\theta \in (.4, .6)$ if the coin appears symmetric. Then the prior is

$$\pi(\theta) = \begin{cases} 5, \ \theta \in (.4, .6) \\ 0 \text{ otherwise} \end{cases}$$

which corresponds to the belief that any values in $(.4, .6)$ is equally likely. Finally, we might only let θ have values $.4, .5, .6$, with $.5$ being twice as likely, giving the prior

$$\pi(\theta) = \begin{cases} \frac{1}{4}, \ \theta = .4, .6 \\ \frac{1}{2}, \ \theta = .5 \\ 0 \text{ otherwise} \end{cases}$$

Note the priors are different and depend on the assumptions we're willing to make about θ. These are often influenced by expert opinion.

The prior choice is subjective. The final result of a Bayes technique is generally dependent on the prior assumed.

6.3 Posterior Distribution

Having obtained a sample $\vec{X} = (x_1, \ldots, x_n)$ we can then get the likelihood for $\vec{x}$ given the value of $\vec{\theta}$.

$$\text{Likelihood} = \pi\left(\vec{x}|\vec{\theta}\right) = \prod_{i=1}^{n} f\left(x_i|\theta\right).$$

By taking a prior on θ we are in essence acting as if the probability law of X is itself a random variable, through its dependence on θ. Hence, we speak of the likelihood as the distribution of $\vec{x}$ conditional on $\vec{\theta}$. Given a prior density for $\theta, \pi(\theta)$ and the conditional density of the elements of the sample (likelihood), $\pi(x|\theta)$, the joint density for the sample and the parameters is simply the product of these two functions.

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

This is the product of the likelihood and the prior. Then the marginal density of the sample values, which is independent of θ is given by the integral of the joint density over the space Θ. Thus

$$\pi(x) = \int\limits_{\Theta} \pi(x, \theta) d\theta = \int\limits_{\Theta} \pi(x|\theta)\pi(\theta) d\theta,$$

which is called the marginal or the marginal likelihood of the sample. The posterior density for θ is the conditional density of θ given the sample values. Thus

$$\pi(\theta|x) = \frac{\pi(x, \theta)}{\pi(x)} = \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)}.$$

The prior density expresses our degree of belief about θ before any experiments while the posterior expresses our beliefs given results of a sample. The marginal likelihood $\pi(x)$ is the normalising constant of $\pi(x|\theta)\pi(\theta)$ so that

$$\int\limits_{\Theta} \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)} d\theta = 1$$

The marginal doesn't depend explicitly on θ. This is often written

$$\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta) \text{ or } \textit{Posterior} \propto \textit{Likelihood} \times \textit{Prior}.$$

Often $\pi(x)$ is not available analytically, but is found using numerical methods, such as MCMC.

Example Suppose $X_1, \dots, X_n$ are i.i.d and are all $N(\mu, \sigma^2)$. Assume a prior for μ which is $N(\xi, \tau^2)$ and a prior for σ^2 which is $InvGamma(\alpha, \beta)$. Then, if $Y\ Gamma(\alpha, \beta)$, α is the shape parameter and β is the rate paramenter. Then $\frac{1}{Y}\ InvGamma(\alpha, \beta)$ and $f_Y(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{-(\alpha+1)} e^{\frac{-\beta}{t}}$. Then we have

$$F_Y(t) = Pr\{Y \le t\} \text{ and } F_{\frac{1}{Y}}(t) = Pr\left\{\frac{1}{Y} \le t\right\} = Pr\left\{Y \ge \frac{1}{t}\right\}$$

$$= 1 - Pr\left\{Y < \frac{1}{t}\right\} = 1 - F_Y\left(\frac{1}{t}\right).$$

And so,

$$f_{\frac{1}{Y}}(t) = \frac{d}{dt}F_{\frac{1}{Y}}(t) = -\frac{d}{dt}F_Y\left(\frac{1}{t}\right) = -\frac{(-1)}{t^2}f_Y\left(\frac{1}{t}\right),$$

$$= \frac{1}{t^2}\frac{\beta^\alpha}{\Gamma(\alpha)}t^{-\alpha+1}e^{\frac{-\beta}{t}} = \frac{\beta^\alpha}{\Gamma(\alpha)}t^{-(\alpha+1)}e^{\frac{-\beta}{t}}.$$

So then we have

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{\frac{-1}{2\tau^2}(\mu - \xi)^2\right\}$$

and

$$\pi(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)}e^{\frac{-\beta}{\sigma^2}}.$$

The likelihood is

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

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

We know the posterior distribution is proportional to the product of the likelihood and the prior, so

$$\pi(\mu, \sigma^2|x) \propto \pi(x|\mu, \sigma^2)\pi(\mu)\pi(\sigma^2)$$

$$\propto \left(2\pi\sigma^2\right)^{\frac{-n}{2}} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^n(x_i - \mu)^2\right\} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{\frac{-1}{2\tau^2}(\mu - \xi)^2\right\} \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)}e^{\frac{-\beta}{\sigma^2}}$$

$$\propto (\sigma^2)^{-\alpha+1-\frac{n}{2}} \exp\left\{\frac{-1}{2\sigma^2}\left[\sum_{i=1}^n(x_i - \mu)^2 + 2\beta\right]\right\} \exp\left\{\frac{-1}{2\tau^2}(\mu - \xi)^2\right\}$$

$$\propto (\sigma^2)^{\frac{-n}{2}-\alpha+1} \exp\left\{\frac{-1}{2}\left[\frac{1}{\sigma^2}\left(\sum x_i^2 + 2\mu\sum x_i + \mu^2\right) + \frac{2\beta}{\sigma^2} + \frac{1}{\tau^2}\left(\mu^2 - 2\mu\xi + \xi^2\right)\right]\right\}$$

$$\propto (\sigma^2)^{\frac{-n}{2}-\alpha+1} \exp\left\{\frac{-1}{2}\left[\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)\mu^2 + 2\left(\sum x_i - \xi\right)\mu + \frac{\sum x_i}{\sigma^2} + \frac{2\beta}{\sigma^2} + \frac{\xi^2}{\tau^2}\right]\right\}.$$

We can compute the marginal likelihood, $\pi(x)$ in this case, however it is rarely possible. This is because $X_1, \dots, X_n$ are normally distributed and we have priors.

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

$$= (2\pi\sigma^2)^{\frac{-n}{2}} \exp\left\{\frac{-1}{2\sigma^2}\sum(x_i - \mu)^2\right\},$$

and since $\pi(\mu) = (2\pi\tau^2)^{\frac{-1}{2}} \exp\left\{\frac{-1}{2\tau^2}(\mu - \xi)^2\right\}$, we have

$$\pi(x) = \int_{-\infty}^{\infty} \pi(x|\mu, \sigma^2)\pi(\mu)\ d\mu$$

$$= \int_{-\infty}^{\infty} (2\pi\sigma^2)^{\frac{-n}{2}}(2\pi\tau^2)^{\frac{-n}{2}} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^n(x_i - \mu)^2 - \frac{1}{2\tau^2}(\mu - \xi)^2\right\}\ d\mu$$

$$= c\int\limits_{-\infty}^{\infty} \exp\left\{\frac{-1}{2\sigma^2}\left[\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right] - \frac{1}{2\tau^2}\left[\mu^2 - 2\xi\mu + \xi^2\right]\right\}\, d\mu$$

$$= c\int\limits_{-\infty}^{\infty} \exp\left\{\frac{-1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\mu^2 - 2\left(\frac{\sum x_i}{\sigma^2}+\frac{\xi}{\tau^2}\right)\mu + \frac{\sum x_i^2}{\sigma^2}+\frac{\xi^2}{\tau^2}\right]\right\}\, d\mu,$$

and then, as the final two terms in the exponential are constant this becomes

$$= c_0\int\limits_{-\infty}^{\infty} \exp\left\{\frac{-\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)}{2}\left[\mu^2 - 2\frac{\left(\frac{\sum x_i}{\sigma^2}+\frac{\xi}{\tau^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)}\mu\right]\right\}\, d\mu.$$

This can then be rearranged and we can complete the square, to get

$$c_0\int\limits_{-\infty}^{\infty} \exp\left\{\frac{-\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)}{2}\left(\left[\mu - \frac{\left(\frac{\sum x_i}{\sigma^2}+\frac{\xi}{\tau^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)}\right]^2 + \left[\frac{\left(\frac{\sum x_i}{\sigma^2}+\frac{\xi}{\tau^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)}\right]^2\right)\right\}\, d\mu,$$

which can be further simplified to the integral over a normal multiplied by a constant.

6.4 Posterior Quantities of Interest

There are many quantities of interest we may want to get from a Bayesian analysis. For example, the mean of the posterior distribution, θ^* is a widely used Bayesian estimator. The mode of the posterior $\tilde{\theta}$ is called the maximum a posteriori estimate of θ. If θ is of dimension p, $(\theta_1, \ldots, \theta_p)$ we may be interested in the marginal density of θ_j.

$$\pi\left(\theta_j | \theta_{-j}, x\right) = \int\limits_{\theta-j} \pi\left(\theta | x\right) d\theta_{-j},\ j = 1, \ldots, p$$

where θ_{-j} is θ with the j^{th} element removed. Consider the posterior expectation of θ^*,

$$\theta^* = \int\limits_{\Theta} \theta\pi(\theta|x)d\theta = E_{\theta|x}\{\theta\}$$

$$\int\limits_{\Theta} \theta\frac{\pi(x|\theta)\pi(\theta)}{\pi(x)}d\theta.$$

This calculation requires knowing $\pi(x)$ which is intractable normally. This occurs in every problem. We aim to simulate values of θ, say $\theta^{(1)}, \ldots \theta^{(N)}$ from $\pi(\theta|x)$. Instead of doing these integrals analytically, we can approximate them numerically,

$$E_{\theta|x}\{\theta\} = \int\limits_{\Theta} \theta\pi(\theta|x)d\theta \approx \frac{1}{N}\sum_{k=1}^{N}\theta^{(k)}.$$

We can use this approach to approximate the prior expectation of any function of $\theta g(\theta)$.

$$E_{\theta|x}\{g(\theta)\} = \int\limits_{\Theta} g(\theta)\pi(\theta|x)d\theta \approx \frac{1}{N}\sum_{k=1}^{N} g\left(\theta^{(k)}\right).$$

The main idea of MCMC is to approximately generate samples from $\pi(\theta|x)$ and use these to approximate integrals.

6.5 MCMC

Wish to generate samples from $\pi(\vec{\theta}|x)$ which can't be done directly. Suppose we can construct a Markov chain (through its transition probabilities with state space $\Theta = \{$ all $\theta\}$) which can be done easily as it has a stable distribution which is the posterior $\pi(\vec{\theta}|\vec{x})$. Set up a Markov chain $\theta^{(0)}, \theta^{(1)}, \ldots$ with transition probabilities (transition kernel) such that $\pi(\vec{\theta}|\vec{x})$ is the stable distribution. Asymptotic results exist which clarify how a sample from a chain with stable distribution $\pi(\vec{\theta}|\vec{x})$ can be used to estimate $g(\theta)$, a function of interest.

If $\theta^{(0)}, \theta^{(1)}, \ldots$ is a realization from an approximate chain, typically

$$\vec{\theta}^{(t)} \to \vec{\theta} \sim \pi(\vec{\theta}|\vec{x}) \text{ in distribution, and}$$

$$\frac{1}{t}\sum_{k=1}^{t} g\left(\vec{\theta}^{(k)}\right) \rightarrow \; E_{\vec{\theta}|\vec{x}}\left\{g\left(\vec{\theta}\right)\right\} \text{ as } t \rightarrow \infty$$

Successive values of $\vec{\theta}^{(t)}$ will be correlated, so we may need to account for this if imagining that the $\vec{\theta}^{(t)}$'s are i.i.d from $\pi(\vec{\theta}|\vec{x})$.

6.6 The Gibbs Sampling Algorithm

Initially introduced by Julian Besag.
Let $\theta = (\theta_1, \ldots, \theta_p)$ and we wish to obtain inferences from $\pi(\theta|x)$, but sampling is difficult. We can recast the problem as one of iterative sampling from appropriate conditional distributions. Consider the full conditional densities

$$\pi\left(\theta_j|x, \theta_{-j}\right), \; j = 1, \ldots, p,$$

where θ_{-j} is as defined above. These are densities of the individual components given the data and the specified values of the other components of θ. They are typically standard densities like normal or gamma.

Suppose we have an arbitrary set of starting values $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_p^{(0)}\right)$ and then implement the following iterative process:

1. For the first iteration, draw

$$\begin{gathered} \theta_1^{(1)} \text{ from } \pi\left(\theta_1|\theta_2^{(0)}, \ldots, \theta_p^{(0)}, x\right) \\ \theta_2^{(1)} \text{ from } \pi\left(\theta_1|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)}, x\right) \\ \vdots \\ \theta_p^{(1)} \text{ from } \pi\left(\theta_p|\theta_1^{(1)}, \ldots, \theta_{p-1}^{(1)}, x\right) \end{gathered} .$$

2. Then for the second iteration, draw

$$\begin{gathered} \theta_1^{(2)} \text{ from } \pi\left(\theta_1|\theta_2^{(1)}, \ldots, \theta_p^{(1)}, x\right) \\ \vdots \end{gathered} .$$

If this procedure is continued through t iterations, we get the sampled vector

$$\theta^{(t)} = \left(\theta_1^{(t)}, \ldots, \theta_p^{(t)}\right),$$

which is a realisation of a Markov chain with transition probabilities

$$p\left(\theta^{(t)}, \theta^{(t+1)}\right) = \prod_{j=1}^{p} \pi\left(\theta_j^{(t+1)}|\theta_l^{(t+1)} \text{ for } l < j \text{ or } \theta_l^{(t)} \text{ for } l > j, x\right).$$

Then as $t \rightarrow \infty$, $\left(\theta_1^{(t)}, \ldots, \theta_p^{(t)}\right)$ tends in distribution to a random variable whose joint density is $\pi\left(\theta|x\right)$. In particular, $\theta_j^{(t)}$ tends in distribution to a random quantity whose density is $\pi\left(\theta_j|x\right)$.

Example The Gibbs sampler is often used in finite mixture models which are used for model based clustering. For Gaussian finite mixtures the density of an observation x is given by

$$f_x(x) = \sum_{g=1}^{G} w_g f(x|\mu_g, \sigma_g^2),$$

where w_g are the mixture weights and $\sum\limits_{g=1}^{G} w_g = 1$, and $f(x|\mu_g, \sigma_g^2) \sim N(\mu_g, \sigma_g^2)$. The likelihood for n observations $x_1, \ldots, x_n$ is

$$\pi(x|\theta) = \prod_{i=1}^{n}\left(\sum_{g=1}^{G} w_g f(x_i|\mu_g, \sigma_g^2)\right).$$

The likelihood is very difficult to work with so we usually observe the data with components labelled $z = (z_1, \ldots, z_n)$, which tell us which component each observation belongs to, i.e $z_i = g$, and the x_i arises from $N(\mu_g, \sigma_g^2)$.

Of course the labels give the clustering of the data, but can't be observed directly. We can include these as unknowns in the Gibbs sampler. Then the likelihood of the complete data is

$$\pi\left(x,|\theta\right)=\prod_{g=1}^{G}\prod_{i:z_i=g}\frac{w_g}{\sqrt{2\pi\sigma_g^2}}\exp\left(-\frac{(x_i-\mu_g)^2}{2\sigma_g^2}\right)$$

$$=\prod_{g=1}^{G}w_g^{ng}(2\pi\sigma_g^2)^{-\frac{ng}{2}}\exp\left(\frac{-1}{2\sigma_g^2}\sum_{i:z_i=g}\frac{(x_i-\mu_g)^2}{2\sigma_g^2}\right),$$

where $ng =$ the number of $i's$ such that $z_i = g$.

We then have a mixture model

$$f(x)=\sum_{g=1}^{G}w_g f(x|\mu_g,\sigma_g^2).$$

The priors are weights. The standard assumption is to assume the weights follow a Dirichlet distribution which is given by

$$\pi(w_1,\ldots,w_g)=\frac{\Gamma(\delta+\ldots+\delta)}{\Gamma(\delta)+\ldots+\Gamma(\delta)}w_1^{\delta-1}\ldots w_G^{\delta-1}=\frac{\Gamma(G\delta)}{\Gamma(\delta)^G}\prod_{g=1}^{G}w_g^{\delta-1}.$$

Usually one assumes that the means μ_g arise from a $N(\xi,\tau^2)$ a priori and do so independently.

$$\pi(\mu_1,\ldots,\mu_G)=\prod_{g=1}^{G}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(\frac{-1}{2\tau^2}(\mu_g-\xi)^2\right).$$

Finally we assume that the variances arise from an inverse gamma distribution independently, so that

$$\pi(\sigma_1^2,\ldots,\sigma_G^2)=\prod_{g=1}^{G}\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma_g^2)^{-(\alpha+1)}\exp\left(\frac{-\beta}{\sigma_g^2}\right).$$

Then we have

$$\pi\left(\theta,z|x\right)\propto\pi(x|\theta)\pi(\theta)$$

$$\propto\prod_{g=1}^{G}w_g^{n_g}(2\pi\sigma_g^2)^{\frac{-ng}{2}}\exp\left(\frac{-1}{2\sigma_g^2}\sum_{i:z_g=g}(x_i-\mu_g)^2\right)W_g^{\delta-1}\exp\left(\frac{-1}{2\tau^2}(\mu_g-\xi)^2\right)(\sigma_g^2)^{-(\alpha+1)}\exp\left(\frac{-\beta}{\sigma_g^2}\right).$$

The next step is to implement a Gibbs sampler for this model to derive the full conditionals. We want to iteratively sample the labels, weights, means and variances.

$$Pr\left(z_i=k|everything\ else\right)\propto W_k(2\pi\sigma_k^2)^{\frac{-1}{2}}\exp\left(\frac{-1}{2\sigma_k^2}(x_i-\mu_k)^2\right)$$

$$\propto\frac{W_k}{\sigma_k^2}\exp\left(\frac{-1}{2\sigma_k^2}(x_i-\mu_k)^2\right).$$

We compute this for each $k = 1,\ldots,G$ then renormalise to get a discrete distribution for the label which we can sample from.

$$\pi(W_1,\ldots,W_G|rest)\propto\prod_{g=1}^{G}W_g^{n_g+\delta-1},$$

which is the form of a Dirichlet $(n_1+\delta,\ldots,n_g+\delta)$ distribution. So we have

$$\pi(\mu_g|everything\ else)\propto\exp\left(\frac{-1}{2\sigma_g^2}\sum_{i:z_i=g}(x_i-\mu_g)^2\frac{-1}{2\tau^2}(\mu_g-\xi)^2\right)$$

$$\propto\exp\left(\frac{-1}{2}\left[\left(\frac{n_g}{\sigma^2}+\frac{1}{\tau^2}\right)\mu_g^2-2\left(\frac{\sum_{i:z_i=g}x_i}{\sigma_g^2}+\frac{\xi}{\tau^2}\right)\mu_g\right]\right)$$

$$\propto\exp\left(\frac{-\left(\frac{n_g}{\sigma_g^2}+\frac{1}{\tau^2}\right)}{2}\left[\mu_g-\frac{\sum_{i:z_i=g}\frac{x_i}{\sigma_g^2}+\frac{\xi}{\tau^2}}{\frac{n_g}{\sigma_g^2}+\frac{1}{\tau^2}}\right]^2\right).$$

So the full conditional for μ_g is

$$N\left(\frac{\sum\limits_{i:z_i=g}\frac{x_i}{\sigma_g^2}+\frac{\xi}{\tau^2}}{\frac{n_g}{\sigma_g^2}+\frac{1}{\tau^2}}\quad\frac{1}{\frac{n_g}{\sigma_g^2}+\frac{1}{\tau^2}}\right).$$

Finally the full conditional for σ_g^2 is

$$\pi(\sigma_g^2|everything\ else)\propto(\sigma_g^2)^{-(\frac{n}{2}+\alpha+1)}\exp\left(\frac{-1}{\sigma_g^2}\left[\frac{1}{2}\sum_{i:z_i=g}(x_i-\mu_g)^2+\beta\right]\right),$$

which is $InvGamma\left(\frac{n}{2}+\alpha,\frac{1}{2}\sum\limits_{i:z_i=g}(x_i-\mu_g)^2+\beta\right)$.

6.7 The Metropolis-Hastings Algorithm

This constructs a markov chain $\theta^{(1)},\dots,\theta^{(t)}$, by defining the transition probability from $\theta^{(t)}$ to $\theta^{(t+1)}$ as follows:

Let $q(\theta,\theta')$ denote a proposal distribution such that if $\theta=\theta^{(t)}$ then θ' is a proposed next value for $\theta^{(t+1)}$. However a further randomization then takes place. with some probability $\alpha(\theta,\theta')$ we actually accept $\theta^{(t+1)}=\theta^{(t)}$. This construction defines a Markov chain with transition probabilities given by

$$p(\theta,\theta')=q(\theta,\theta')\alpha(\theta,\theta')+\Pi(\theta'=\theta)\left[1-\int q(\theta,\theta'')\alpha(\theta,\theta'')\,d\theta''\right],$$

where Π is an indicator function.

If not we set

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

and then one can show that

$$\pi(\theta|x)q(\theta,\theta')=\pi(\theta'|x)q(\theta',\theta).$$

This is called detailed balance and it is a sufficient condition to ensure that $\pi(\theta|x)$ is the stable distribution of the chain. In practice we generally assume $q(\theta,\theta')$ is a normal distribution which is $N(\theta,\sigma_{prop}^2I)$. The behaviour of the chain will depend on the value of σ_{prop}^2, but generally we tune this to give 25% $\sim$ 40%.

Example Consider an auto-regressive process of order 1, i.e AR(1), defined by

$$X_{t+1}=\phi X_t+\epsilon_t$$

where $\epsilon\sim N(0,\sigma^2)$ and $X_0\sim N(0,1)$. Consider a realization of this process $x_0,\dots,x_n$. Then the likelihood for $\phi,\sigma^2,\pi\left(x|\phi,\sigma^2\right)$ is

$$\frac{1}{\sqrt{2\pi}}\exp\left\{\frac{-X_0^2}{2}\right\}\prod^n\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-1}{2\sigma^2}\left(x_t-\phi x_{t-1}\right)^2\right\}.$$

To ensure stationarity we enforce that $\phi\in(0,1)$ and $|\phi|<1$. We take a uniform $(-1,1)$ prior on ϕ and an Inverse Gamma $(1,.01)$ prior on σ^2.

$$\pi\left(\phi\sigma^2|x\right)\propto(\sigma^2)^{\frac{n}{2}+1+1}\exp\left\{\frac{-1}{2\sigma^2}\sum_{t=1}^n\left(x_t-\phi x_{t-1}\right)^2-\frac{0.01}{0.2}\right\}.$$

If we look at the full conditional of ϕ it looks like a Gaussian, but we're restricting $\phi\in(-1,1)$ so we need to sample from a truncated Gaussian which is hard.

We'll use Metropolis Hastings to update ϕ and then use a Gibbs step to update σ^2, so this is a hybrid algorithm, which works as both updates preserve the Markov property.

To update ϕ we'll use a Gaussian proposal, centred at the current value with standard deviation σ_{prop}^2 so then $\phi'\sim N\left(\phi^{(t)},\sigma_{prop}^2\right)$. We tune σ_{prop}^2 to give acceptances between 25% and 40%. This new value is accepted as the next value in the chain with probability

$$\alpha=min\left(1,\frac{\exp\left\{\frac{-1}{2\sigma^2}\sum\limits_{t=1}^n(x_t-\phi' x_{t-1})^2\right\}\frac{1}{\sqrt{2\pi\sigma_{prop}^2}}\exp\left\{\frac{-1}{2\sigma_{prop}^2}(\phi-\phi')^2\right\}}{\exp\left\{\frac{-1}{2\sigma^2}\sum\limits_{t=1}^n(x_t-\phi x_{t-1})^2\right\}\frac{1}{\sqrt{2\pi\sigma_{prop}^2}}\exp\left\{\frac{-1}{2\sigma_{prop}^2}(\phi-\phi')^2\right\}}\right).$$

The case where we use a Gaussian proposal is called the random walk Metropolis algorithm. Then σ^2 is updated by drawing from the full conditional $InvGamma\left(\frac{n}{2}+1,\frac{1}{2}\sum\limits_{t=1}^n(x_t-\phi x_{t-1})^2+0.01\right)$.