BAYES METHODS: DRAFT LECTURE NOTES HT18

GEOFF K NICHOLLS

Course: SC7/SM6 Bayes Methods HT18
Teachers: Geoff Nicholls (Lect & Tutor) Chris Carmona (Tutor & TA) Soufiane Hayou (TA)
Material: Notes and Problem sheets are available at
http://www.stats.ox.ac.uk/~nicholls/BayesMethods/
and via the MSc weblearn pages.
Lectures: Sixteen lectures in LLT.
Classes: Part C classes: LG.05, Thurs 2-3:30/Fri 10:30-12, week 3,5,7, TT1 (due
Tues 5pm at department)
MSc classes: LLT, Thursday 5-6pm weeks 3,5,7,TT0/1
MSc practicals: in the IT teaching Suite, Friday 11-1pm weeks 2, 5 (assessed)
and 7.

There follow some draft lecture notes for the MSc and Part C Bayes Methods course. These lecture notes have been assembled from my HT 2018 lecture slides, with a few extra remarks thrown in. They are only draft notes, because the conversion from slide to notes is not complete. The Figures need numbers and captions, and the text adjusted somewhat. I hope you find them useful. GKN
1. The Bayesian inferential pipeline

Some unknown real-world quantity $\Theta$ takes values in $\Omega$. Let $S \subseteq \Omega$. We can imagine placing bets (ie, giving odds $O_S$ say) on the event that $\Theta \in S$ prior to seeing the data.

Prior odds $O_S = \Pr(\Theta \in S)/\Pr(\Theta \notin S)$ measure the strength of our belief that the event will occur.

If our beliefs are coherent (see lecture 11 on Savage Axioms) then they determine a subjective prior distribution (represented by a density $\pi(\theta)$ on $\Omega$ say) which exists and is unique.

$$\int_S \pi(\theta)d\theta = \Pr(\Theta \in S).$$

Observations $Y = (Y_1, ..., Y_n)$ have a distribution (informally, the likelihood) determined by $\Theta$: $Y \sim L(\theta; y)$ when $\Theta = \theta$. Suppose we observe $Y = y$. How do our odds for $\Theta \in S$ change? If $\pi(\theta|y)$ is the posterior density of the new posterior distribution $\Pr(\Theta \in S|Y = y)$ then by Bayes rule

$$\pi(\theta|y) = \frac{L(\theta; y)\pi(\theta)}{m(y)}$$

with $m(y) = \int_{\Omega} L(\theta; y)\pi(\theta)d\theta$ the normalising marginal likelihood. $m(y)$ is also the prior predictive distribution of the data - if $\theta \sim \pi(\theta)$ and $y' \sim L(\theta; \cdot)$ then marginally $y' \sim m(y')$.

The posterior predictive distribution of the data

$$p(y'|y) = \int_{\Omega} L(\theta; y')\pi(\theta|y)d\theta$$

is useful in goodness of fit - our data should predict new data that resembles itself - the simulated data $y' \sim p(y'|y)$ should cover the real data $y$.

[Remark - FCBSM cites RT Cox (1946, 1961) who shows subjective probabilities should be computed in the same way as probabilities defined from frequency or the reasoning will contradict “reasonable belief”]

Answers to questions about $\Theta$ can be given in terms of $\pi(\theta|y)$. In particular

$$\Pr(\Theta \in S|Y = y) = \int_S \pi(\theta|y)d\theta$$

is the posterior probability the truth lies in $S$, given the data and the prior probabilities expressed in $\pi(\theta)$.

When we are working with familiar parametric models (normal linear, GLM, hierarchical random effects) in a frequentist setting, statistical inference has a “pipeline”. For example, if the aim is to test a scientific hypothesis we might have:

1) EDA, 2) data modeling, 3) parameter estimation and model selection using MLE’s and LRT’s, 4) goodness of fit checking, 5) summary statement of outcome.

In a Bayesian setting for the same problem we might have:

1) prior elicitation, 2) EDA, 3) data modeling, 4) parameter estimation and model selection using the posterior mean and Bayes Factors, 5) goodness of fit, 6) conclusions.

In this course we will run through this inferential pipeline using Bayes methods. We will make many passes along the same sequence, developing the modeling and computation tools we need to be able to apply Bayes methods in different settings. One of the key themes of this course will be the importance (think of it as an
opportunity) of careful prior modeling. In the example which follows we run through the pipeline with a conjugate prior. The idea is to display the structure with a relatively simple example.

Example: (from FCBSM)

Data and question: How does the distribution of the number of children differ in a sample population of USA women with and without college degrees? Is the mean number greater for those without degrees?

Number of children recorded for 155 US women aged 40. Two groups: in group 1, \( n_1 = 111 \) without degrees; group 2, \( n_2 = 44 \) with degrees.

Let \( Y_{i,j} \) give the number children for woman \( j = 1, ..., n_j \) in group \( i = 1, 2 \). Let \( \Theta_i, i = 1, 2 \) give the mean of the distribution of the number of children for each of the two populations. Let \( Y_i = (Y_{1,i}, ..., Y_{n_1,i}), i = 1, 2, Y = (Y_1, Y_2), \Theta = (\Theta_1, \Theta_2) \).

Data model: Poisson model with independent responses \( Y_{i,j} \sim \Pi(\Theta_i) \). The likelihood for \( \Theta = \theta \) given \( Y = y \) is

\[
L(\theta; y \propto e^{-n_1\theta_1 - n_2\theta_2})\theta_1^{n_1\bar{y}_1}\theta_2^{n_2\bar{y}_2}.
\]

Prior Model: the gamma(\( \alpha, \beta \)) density is conjugate to the Poisson likelihood. Elicitation: I expect a prior mean \( E(\Theta_i) \) around 2. Taking \( \alpha = 2 \) and \( \beta = 1 \) for both groups, the upper tail of the prior predictive distribution \( m(y') \) for unseen data corresponds reasonably well to my experience:

\[
K=10000; \text{psim} = \text{rep}(0,K); \text{for (k in 1:K) \{psim[k]=rpois(1,lambda=rgamma(1,2,1))\}} \text{mean(psim>5)}
\]

[1] 0.0623

With these parameter choices, the prior is

\[
\pi(\theta) \propto e^{-\beta\theta_1 - \beta\theta_2})(\theta_1\theta_2)^{\alpha-1}
\]

up to a constant not depending on \( \theta \). This prior is non-informative with respect to a scientific hypothesis about the relative magnitude of the population means \( \theta_1, \theta_2 \).

Posterior Distribution:

\[
\pi(\theta|y) \propto e^{-(n_1+\beta)\theta_1-(n_2+\beta)\theta_2}\theta_1^{n_1\bar{y}_1+\alpha-1}\theta_2^{n_2\bar{y}_2+\alpha-1}
\]
from which we read off $\theta_1 \perp \theta_2$ and

$$\theta_i | y \sim \text{gamma}(n_i \bar{y}_i + \alpha, \beta + n_i) \text{ for } i = 1, 2.$$  

The 95% posterior equal-tail intervals for $\theta$ are $\theta_1 \in [1.7, 2.2], \theta_2 \in [1.2, 1.9]$ (the HPD intervals are almost identical).

We want to measure support for $\theta_1 > \theta_2$, ie the mean number of children being greater in the sample population without degrees. We want an estimate of $\Pr(\Theta \in S|y)$ with

$$S = \{(\theta_1, \theta_2) \in \mathbb{R}^2_{+}; \theta_1 > \theta_2\}$$

The posterior probability for $\theta_1 > \theta_2$ is $q = \Pr(\Theta_1 > \Theta_2|Y = y) \simeq 0.97$, conveniently estimated by Monte Carlo:

1. Simulate $\theta^{(k)} \sim \pi(\theta|y)$ for $k = 1, ..., K$;
2. Since $q = E(\mathbb{1}_{\theta_1 > \theta_2}|Y = y)$,

$$\hat{q}_K = K^{-1} \sum_k \mathbb{1}_{\theta_1^{(k)} > \theta_2^{(k)}}$$

is an unbiased and consistent estimator for $q$.

Goodness of fit: we have a model for the distribution of offspring number for a woman sampled from the population in either group. Our model determines a predictive distribution for this number. How well does our predictive distribution match the actual data?

Let $Y' \sim p(y'|y)$ be a simulated data set. Take a function $T(Y')$ chosen to be sensitive to likely problems with fit. How does $T(Y')$ compare to the actual value $T(y)$ we saw in the data? Report an estimate of

$$\xi = \Pr(T(Y') > T(y)|y)$$

as a measure of how unusual the data are. If $\xi$ is very small or very large, there could be a problem: the prior may be weighting against the true parameter value, the observation model may be poor, or both.
Let $Y'_1$ be the offspring-count for a new woman sampled from the group 1 population. The posterior predictive distribution of $Y'_1$ is:

$$p(y'_1|y) = \int L(\theta_1; y'_1)\pi(\theta_1|y)d\theta$$

$$\propto \int \frac{\theta^{n_1}e^{-\theta_1}}{y'_1!}\theta^{\alpha+n_1\bar{y}-1}\exp(-n_1\theta_1) d\theta_1$$

$$= \frac{1}{y'_1!} \int \theta^{\alpha+n_1\bar{y}+y'_1-1}\exp(-n_1\bar{y}+y'_1+1)\theta_1 d\theta_1$$

$$\propto \frac{1}{\Gamma(y'_1+1)\Gamma(\alpha+n_1\bar{y}+y'_1)}\frac{\Gamma(x+s)}{\Gamma(x+1)\Gamma(s)}p^x(1-p)^s$$

This is a negative binomial distribution (R-notation, $s$ is size and $p$ is prob)

$$\text{NegBin}(x;s,p) = \frac{\Gamma(x+s)}{\Gamma(x+1)\Gamma(s)}p^x(1-p)^s$$

with parameters

$$Y_1|y \sim \text{NegBin}(s = \alpha + n_1\bar{y}, p = (\beta + n_1)/(\beta + n_1 + 1)),$$

[the distribution of the number of failures $x$ in a sequence of Bernoulli($p$) trials before $s$ successes]. The posterior predictive mean $s(1-p)/p$ is

$$E(Y'_1|y) = (\alpha + n_1\bar{y})/(\beta + n_1)$$

which is equal $E(\Theta_1|y)$ as you might hope. Looking at the distribution for the group 1 population, the match doesn’t seem great for small family sizes. However the data are random and might not match by chance. It would be good to measure how unusual the real data is.

Focusing again on just the group 1 data, take $T(Y') = \hat{p}_2(Y')/\hat{p}_1(Y')$, where $\hat{p}_j = n_1^{-1}\sum_1 I_{Y'_1=j}$ is an estimate of the posterior predictive probability to see $j$ children. This ratio seemed out of line.

Step (1) Simulate $k = 1,2,\ldots,K$ synthetic data sets $y^{(k)} = (y^{(k)}_1,\ldots,y^{(k)}_{n_1})$ from the Posterior Predictive Distribution.
Step (2) For each simulated data set we compute the ratio
\[ r^{(k)} = \hat{p}_2(y^{(k)})/\hat{p}_1(y^{(k)}). \]
Step (3) Compute \( r^* = \hat{p}_2(Y_1)/\hat{p}_1(Y_1) \) for the real data. Make a histogram of the simulated \( r^{(k)}, k = 1, ..., K \)-values. Compute \( \hat{\xi} = n_1^{-1} \sum_k I_{r^{(k)} > r^*}. \)

The data are very unlikely. Only about 2 in 1000 simulated data sets showed such a large ratio.

Conclusion: fitting a Poisson model to the child counts, and using a gamma(2, 1) prior, we estimate the posterior probability for the group one mean to exceed the group two mean to be about 0.97. However the model fit to the data does not seem adequate, perhaps suggesting some problem with the data model.

Closing remarks: Notice that, even in this very simple setting I found it convenient to resort to Monte Carlo. In the next lecture we will develop some MC-skills to allow us to handle complex inference tasks in a generic way.
2. Markov Chain Monte Carlo methods

Why Monte Carlo? Many of the important quantities in Bayesian inference are expectations over the posterior distribution: posterior probabilities, marginal likelihoods, many point estimates and interval estimates - all given in terms of typically intractable integrals.

Suppose \( \Theta \in \Omega \) is a parameter vector with posterior density \( \Theta \sim \pi(\theta|y) \) given data \( Y = y \), suppose \( f : \Omega \to \mathbb{R} \) is some function and we want to evaluate the posterior expectation

\[
E_{\Theta|y}(f(\Theta)|Y = y) = \int_{\Omega} f(\theta) \pi(\theta|y) d\theta.
\]

If we have \( X_{t} \in \Omega, t = 1, 2, ..., T \) with \( X_{t} \sim \pi(\cdot|y) \) then

\[
\hat{f}_{T} = T^{-1} \sum_{i=1}^{T} f(X_{t})
\]

is an unbiased estimator for \( E_{\Theta|y}(f(\Theta)|Y = y) \).

Why MCMC? MCMC is a family of algorithms for simulating \( X_{0}, X_{1}, X_{2}, ... \) so that \( X_{t} \sim p \) (or at least \( X_{t} \) converges to \( p \) in distribution) for a user-defined probability distribution \( p \).

MCMC methods are one of the most versatile classes of Monte Carlo algorithms we have, and are in routine use across statistics.

I will set out theory for the case that \( \Omega \), the space of states of \( X \), is finite (and therefore discrete) because it is simpler. However, it also captures many of the essential issues. When we work on a computer we approximate any continuous quantities like \( \theta \), \( L(\theta; y) \), \( \pi(\theta) \) and \( \pi(\theta|y) \) using finite precision arithmetic so we are really working with finite \( \Omega \) anyway.

2.1. Markov chains. Let \( \{X_{t}\}_{t=0}^{\infty} \) be a homogeneous Markov chain of random variables on \( \Omega \) with starting distribution \( X_{0} \sim p^{(0)} \) and transition probability

\[
P_{i,j} = \mathbb{P}(X_{t+1} = j|X_{t} = i).
\]

Denote by \( P_{i,j}^{(n)} \) the \( n \)-step transition probabilities

\[
P_{i,j}^{(n)} = \mathbb{P}(X_{t+n} = j|X_{t} = i)
\]

and by \( p^{(n)}(i) = \mathbb{P}(X_{n} = i) \).

The transition matrix \( P \) is irreducible if and only if, for each pair of states \( i, j \in \Omega \) there is \( n \) such that \( P_{i,j}^{(n)} > 0 \). The Markov chain is aperiodic if \( P_{i,j}^{(n)} \) is non zero for all sufficiently large \( n \).

2.1.1. The Stationary Distribution and Detailed Balance. In discussing Markov chains we will work with a generic “target” distribution \( p(i), i \in \Omega \). This is the distribution we will try to sample. When we come to apply the MCMC methods to Bayesian inference, the target distribution will be the posterior \( p(\theta) = \pi(\theta|y) \).

The probability mass function (PMF) \( p(i), i \in \Omega \), \( \sum_{i \in \Omega} p(i) = 1 \) is a stationary distribution of \( P \) if \( pP = p \). If \( p^{(0)} = p \) then

\[
p^{(1)}(j) = \sum_{i \in \Omega} p^{(0)}(i) P_{i,j},
\]
so \( p^{(1)}(j) = p(j) \) also. Iterating, \( p^{(t)} = p \) for each \( t = 1, 2, \ldots \) in the chain, so the distribution of \( X_t \sim p^{(t)} \) doesn’t change with \( t \), it is stationary.

We want \( X_t \overset{D}{\to} p \). The convergence theorem for finite irreducible Markov chains tells us that if \( \Omega \) is finite, \( pP = p \), and \( P \) is irreducible and aperiodic, then indeed \( X_t \overset{D}{\to} p \). To show this works for given \( P \) we need to check \( pP = p \), but that is hard, as we have to sum over all \( \Omega \) to evaluate \( pP \). However...

2.1.2. **Detailed Balance.** If there is a probability mass function \( p(i), i \in \Omega \) satisfying
\[
p(i) \geq 0, \sum_{i \in \Omega} p(i) = 1 \text{ and }\]

"Detailed balance": \( p(i)P_{i,j} = p(j)P_{j,i} \) for all pairs \( i, j \in \Omega \),

then \( p = pP \) so \( p \) is stationary for \( P \).

Detailed balance is sufficient for stationarity, and much easier to check. A Markov chain satisfying DB is “reversible”.

Exercise: prove that if \( p(i)P_{i,j} = p(j)P_{j,i} \) for all \( i, j \in \Omega \) then \( p = pP \). Give the corresponding expressions for the continuous case where \( \Omega = \mathbb{R}^n \) say, expressing stationarity as an integral equation satisfied by \( p \), and write DB in terms of probability densities.

2.1.3. **Convergence and the Ergodic Theorem.** We choose some “start state” \( X_0 \sim p^{(0)} \) to initialise the Markov chain. If the chain converges to the target distribution \( p \), then \( X_t \sim p^{(t)} \) with \( p^{(t)} \approx p \) at large \( t \), so when we look at our Markov chain samples \( X_0, X_1, \ldots, X_T \), “most” of the samples are “nearly” distributed according to \( p \). Is this good enough?

Let \( \hat{f}_T \) be the estimator for \( E(f(X)), X \sim p \) we defined above.

**Theorem.** If \( \{X_t\}_{t=0}^\infty \) is an irreducible and aperiodic Markov chain on a finite space of states \( \Omega \) satisfying detailed balance with respect to the probability distribution \( p \), then as \( T \to \infty \)

\[
P(X_T = i) \to p(i) \text{ and } \hat{f}_T \xrightarrow{a.s.} E(f(X))
\]

for any bounded function \( f : \Omega \to R \). [For proof see eg Norris Markov Chains, CUP, (1997)]

We refer to such a chain as ergodic with target \( \pi \).

A more general statement asks for a positive or Harris recurrent chain. The conditions are simpler here because we are assuming a finite state space for the Markov chain.

We would really like to have a CLT for \( \hat{f}_n \) formed from the Markov chain output, so we have confidence intervals \( \pm \sqrt{\text{var} \hat{f}_n} \) as well as the central point estimate \( \hat{f}_n \) itself. CLT’s hold for all the examples in this course. [See eg Part C Advanced Simulation]

2.2. **Metropolis-Hastings Algorithm.** Suppose we need samples from a pmf \( p(i), i \in \Omega \) with \( \Omega \) a finite set. We give an algorithm simulating a Markov chain targeting \( p \). It is enough to give a rule simulating \( X_{t+1} \) given \( X_t \). The algorithm determines the transition probabilities \( P(X_{t+1} = j|X_t = i) \) and the transition matrix \( P \).

The basic idea here is to simulate a random walk \( X_0, X_1, X_2, \ldots \) in \( \Omega \) by accepting or rejecting proposals from a simple irreducible transition matrix \( Q_{i,j}, i, j \in \Omega \) which we get to choose. If \( p \) was stationary for \( Q \) we could just simulate the chain with
transition matrix $Q$, as the chain would target $p$. However we don’t know how to choose such $Q$. The trick is to “correct” proposals drawn from $Q$ to get a new effective transition matrix $P$ which satisfies DB for $p$.

Metropolis Hastings MCMC: the algorithm simulates a Markov chain. Let $q(j|i) = Q_{i,j}$ be a proposal distribution satisfying $q(j|i) > 0 \iff q(i|j) > 0$. If the chain is irreducible and aperiodic then it is ergodic with target distribution $p$.

Let $X_t = i$. $X_{t+1}$ is determined in the following way.

1. Draw $j \sim q(\cdot|i)$ and $u \sim U[0,1]$.
2. If
   \[
   u \leq \alpha(j|i) \quad \text{where} \quad \alpha(j|i) = \min \left\{ 1, \frac{p(j)q(i|j)}{p(i)q(j|i)} \right\}
   \]
   then set $X_{t+1} = j$, otherwise set $X_{t+1} = i$.

We initialise this with some $X_0 = i_0, p(i_0) > 0$ and iterate for $t = 1, 2, 3, ... T$ to simulate the samples we need.

2.2.1. Example: Simulating the hypergeometric distribution. The hypergeometric distribution $\text{HyperGeom}(k; K, N, n)$ with parameters $K = 10, N = 20, n = 10$ gives the probability for $k$ successes in $n$ draws from a population of size $N$ containing $K$ successes. If $p(k) = \text{HyperGeom}(k; K, N, n)$ then

\[
p(k) = \binom{K}{k} \frac{(N-K)}{n-k} / \binom{N}{n}
\]

Give a MH MCMC algorithm ergodic for $p(k)$,

\[
\max \{0, n + K - N\} \leq k \leq \min \{n, K\}
\]

(i.e $k = 0, 1, 2, ..., 10$ here).

Step 1: Choose a proposal distribution $q(j|i)$. It needs to be easy to simulate and determine an irreducible chain. A simple distribution that ‘will do’ is

\[
q(j|i) = \begin{cases} 
1/2 & \text{for } j = i \pm 1 \\
0 & \text{otherwise,}
\end{cases}
\]

i.e. toss a coin and add or subtract 1 to $i$ to obtain $j$. This is irreducible (we can get from any state $A$ to any other state $B$ by adding or subtracting 1’s).

Notice that we can “walk out” of the state space $B^-, B^+$ given above. If for eg $i = B^+$ and we propose $j = i + 1$ then we have proposed a state $j$ with zero probability in the target distribution. One transparent way to deal with this is give these states probability zero in the target, setting $p(j) = 0$ for all $j \not\in \Omega$.

Step 2: Write down the algorithm.
If $X_t = i$, then $X_{t+1}$ is determined in the following way.

1. Simulate $j \sim U\{i - 1, i + 1\}$ and $u \sim U[0,1]$.
2. If $B^- \leq j \leq B^+$ and
   \[
   u \leq \min \left\{ 1, \frac{p(j)q(i|j)}{p(i)q(j|i)} \right\} = \min \left\{ 1, \frac{K}{i} \frac{(N-K)}{n-k} \right\}
   \]
   then set $X_{t+1} = j$, otherwise (i.e if either condition fails) set $X_{t+1} = i$. 

The point here is that if $j < B^-$ or $j > B^+$ then $p(j) = 0$ so $\alpha = 0$ and we are bound to reject.

```r
# MCMC simulate $X_t \sim \text{HyperGeom}(K=10, N=20, n=10)$.
K<-10; N<-20; n<-10; Bm<-min(0, n+K-N); Bp<-min(n, K);
T<-1000; X<-rep(NA, T);
X[1]<-Bm # start state at lower bound, here $X[1]=0$
for (t in 1:(T-1)) {
  i<-X[t]
  j<-sample(c(i-1, i+1), 1)
  if (j<Bm | j>Bp) {
    X[t+1]<-i # must reject if outside SP
  } else {
    a<-min(1, (choose(K,j)*choose(N-K,n-j))/
           (choose(K,i)*choose(N-K,n-i)))
    U<-runif(1)
    if (U<=a) {X[t+1]<-j} else {X[t+1]<-i}
  }
}
```

2.2.2. 
Ergodicity proof for Metropolis Hastings. We assume the chain is irreducible and aperiodic - this must be checked separately for each MH MCMC algorithm. Under this assumption, it is sufficient (for ergodicity) to show that the Markov chain determined by the random MCMC update has $p$ as a stationary distribution. We will compute the transition matrix $P$ and show it satisfies detailed balance, $P_{i,j}p(i) = P_{j,i}p(j)$, since that implies $p = pP$.

We don’t need to calculate $P_{i,j}$ when $i = j$ as DB is clear. Suppose $j \neq i$. If $X_t = i$, then the probability $P_{i,j}$ to move to $X_{t+1} = j$ at the next step is the probability to propose $i$ at step 1 times the probability to accept it at step 2, so

$$P_{i,j} = P(X_{t+1} = j | X_t = i) = q(j|i)\alpha(j|i).$$

Now check DB:

$$p(i)P_{i,j} = p(i)q(j|i)\alpha(j|i) = p(i)q(j|i)\min\left\{1, \frac{p(j)q(i|j)}{p(i)q(j|i)}\right\} = \min\{p(i)q(j|i), p(j)q(i|j)\} = p(j)q(i|j)\alpha(i|j) = p(j)P_{j,i}$$

and we are done.

2.3. MCMC for state spaces which are not finite. Does this work for continuous rv? Computers use the “computer measure”. The reals are discretised. Let $x^*$ be the computer truncation of $x$ and $\delta x = \{y : x^* = x\}$. The length $|\delta x|$ of the cell containing $x$ is not constant. Roughly $|\delta x|/x \approx 10^{-15}$. $\pi(x)$ is approximated by $[\pi(x)]^*$. 

Figure 1. Left: $x$-axis is step counter $t = 1, 2, 3...200$. The $y$-axis is Markov chain state $X_t$ for $p(k) = \text{HyperGeom}(k; K = 10, N = 20, n = 10)$, $k = 0, ..., 10$. Right: histogram of $X_1, X_2, ..., X_n$ for $T = 1000$. 
The Hastings ratio we compute is
\[
\frac{\tilde{p}(y)[q(x|y)]^*}{\tilde{p}(x)[q(y|x)]^*} = \frac{[p(y)]^* \delta y [q(x|y)]^* \delta x}{[p(x)]^* \delta x [q(y|x)]^* \delta y}
\]
\[
\approx \frac{\Pr_p(Y \in \delta y) \Pr_q(X \in \delta x | Y = y)}{\Pr_p(X \in \delta x) \Pr_q(Y \in \delta y | X = x)}
\]
since \( \Pr_p(X \in \delta x) \approx [p(x)]^* \delta x \) etc. If we apply this to densities, we simulate the approximate distribution. Our discussion of Markov chains on finite spaces is relevant.

2.3.1. MH example: an equal mixture of bivariate normals.
\[
\pi(\theta) = (2\pi)^{-1} \left( 0.5e^{-(\theta-\mu_1)^T \Sigma_1^{-1} (\theta-\mu_1)/2} + 0.5e^{-(\theta-\mu_2)^T \Sigma_2^{-1} (\theta-\mu_2)/2} \right)
\]
with \( \theta = (\theta_1, \theta_2) \). Use \( \mu_1 = (1,1)^T \), \( \mu_2 = (4,4)^T \) and \( \Sigma_1 = \Sigma_2 = I_2 \) for this illustration.

Step 1. For a proposal distribution \( q \) we want something simple to sample. The simplest thing I can think of is the same as before:
\[
\theta_i' \sim U(\theta_i - a, \theta_i + a)
\]
with \( a \) a fixed constant. Note that this time we are proposing in a box of side \( 2a \).

That is easy to sample, and certainly \( q(\theta' | \theta) > 0 \) if \( q(\theta | \theta') > 0 \) since \( q(\theta' | \theta) = q(\theta | \theta') = 1/4a^2 \).

Step 2. The algorithm is, given \( \theta^{(n)} = \theta \),
[1] for \( i = 1, 2 \) simulate \( \theta_i' \sim U(\theta_i - a, \theta_i + a) \)
[2] with probability
\[
\alpha(\theta' | \theta) = \min \left\{ 1, \frac{\pi(\theta')}{\pi(\theta)} \right\}
\]
set \( \theta^{(n+1)} = \theta' \) otherwise set \( \theta^{(n+1)} = \theta \).

This algorithm is ergodic for any \( a > 0 \) but we will see that the choice of \( a \) makes a difference to efficiency.

```r
#MCMC simulate X_t according to a mixture of normals
f<-function(x,mu1,mu2,S1i,S2i,p1=0.5) {
  #mixture of normals, density up to constant factor
  c1<-exp(-t(x-mu1)%*%S1i%*%(x-mu1))
  c2<-exp(-t(x-mu2)%*%S2i%*%(x-mu2))
  return(p1*c1+(1-p1)*c2)
}
a=3; n=2000
mu1=c(1,1); mu2=c(4,4); S=diag(2); S1i=S2i=solve(S);
X=matrix(NA,2,n); X[,1]=x=mu1
for (t in 1:(n-1)) {
  y<-x+(2*runif(2)-1)*a
  MHR<-f(y,mu1,mu2,S1i,S2i)/f(x,mu1,mu2,S1i,S2i)
  if (runif(1)<MHR) x<-y
  X[,t+1]<-x
}
(see the associated R-file for plotting commands)
```
2.4. Convergence and mixing. We want to estimate $E_p(f(X))$ using our MCMC samples $X_0, X_1, X_2, ..., X_n$ targeting $p(x)$ and calculate the estimate $\bar{f}_n = n^{-1} \sum f(X_t)$. The ergodic theorem tells us this estimate converges in probability to $E_p(f(X))$.

How large should we take $n$? There are two issues. First, suppose $p^{(0)}(x) = p(x)$, so we start the chain in equilibrium. The variance, $\text{var}(\bar{f}_n)$, of $\bar{f}_n$ will get smaller as $n$ increases. We should choose $n$ large enough to ensure $\text{var}(\bar{f}_n)$ is small enough so that $\bar{f}_n$ has useful precision. However, calculating $\text{var}(\bar{f}_n)$ wont be completely straightforward as the MCMC samples are correlated. Second, we dont start the chain in equilibrium. The samples in the first part of the chain are biased by the initialization. It is common practice to drop the first part of the MCMC run (called “burn-in”) to reduce the initialization bias. We know $p^{(t)} \to p$ as $t \to \infty$ and want to choose a cut-off $T$ beyond which $p^{(t)} \simeq p$ to a good approximation. We need $n \gg T$ so that most of the samples are representative of $p$.

Note that if $n \gg T$ then the bias from burn-in will be slight anyway. One observation here is that if you need to drop states from the start of the chain to reduce this bias, you probably havnt run the chain long enough.

The following figures show autocorrelations for two MCMC runs for the bivariate normal mixture, with different values of the jump size $a = 2, 4$.

2.4.1. MCMC variance in equilibrium. $X_0, X_1, X_2, ...$ are correlated so $\text{var}(\bar{f}_n) \neq \text{var}(f(X))/n$ in general.

Correlation at lag $s$

$$\rho_s(f) = \frac{\text{cov}(f(X_i), f(X_{i+s}))}{\text{var}(f(X_i))}$$

(so $\rho_0 = 1$). Let $\sigma^2 = \text{var}(f(X_i))$. This doesnt depend on $i$ because the chain is stationary, because it was started in equilibrium.

Express $\text{var}(\bar{f}_n)$ in terms of $\rho_s(f)$. This gives insight and leads to an estimator for $\text{var}(\bar{f}_n)$, since we can estimate $\rho_s(f)$. 

\[
\text{var}(\bar{f}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{cov}(f(X_i), f(X_j))
\]
\[
= \sigma^2 n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{|i-j|}
\]
\[
= \sigma^2 n^{-1} \left[ 1 + 2 \sum_{s=1}^{n-1} \left( 1 - \frac{s}{n} \right) \rho_s \right]
\]
\[
\simeq \sigma^2 n^{-1} \left[ 1 + 2 \sum_{s=1}^{n-1} \rho_s \right]
\]
\[
= \sigma^2 \tau_f / n,
\]
if as usual \( \rho_s \) is small when \( s \) is large. \( \tau_f \) is called the integrated autocorrelation time. The quantity \( \text{ESS} = n/\tau_f \) is called the effective sample size - the number of
independent samples that would give the same precision for $\mathcal{F}$ as the $n$ correlated samples we actually have.

We can estimate $\gamma_s = \text{cov}(f(X_i), f(X_{i+s}))$ using

$$\hat{\gamma}_s = \frac{1}{n} \sum_{i=1}^{n-s} (f(X_i) - \hat{f})(f(X_{i+s}) - \hat{f}),$$

and $\gamma_0 = \text{var}(f(X_i))$ (as usual) from the sample output, and compute $\hat{\rho}_s = \hat{\gamma}_s/\hat{\gamma}_0$.

We get an estimate of $\tau_f$,

$$\hat{\tau}_f = 1 + 2 \sum_{s=1}^{M} \hat{\rho}_s,$$

with $M$ a cut-off on the sum. $\hat{\rho}_s$ goes to zero with $s$ and is dominated by noise at large $s$; beyond some value of $s$ we are actually making our estimate worse by adding more terms to form the sum of $\hat{\rho}_s$ over $s$. We have to truncate the sum over $s$, at $s = M$.

2.4.2. MCMC convergence. There is no simple generic sufficient condition we can test for convergence. Here some checks we can run to detect poor mixing and identify a burn-in and run length.

1. Make multiple runs from different start states and check marginal distributions agree.
2. Plot the autocorrelation function. Check that it falls off to vary around zero. Calculate the ESS and check it is reasonably large.
3. Plot MCMC traces of the variables and key functions. The chain should be stationary after burn-in.

Here is an example of some of the plots I would use for convergence checking on the normal mixture MCMC sampler. In this example the convergence is quite poor: the three histograms differ by more than the precision I would typically be aiming for. See associated R-file for further examples.

2.5. Data Augmentation and the Gibbs sampler.
2.5.1. The Gibbs sampler. The Gibbs sampler is particularly natural. It works well if we can easily sample conditional distributions. Suppose for eg $\pi(\theta_1, \theta_2)$ is a bivariate density we want to sample and we have $(\theta_1, \theta_2) \sim \pi$.
Simulate a new $\theta_1' \sim \pi(\theta_1'|\theta_2)$ and then simulate $\theta_2' \sim \pi(\theta_2'|\theta_1')$ (using the new $\theta_1'$). The distribution of $(\theta_1', \theta_2')$ is

$$p(\theta_1', \theta_2') = \int \pi(\theta_1, \theta_2) \pi(\theta_1'|\theta_2) \pi(\theta_2'|\theta_1') d\theta_1 d\theta_2$$

$$= \int \pi(\theta_1, \theta_2) \pi(\theta_1'|\theta_2) \pi(\theta_2'|\theta_1') \pi(\theta_1') \pi(\theta_2') d\theta_1 d\theta_2$$

$$= \int \pi(\theta_1|\theta_2) \pi(\theta_2|\theta_1') \pi(\theta_1', \theta_2') d\theta_1 d\theta_2$$

$$= \int \pi(\theta_1|\theta_2) \pi(\theta_2|\theta_1') d\theta_2 \int \pi(\theta_2|\theta_1') d\theta_2$$

$$= \pi(\theta_1', \theta_2')$$ since the integrals are each equal one.

so if we start with $(\theta_1, \theta_2) \sim \pi$ then after these two steps we have a new correlated sample $(\theta_1', \theta_2') \sim \pi$. If $\theta^{(0)} = (\theta_1, \theta_2)$, then $\theta^{(1)} = (\theta_1', \theta_2')$ and we can iterate to simulate $\theta^{(2)}, \theta^{(3)}...$

The Gibbs sampler comes in several flavors. In the multivariate case $\theta = (\theta_1, \theta_2, ..., \theta_K)$ we can sample $\pi(\theta_i|\theta_{-i})$ for $i = 1, 2, 3, ..., K$ in turn, or select $i \sim U\{1, 2, 3, ..., K\}$ at each step. We need to check the ergodicity conditions when we construct a new MCMC sampler.

Exercise: consider a Metropolis Hastings algorithm targeting $\pi(\theta_1, \theta_2)$ with proposal $q(\theta'|\theta) = \pi(\theta'|\theta_2)$. Show that the acceptance probability is equal one. Hence show that the Gibbs sampler is a MH sampler.

2.5.2. Data Augmentation. Some important early applications of the Gibbs sampler arise in the context of missing data. This is also called “data augmentation". DA is convenient when the likelihood on the full data is much simpler than the likelihood on the observed data.

Suppose the observation process is $z \sim p(z|\theta)$, $y \sim p(y|z, \theta)$ and we observe $y$. The posterior $p(\theta|y)$ is awkward as the likelihood is an integral,

$$L(\theta; y) = \int p(y|z, \theta)p(z|\theta) dz$$

In data augmentation we work with the joint posterior density $p(\theta, z|y)$, thinking of the missing data as another parameter. The likelihood is

$$p(\theta, z|y) \propto p(y|z, \theta)p(z|\theta)p(\theta)$$

The original DA algorithm was a Gibbs sampler

[1] $z' \sim p(z'|\theta, y)$ with $p(z'|\theta, y) \propto p(y|z', \theta)p(z'|\theta)$

[2] $\theta' \sim p(\theta'|y, z')$

2.5.3. Example of DA: Probit regression. In probit regression we have, for the $i$th observation $y_i, i = 1, ..., n$, covariates $x_i = (x_{i,1}, ..., x_{i,p})$, parameters $\theta = (\theta_1, \theta_2, ..., \theta_p)$, a linear predictor $\eta_i = \sum_j \theta_j x_{i,j}$, and observation model

$$y_i \sim Bernoulli(\Phi(\eta_i))$$

with $\Phi$ the cdf of a standard normal.
If the prior for $\theta$ is $\pi(\theta)$ then the posterior for $\theta | y$ is
\[
\pi(\theta | y) \propto \pi(\theta) \prod_i \Phi(y_i) \Phi(y_i + 1 - \Phi(y_i))^{1 - y_i},
\]
with $\eta_i = \eta_i(\theta), i = 1, \ldots, n$.

There is another way to represent this model. Let $z_i \sim N(\eta_i, 1), i = 1, \ldots, n$ be scalar normal random variable. Set $y_i = 1$ if $z_i > 0$ and $y_i = 0$ if $z_i \leq 0$. This gives $y_i \sim Bernoulli(\Phi(\eta_i))$ again.

To see this, write $Pr(\theta | y) = Pr(\eta | y + W > 0)$ with $W$ a standard normal. Now $Pr(\eta + W > 0) = Pr(W > -\eta)$. Since a standard normal is symmetrical, $Pr(W > -\eta) = Pr(W < \eta)$ which is $\Phi(\eta)$.

The joint posterior is
\[
\pi(\theta, z | y) \propto p(y | z, \theta) \pi(z | \theta) \pi(\theta)
\]
with $p(y_i | z_i) = p(y_i | z)$ and
\[
p(y_i | z_i) = I_{y_i = 1} I_{z_i > 0}.
\]

The marginal distribution of $\theta | y$ is just $\pi(\theta | y)$ from the previous slide, so nothing has changed there. In this representation we have a latent 'propensity' score $z$ for each observation $y_i$ and we effectively observe the sign of $z$.

Suppose we are doing Bayesian inference for $\theta$ with normal priors.

Now $p(y_i | z, \theta)$ is 0/1 as $z$ agrees/disagrees with $y_i$ and $\pi(z | \theta)$ is $N(\eta_i, 1)$. For the same reason, $\pi(\theta | y, z) = \pi(\theta | z)$. If $\pi(\theta)$ is normal, it is conjugate to $\pi(z | \theta)$ and $\theta | y, z$ is also normal for each component. In this case we have
\[
\pi(\theta, z | y) \propto \pi(\theta) \prod_i \pi(z_i | \theta) I_{y_i = 1} I_{z_i > 0}
\]

Our DA/Gibbs sampler becomes

1. For each $i = 1, \ldots, n$, simulate $z'_i \sim N(\eta_i, 1)$ conditioned on the sign of $z_i$ (+/− as $y_i = 1/0$)
2. simulate $\theta' \sim p(\theta | z')$ where
\[
p(\theta | z') \propto \pi(\theta) \prod_i N(z_i; \eta_i(\theta), 1).
\]

**Exercise** Let $X$ be an $n \times p$ design matrix with rows $x_i, i = 1, 2, \ldots, n$ and $\theta$ a $p$-component vector of parameters. Suppose we have binary observations $y_i, i = 1, 2, \ldots, n$ with $Pr(Y_i = 1 | x_i) = \Phi(x_i \theta)$. Suppose our prior for $\theta$ is $\theta \sim N(0, \Sigma)$.

Show
\[
\pi(\theta | z) = N(\theta; \mu, V)
\]
with $\mu = VX^Tz$ and $V = (\Sigma^{-1} + X^TX)^{-1}$ and comment briefly on how you might simulate $\theta | z$.

Show that
\[
\pi(z | y, \theta) \propto \prod_{i:y_i=0} N(z_i; x_i \theta, 1) I_{z_i < 0} \prod_{i:y_i=1} N(z_i; x_i \theta, 1) I_{z_i \geq 0},
\]
and explain how to sample $z_i | y_i, \theta$ by the method of inversion.
3. Model selection

What do we mean by a model? In Bayesian inference we have a prior model \( \Theta \sim \pi(\theta) \), \( \Theta \in \Omega \) and an observation model \( Y \sim L(\theta; y) \) when \( \Theta = \theta \) and this determines a posterior probability distribution \( \pi(\theta | y) \) for the unknown true parameter value \( \Theta \) given that we see data \( Y = y \).

The model is specified by the two distributions \( L(\theta; y) \) and \( \pi(\theta) \) (using densities as a shorthand for distributions).

Suppose we are considering a discrete set \( \mathcal{M} \) of models indexed by integers \( m \in \mathcal{M} \).

In the most general setup the parameter space might vary from model to model. My notation below assumes that the functions \( L \) and \( \pi \) may change but \( \Omega \) remains the same.

**Example:** In lecture 1 \( Y_{i,j} \) = number of children for \( i \)th woman in group \( j = 1/2 \) with/without degrees. Group sizes were \( n_1 = 111 \) and \( n_2 = 44 \). The observation model was \( Y_{i,j} \sim \text{Poisson}(\theta_i) \) jointly independent. The prior was \( \theta_1, \theta_2 \sim \text{Gamma}(\alpha, \beta) \) independently.

I suggested \( \alpha = 2, \beta = 1 \) was a reasonable prior, but you might have other knowledge and decide the prior \( a = 1, b = 1 \) better represents your beliefs for this group of people.

Let \( \mathcal{M} = \{1, 2\} \) be the set of model indices. When \( m = 1 \) we have the prior model \( \pi_1(\theta) = \text{Gamma}(\theta_1; \alpha_1, \beta_1) \) with \( (\alpha_1, \beta_1) = (2, 1) \) and when \( m = 2 \) we have the corresponding model with \( (\alpha_2, \beta_2) = (1, 1) \). The prior varies across models. The Likelihood stays the same.

**Example:** When we do regression in a GLM with \( p \) covariates, a \( p \)-component vector \( \theta = (\theta_1, \theta_2, ..., \theta_p) \) of effects, a linear predictor \( x\theta \), an (inverse) link function \( \mu(x\theta) \) and a binary response \( Y \sim \text{Bern}(\mu(x\theta)) \), there are several classical choices for the link function: logistic \( \mu_1(\eta) = \exp(\eta)/(1 + \exp(\eta)) \), probit \( \mu_2(\eta) = \Phi(\eta) \) and (less used?) log-log \( \mu_3(\eta) = 1 - \exp(-\exp(\eta)) \). If we have \( n \) observations \((y_i, x_i), i = 1, 2, ..., n\) then we have 3 different likelihoods \( m \in \mathcal{M} = \{1, 2, 3\} \)

\[
L_m(\theta; y) = \prod_{i=1}^{n} \mu_m(x_i\theta)^{y_i}(1 - \mu_m(x_i\theta))^{1-y_i}
\]

The prior \( \pi(\theta) \) for \( \theta \) is the same for all three models in this example.

3.1. Why do model selection? Model Construction/Improvement - We have a few models and we want to find the one that best fits the data (this is the case for the link function example).

Model comparison - two scientists have different beliefs about \( \theta \) and want to decide which is more in line with reality (for eg the birth rate example).

Hypothesis testing - we have a small number of specific hypotheses developed from physical models of reality. We think one of them is a true description of the generate process for the data. Which one?

Goodness of fit/Model expansion - we want to check a model \( M_0 \) is adequate. We define a model \( M_1 \) incorporating likely model extensions (we might replace a linear covariate with a polynomial) and compare \( M_0 \) and \( M_1 \). In these examples the best (correct? true?) choice for the model index \( m \) is unknown. We can treat it as just one more parameter we dont know and work with the extended posterior \( \pi(\theta, m | y), \ (\theta, m) \in \Omega \times \mathcal{M} \).
The marginal model probability \( \pi(m|y) = \int_\Omega \pi(\theta, m|y)d\theta \) can be written
\[
\pi(m|y) \propto p(y|m)\pi_M(m),
\]
where
\[
p(y|m) = \int_\Omega L_m(\theta; y)\pi_m(\theta)d\theta
\]
is the marginal likelihood under model \( m \), and \( \pi_M(m) \) is the prior probability that \( m \) is the correct model. If we want to discover the true model, \( M \) say, our loss function might be \( L(M, m) = 1_{M=m} \), the 0-1 loss function. We should choose the model with the largest posterior probability. If \( m \) and \( m' \) are two models we favor \( m \) if \( A = \pi(m|y)/\pi(m'|y) > 1 \). This number has the simple meaning that model \( m \) is \( A \) times more probable a posteriori than model \( m' \).

We may be concerned that our prior weighting \( \pi_M(m) \) is distorting this ratio. The Bayes factor
\[
B_{m,m'} = p(y|m)/p(y|m')
\]
is equal to \( \pi(m|y)/\pi(m'|y) \) if \( \pi_M(m) = \pi_M(m') \), ie, if the prior weighting is equal. Notice that \( p(y|m) \) is the prior predictive distribution (under model \( m \)) for the data \( y \) we observed. So the Bayes factor also tells us how much more likely the data is under model \( m \) than \( m' \). In this setting the model \( m \) we select is just the one that best predicts the data we saw, not necessarily the true model. However in this misspecified setting, where the true model is not in the set under consideration, we cannot interpret \( \pi(m|y) \) straightforwardly (it should be zero for all models).

Bayes factors (and posterior odds) have a built in penalty on complexity. As the prior becomes more diffuse, the probability mass it puts on the support of the likelihood goes down. On the other hand the maximum value of likelihood itself increases.

Marginal likelihoods are hard in practice to estimate (\( L \) is concentrated in \( \Omega, \pi \) is diffuse). Also, because they are averages (\( L \) averaged over \( \pi \)) ML’s depend on \( L \) and \( \pi \) everywhere in \( \Omega \) (not just in the vicinity of of the MLE) so if we have model mispecification anywhere (even in the tails, at physically uninteresting values of \( \theta \)) then \( B_{m,m'} \) may be distorted.

Example: in the first example the likelihood doesn’t change, we just change the prior. If each model is a priori of equal probability \( \pi_M(m) = 1/2, m=1,2 \). The marginal likelihood for group \( i \) (data \( y_i = (y_{i,j}), j=1,...,n_i \) in model \( m \) is
\[
p(y_i|m) \propto \int \exp(-n_i\theta_i)\theta_i^{n_i\bar{y}_i}\text{Gamma}(\theta_i; \alpha_m, \beta_m)d\theta_i
\]
leaving off factors of \( y_{i,j}! \), so the ML is
\[
p(y|m) = \prod_{i=1,2} \frac{\Gamma(\alpha_m + n_i\bar{y}_i)}{\Gamma(\alpha_m)} \frac{\beta_m^{\alpha_m}}{\beta_m^{\alpha_m + n_i\bar{y}_i}}
\]
and the Bayes Factor is
\[
B_{1,2} = p(y|m=1)/p(y|m=2)
\]
Plugging in the number (see R code) I get \( B_{1,2} \approx 1.94 \) mild evidence in favor of \( a=2 \) over \( a=1 \).

3.2. Estimation methods. We noted that marginal likelihoods are hard to estimate. Here are some consistent estimators, in order of increasing stability.

Naive: \( \hat{\theta}(t) \sim \pi(\theta), t=1...T \) and \( \hat{p} = T^{-1} \sum L(\theta(t); y) \)
In our setting the choice above gives $hertiors.$ and we can estimate numerator and denominator using samples from the two pos-

estimators exploit the identity $θ$ so if $Bridge Estimate: let $Harmonic Mean: this is importance sampling using the posterior. $θ(t) ∼ π(θ|y), t = 1...T$ and

$$\hat{p} = \sum_t w_t L(θ(t); y) \sum_t w_t$$

with $w_t ∝ π(θ(t))/π(θ(t)|y)$, leading to

$$\hat{p} = \left(T^{-1} \sum_t L(θ(t); y)^{-1}\right)^{-1}$$

Bridge estimator: let $h : Ω → R$ be a parameter function we can choose. This estimator exploits the identity

$$p(y|m) = \frac{E_θ(π(θ)L(θ; y)h(θ))}{E_θ[π(θ)h(θ)]}$$

so if $θ^{(1,t)} ∼ π(θ), θ^{(2,t)} ∼ π(θ|y), t = 1...T$, the estimator is

$$\hat{p} = \frac{\sum_t π(θ^{(1,t)})L(θ^{(1,t)}; y)h(θ^{(1,t)})}{\sum_t π(θ^{(2,t)})h(θ^{(2,t)})}.$$ 

The idea is to choose $h(θ)$ to minimise the mean square error of $\hat{p}$. Meng and Wong (1996) give the detail. A good rule of thumb for bridging two distributions with densities $π_1/\pi_2$ is $h ∝ 1/√\pi_1/\pi_2$.

For the bridge estimate we can more directly estimate the Bayes factor (with $p(θ|y, m) ∝ π_m(θ)L_m(θ; y)$ the posterior under model $m$) since

$$\frac{p(y|m = 1)}{p(y|m = 2)} = \frac{E_θ[π_2(θ)h(θ)]}{E_θ[π_1(θ)h(θ)]}$$

and we can estimate numerator and denominator using samples from the two post-

eriors.

In our setting the choice above gives $h(θ) = 1/√π_1L_1π_2L_2$ so if $θ^{(1,t)} ∼ π(θ|y, m = 1), θ^{(2,t)} ∼ π(θ|y, m = 2), t = 1...T$, the estimator is

$$\hat{p} = \frac{\sum_t \left(π_1(θ^{(1,t)})L_1(θ^{(1,t)}; y)π_2(θ^{(2,t)})L_2(θ^{(2,t)}; y)\right)^{1/2}}{\sum_t \left(π_1(θ^{(1,t)})L_1(θ^{(1,t)}; y)π_2(θ^{(2,t)})L_2(θ^{(2,t)}; y)\right)^{1/2}}.$$ 

Example: compare the probit and logit link functions for the Challenger O-ring data. I scaled the temperatures to mean zero sd one prior to fitting. If $m = 1$ is logistic and $m = 2$ is probit the likelihoods are

$$L_m(β; y) = \prod_{i=1}^n μ_m(β_1 + β_2 x_i)^{y_i}(1 - μ_m(β_1 + β_2 x_i))^{1-y_i}$$

with $μ_m$ logistic or probit. I used the prior $β_1, β_2 ∼ N(0, 3^2)$ as this gives sensible variation in $μ$. The marginal likelihoods are

$$p(y|m) ∝ \int \prod_{i=1}^n L_m(β; y) \exp(-β^T β/18) dβ_1 dβ_2.$$ 

See attached code. Estimates of Bayes Factors: 3.5 (! naive) 0.73 (harmonic) 0.552 (bridge) 0.554 (R-built-in Laplace) The data provides weak evidence “barely worth mentioning” in favor of the Probit over the Logit.
4. Model averaging

If our focus is on estimating a parameter (say, finding a credible interval for $\theta$) we might want to allow for the uncertainty in which model is the right model. Instead of selecting a model and estimating $\theta$ in that model, we could integrate over the model uncertainty and work with

$$\pi(\theta|y) = \sum_{m \in M} \pi(\theta, m|y).$$

In terms of our model distributions

$$\pi(\theta|y) = \sum_m \pi(\theta, m, y) / p(y) = \sum_m \pi(\theta|m, y) p(y|m) \pi_M(m) / p(y)$$

In this expression $\pi(\theta|m, y)$ is our old posterior given model $m$. The denominator $p(y) = \sum_m p(y|m) \pi_M(m)$ is a normalising constant (no $\theta, m$-dependence).

If our loss function is the MSE, then the model-averaged posterior mean $E_{\theta, M}(\theta|Y = y)$ is the best estimator. Selecting a model $m^*$ say first and using $E_{\theta}(\theta|Y = y, M = m^*)$ will not in general minimise the expected loss allowing for uncertainty in the model. Note that we are assuming in this discussion that the parameter space doesn’t change as we vary the model.

Example (choice of link functions): Using model averaging in the link function setting the average is

$$\pi(\beta|y) = \pi(\theta|m = 1, y) w_1 + \pi(\theta|m = 1, y) w_2$$

I have taken $\pi_M(m = 1) = 1/2$ etc so that

$$w_1 = p(y|1)/(p(y|1) + p(y|2))$$

and $w_2 = 1 - w_1$. In terms of the Bayes Factor (estimated using Bridge estimator)

$$w_1 = B_{1,2}/(B_{1,2} + 1) \approx 0.55/(1 + 0.55)$$

approximately. In the figure below the posterior densities for $\beta_2|y$ have been estimated for each model using a histogram. The model average is slightly weighted towards Model 2/Probit.

4.1. Example: model averaging in a normal linear model. [see Hoff 9.3.1 - I use simulation and different example]

Consider model averaging in polynomial regression,

$$Y \sim N(X \theta_z, \sigma^2), \quad \theta_z = (z_1 \theta_1, ..., z_p \theta_p)$$

with $z = (z_1, ..., z_p)$ a vector of binary indicator variables, and

$$X = [X_1, X_2, ..., X_p], \quad X_1 = (1, ..., 1)^T, \quad X_2 = (x_1, ..., x_n)^T,$$

and $X_i = X_2^{i-1}$, $i = 2, 3, ..., p$, so the highest power is $p - 1$.

Here $z_i$ switches on and off the contribution $x_i^{r-1}$ from $X_i \theta_i$ to the mean of $Y$. For example $z = (1, 1, 0, 0, ..., 0)$ gives linear regression, with $Y = \theta_1 + \theta_2 x + \epsilon$.

There are $2^p$ models. The model index $z$ has prior $\pi(z)$. The joint posterior is

$$\pi(\theta, z, \sigma|y) \propto L(\theta_z, \sigma; y) \pi(\theta, \sigma) \pi(z),$$

with $L(\theta_z, \sigma; y) = N(y; X \theta_z, \sigma^2)$. 
This is the joint posterior distribution of the model (indexed by $z$) and the parameters $\theta$ and $\sigma$.

**Priors:** In this example we focus on the model averaging, so we take the priors as given.

We scale the covariates and take $\beta_i \sim N(0, 3)$ and $\sigma \sim 1/\sigma$ in the parameter priors.

We take the model prior

$$
\pi(z) \propto \xi |z| \left(1 - \xi\right)^{p-|z|},
$$

with say $\xi = c/p$ (with $c$ small, I use $c = 3$ for $p = 5$ below). This says we expect about $c$ of the covariates to play a role.

**Posterior:** Joint distribution of model index and parameters is

$$
\pi(\theta, z, \sigma | y) \propto N(y; X\theta, \sigma^2) \times N(\theta; 0, I_p) \times \sigma^{-1} \times \xi^{|z|}(1 - \xi)^{p-|z|}
$$

**MCMC targeting $\pi(\theta, z, \sigma | y)$:** I use MH MCMC and random-walk proposals. Suppose $X_t = (\theta, z, \sigma)$ is the state of the Markov chain at step $t$. I cycle through a $\theta$-update, a $z$-update and a $\sigma$-update sequentially. At each stage I sample the index for the parameter to be updated $i \sim U\{1, 2, ..., p\}$.

If we are updating $\theta_i$ then $\theta_i' \sim U(\theta_i - a, \theta_i + a)$. The $z$’s and the other $\theta$’s dont change so

$$
\theta_z' = (z_1 \theta_1, ..., z_{i-1} \theta_{i-1}, z_i \theta_i', z_{i+1} \theta_{i+1}, ..., z_p \theta_p)
$$

and the acceptance probability is

$$
\alpha_{\theta_i}(\theta' | \theta) = \min\{1, \frac{N(y; X\theta_i', \sigma^2)}{N(y; X\theta_i, \sigma^2)}\frac{N(\theta_i'; 0, 9)}{N(\theta_i; 0, 9)}\}.
$$

The proposal probabilities $q(\theta_z' | \theta_z) = q(\theta_z' | \theta_z)$ cancel as do factors in $\pi(\theta, z', \sigma | y) / \pi(\theta, z, \sigma | y)$ not depending on $\theta_i$.

To update a $z$ I use $z_i' = 1 - z_i$ and leave everything else unchanged giving

$$
\theta_z' = (z_1 \theta_1, ..., z_{i-1} \theta_{i-1}, z_i' \theta_i, z_{i+1} \theta_{i+1}, ..., z_p \theta_p)
$$
and
\[ \alpha_z(\theta_z|\theta_z) = \min\{1, \frac{N(y; X\theta_z, \sigma^2)\xi|\theta_z'(1 - \xi)^{\theta_z}}{N(y; X\theta_z, \sigma^2)\xi|\theta_z'(1 - \xi)^{\theta_z}} \}. \]

There is also a \( \sigma \) update. This could be done using rw-proposal (rejecting if negative). In the attached scheme I use a “random walk on a log scale” which we will justify later in the course.

4.2. Summarising the Posterior over models and parameters for the NLM regression. We illustrate the method\(^1\) for model averaging polynomial regression on data for average claims paid per policy for automobile insurance in New Brunswick in the years 1971-1980. I scaled both \( x \) and \( y \) to mean zero and standard deviation one and allow up to quartic terms, \( X = (1, x, x^2, x^3, x^4) \).

The R-code and further detail of the algorithm are available on the course website.

We ran the code and generated samples
\[ (\theta_z(t), z(t), \sigma(t)) \sim \pi(\theta, z, \sigma|y) \quad t = 1, 2, ..., T. \]

The traces show the model variation. The posterior predictive distribution is represented by plotting the function \( v(x)\theta_z(t) \) against \( x \) with \( v = (1, x, x^2, x^3, x^4) \). Notice that when \( z_i = 0 \), \( \theta_i \) is disconnected from the data and “floats freely” in the prior, \( \theta_i|z_i = 0, y \sim N(0, 9) \). There is little evidence for a quartic term, so \( z_4 = 0 \) frequently and we see \( \theta_4 \) has about the same variation as \( N(0, 9) \).

The data are black dots, the samples are grey, and the red curve \( \hat{\mu}(x) \) say, is an estimate of the expectation of the posterior predictive distribution over \( y' \sim p(y'|y) \) at \( x \),
\[
\mu(x) = E_{\theta, z, \sigma|y}(E_{Y'|\theta, z, \sigma}(Y'|\theta, z, \sigma))
\]
which is just \( \hat{\mu}(x) = v(x)\hat{\theta}_z \) with \( \hat{\theta}_i \) the posterior mean \( z_i \theta_i \) averaged over models.

The posterior probabilities for the models are estimated from the \( z(t) \) values. The marginal probability for model \( z_0 \)
\[ \pi(z = z_0|y) = E_{\theta, z, \sigma|y}(I_{z = z_0}), \]
so we just count up the proportion of times \( z \) appears in the MCMC output,
\[ \hat{\pi}_{z_0} = T^{-1} \sum_i I_{z(t) = z_0} \]
to get an unbiased estimate for \( \pi(z = z_0|y) \). The results (as percentages) sorted by magnitude are
\[
\begin{align*}
11110 & 00010 10110 00110 01010 00011 10011 26.2 21.7 20.0 5.8 5.5 4.5 3.1 \\
01000 & 11011 00111 01011 01110 10010 11111 2.7 2.1 1.9 1.7 1.3 1.1 1.0 \\
01100 & 10111 01111 11100 01001 11001 0.4 0.4 0.2 0.2 0.1 0.1 \\
0.4 & 0.4 0.2 0.2 0.1 0.1 0.1 \\
\end{align*}
\]
The top ranked models are cubics. The MAP model is a cubic.

---

\(^1\)The analysis below might be improved by restricting analysis to nested models in which lower order terms are included automatically, so \( z \in \{00000, 10000, 11000, ..., 11111\} \).
5. Prior Elicitation

Think about the prior! When Bayesian reasoning leads to nonsensical answers, it is almost always the result of careless prior specification. The issue is obviously important when the data are only weakly informative of the parameter. The problem of prior specification becomes acute in high dimensional problems.

*Things to think about when building a prior...*

- Is there a key scientific hypothesis or parameter? If so we may wish to construct a prior which is non-informative with respect to this hypothesis/parameter. [example - the span variable in Radiocarbon dating]
- Is the parameter \( \theta \) generated by some process we can model? If so then the distribution over \( \theta \) determined by the process is the prior. [example - the ancestry of a sample of individuals from a population is modeled by the Kin]
- Is there some physically interpretable function \( f(\theta) \) of the parameter? The distribution of \( f(\theta) \) is determined by the prior so the prior is constrained to realise a priori plausible \( f \)-values. [example - choosing priors for logistic effects]
- How reliable is the information you are using to build a prior? If it is unreliable, you may wish to downweight it, taking care to ensure that carelessly imposed prior structure doesn't overwhelm data-information for parameters which are poorly informed by the data [this leads to objective Bayes]
- Is the number of things you don't know one of the things you don't know? In this case you may need to put a prior on the number of unknowns! [example - galaxy data and reversible jump]

*And once you have built a prior...*

- The prior density you write down is meant to model your prior knowledge. Once you are done, simulate the prior, and check the realised samples and physically meaningful functions of the samples are distributed as intended.
- It isn't necessary (or even sensible) to analyse the data with just one prior. We typically check results are insensitive to a range of priors representing different states of knowledge.

5.1. Using Simulation to check a prior. Example I used the Challenger data to illustrate bridge estimation and model averaging. The linear predictor is \( \eta_i = \beta_1 + \beta_2 x_i \), with \( x_i = \text{scale(temp)}[i] = (\text{temp}[i] - \text{mean(temp)})/\text{sd(temp)} \) the scaled temperature. Consider the logistic link setting. The prior I used was \( \beta_1, \beta_2 \sim N(0, 3^2) \). Is this remotely plausible?

One quantity we might have some feeling for is \( \mu(\eta) \), the probability for O-ring failure. We know there was a failure, we know they dont always fail. The probability for failure shouldn't vary too sharply with temperature but must depart substantially from zero and one. We simulate \( \beta_1, \beta_2 \sim N(0, v) \) for a few values of \( v \) and look at \( \mu(\beta_1 + \beta_2 x) \) as a function of \( x \).

5.1.1. Careful prior elicitation avoids hazard for multiple testing (yes, in Bayesian Inference!) The following thoughts are lifted from Cox *Principles of Statistics*. Suppose we have a baseline model \( M_0 \) and \( K \) alternatives \( M_k \), \( k = 1 \ldots K \), which are tested using \( K \) data sets (one each) and we report the model with the smallest \( P \)-value, \( P' \) say. This is similar to the situation we get when surveying the literature. If we use \( p \)-values then there is a correction to allow for multiple testing. Under \( M_0 \), \( P(P' < \alpha) = 1 - (1 - \alpha)^K \) which is approximately \( K\alpha \) if \( \alpha \) is small and \( K \) is
not too large. It follows that we should target significant level $\alpha/K$ to achieve level $\alpha$ overall. This is called a Bonferroni correction.

If we have $K$ Bayes factors $B_{k,0}, k = 1...K$ and report the largest there is a similar issue, which solves itself by careful consideration of the prior $\pi_M(m)$ over models. This may lead to some down-weighting of the evidence.

For example, if the $K$ models are competing alternatives, we might assert $\pi_M(0) = 1/2, \pi_M(k) = 1/2K, k = 1..K$. The posterior odds $p(y|k)\pi_M(k)/p(y|k)\pi_M(0)$ for model $k$ over model 0 are then $B_{k,0}/K$, similar to the Bonferroni correction. When many models are compared against a baseline, the posterior odds make a better model selection tool than the Bayes Factor, and care must be taken over the choice of $\pi_M$.

Note that data dredging is OUT in both settings. If the test data informs the hypothesis, then the prior depends on the data and the posterior is not given by Bayes formula.

5.2. Prior Elicitation case study: Radiocarbon dating. We dig a hole in the ground. We dig through a habitation layer. Above and below this layer there is no evidence for dwellings on the site. We take $n$ charcoal samples from the habitation layer and get them radiocarbon dated. The key questions of interest are, when was the site settled, for how long was it settled and when did the settlement cease?

There is a suggestion that this camp was settled for just weeks or months rather than years.

The data we have are 7 dates from a ancient settlement in NZ. The strata are very clear. As we dig down we encounter artifact-free sand, then obvious settlement remains, and then below that artifact-free earth. The researchers are very confident (before see the RCD’s) there was no settlement prior to 1000 years BP (in fact 700 BP would still be reasonable), and that the settlement had been abandoned prior to 500BP.

5.2.1. Observation model. An uncalibrated radiocarbon age $y_i$ is, for $i = 1, 2, ..., n$, a noisy biased measurement of the unknown true age $\theta_i$ of the $i$’th dated specimen.
The observation model for the data is

\[ y_i = [\mu(\theta_i) + \zeta_i] + \epsilon_i \]

with

\[ \zeta_i \sim N(0, \sigma_c(\theta_i)^2) \]

modeling uncertainty in \( \mu \) and

\[ \epsilon_i \sim N(0, \sigma_i^2) \]

the error associated with the measurement of specimen \( i \) itself. The likelihood is

\[ L(\theta; y) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{- (y_i - \mu(\theta_i))^2}{2(\sigma_c(\theta_i)^2 + \sigma_i^2)}\right) \]

5.2.2. Priors. A simple uniform prior \( \pi_u(\theta) \) for \( \theta = (\theta_1, ..., \theta_n) \) might set lower and upper bounds \( L = 500 \) and \( U = 1000 \) on the true ages of the specimen, and assert that all other values are equally probable:

\[ \pi_u(\theta) \propto \prod_i \mathbb{I}(L \leq \theta_i \leq U). \]
We could estimate the start and end of settlement using \( \theta_{(1)} = \min(\theta) \) and \( \theta_{(n)} = \max(\theta) \) the oldest and most recent dates, and estimate the span using
\[
S_u = \max(\theta) - \min(\theta),
\]

Does this prior meet our elicitation criteria? (A) can we think of the dates \( \theta_i, i = 1, 2, \ldots, n \) as generated by some “physical” process? (B) arrival date, departure date, length of stay are linked to scientific questions - make sure prior is non-informative on these variables.

Does \( \pi_u \) meet these criteria? Certainly not (A). For (B) the marginal prior density of \( \theta_{(n)} \) (order statistics of a uniform) is
\[
f(\theta_{(n)}) \propto \left[\frac{\theta_{(n)} - L}{U - L}\right]^{n-1}
\]
which is not uniform. You may prefer to check using simulation.

These marginal priors \( \pi_u(\theta_{(1)}), \pi_u(\theta_{(n)}) \) and \( \pi_u(S_u) \) are strongly informative and we don’t want this.

5.2.3. A prior from process generating \( \theta \). The settlement starts at an unknown age \( \psi_2 \) and ends at a time \( \psi_1 \). These are some of the things our prior knowledge directly informs so we should control the information we provide by making them an explicit part of our prior. The span is
\[
S_s = \psi_2 - \psi_1.
\]

Between \( \psi_2 \) and \( \psi_1 \) the inhabitants generated datable material in the interval \( dt \) with probability \( \lambda dt \). This gives a Poisson process generating datable material. The material was thinned (by erosion etc) so it is datable with probability \( p \). Assume \( p \) is a constant not depending on time (so older material is equally likely to survive). The dates \( \theta \) are a realisation of a Poisson process with rate \( p \lambda \). We choose the number of dates so we condition on \( n \). It follows that \( \theta_i \sim U(\psi_1, \psi_2) \) (conditioning a Poisson process on the number of events gives a uniform distribution for the event times). Our prior for \( \theta|\psi \) is therefore
\[
\pi_s(\theta|\psi) = (\psi_2 - \psi_1)^n.
\]

We have to specify the prior for \( \psi \). We would like the span to be uniform. The prior
\[
\pi_s(\psi) \propto \frac{1}{(U - L - (\psi_2 - \psi_1))}
\]
has a uniform distribution on values of \( \psi_2 - \psi_1 \), so this is non-informative with respect to the span. Since \( \pi_s(\theta, \psi) = \pi_s(\theta|\psi)\pi_s(\psi) \) we get
\[
\pi_s(\theta, \psi) \propto \frac{1}{(\psi_2 - \psi_1)^n} \frac{1}{(U - L - (\psi_2 - \psi_1))}
\]
models the prior information we actually have. These marginal priors $\pi_s(\theta(1))$, $\pi_s(\theta(n))$ and $\pi_s(S_u)$ better represent the prior information we have. The prior on span is uniform, and the priors $\pi_s(\theta(1))$, $\pi_s(\theta(n))$ distribute probability mass more evenly over the parameter domain.

5.2.4. Sampling the posterior. The two posterior distributions have the same likelihood, and prior models $m = 1$ ($\pi_u(\theta)$) and $m = 2$ ($\pi_u(\theta, \psi)$).

$$\pi_u(\theta|y) \propto L(\theta; y)\pi_u(\theta)$$

and

$$\pi_s(\theta, \psi|y) \propto L(\theta; y)\pi_s(\theta, \psi).$$

I used simple random walk MH taking each variable in turn. For example, targeting $\pi_s(\theta, \psi|y)$, propose $\theta_i' \sim U(\psi_1, \psi_2)$ and accept $w_p \alpha(\theta', \psi|\theta, \psi) = \min\{1, \frac{L(\theta'; y)}{L(\theta; y)}\}$ since $\pi_s(\theta, \psi)$ depends only on $\psi$ which hasnt changed.

For $\psi_1$, $\psi_1' \sim U(L, \min(\theta))$, set $\psi' = (\psi_1', \psi_2)$ and

$$\alpha(\theta, \psi'|\theta, \psi) = \min\{1, \frac{\pi_s(\theta, \psi')}{\pi_s(\theta, \psi)}\}$$

since the likelihood doesn’t depend on $\psi$. The step for $\psi_2$ proposes $\psi_2' \sim U(\max(\theta), U)$ and accepts with the same formulae as $\psi_1$ (but now $\psi' = (\psi_1', \psi_2')$). See attached R-code for details.

5.2.5. Summarising the results. There are many things we test for. Here I focus on the span, and the two models $m = 1, 2$. I ran the MCMC, and checked convergence. I obtained HPD estimates for the span: Model 2 (shrink/$\pi_s$), [0,160]; Model 1 (unif/$\pi_u$), [70,160]. The marginal posterior histograms differ in shape and support. We would like to compare the two models using a Bayes Factor. This is not testing a pre-defined scientific hypothesis. We are using the test to establish our prior modelling is sound (perhaps for use in other contexts). We cannot use our preferred Bridge estimator for the Bayes factor itself, because the two models have different parameter spaces $\theta$ and $\theta, \psi$ but we could use it for the marginal likelihoods, as the prior and posterior at least have the same parameter space. But we can use the Bridge estimator for the marginal likelihood. We check the result is stable across multiple runs, with very large sample sizes (ESS $\gg 1000$).
If in model $m = 1, 2$ the posterior samples for $\theta$ are $\theta^{(m,t)}, t = 1, 2, ..., T$ and the prior samples are $\tilde{\theta}^{(m,t)}$ then the bridge estimator identity (using $h = 1/\sqrt{\pi \pi L}$) for $p(y|m)$ is

$$p(y|m) = \frac{E_{\theta|m}(\pi_m(\theta)L(\theta; y)h(\theta))}{E_{\theta|m}(\pi_m(\theta)h(\theta))} = \frac{E_{\theta|m}(L(\theta; y)^{1/2})}{E_{\theta|m}(L(\theta; y)^{-1/2})}$$

and the estimator itself is

$$\hat{p}_m = \frac{\sum_{t=1}^{T} L(\hat{\theta}^{(m,t)}; y)^{1/2}}{\sum_{t=1}^{T} L(\theta^{(m,t)}; y)^{-1/2}}$$

We find $\hat{p}_1 \simeq 4 \times 10^{-21}$ and $\hat{p}_2 \simeq 8 \times 10^{-19}$ so the Bayes factor for shrinking over uniform priors is about $B_{2,1} \simeq 200$, so the shrinkage prior is clearly favoured.

5.2.6. Radiocarbon dating example: Conclusions. We don’t have much data (7 noisy numbers!) so the conclusions show some sensitivity to the choice of prior. The process-model based prior allows very small spans close to zero, while the uniform prior rules them out. This is clearly a case where we don’t want the prior to impose structure we can’t support on prior grounds. We might go on to test for $\psi_1 = \theta_1 = ... \theta_n = \psi_2$ and consider goodness of fit.
6. ABC: Why work hard to fit a model that is wrong anyway?

Approximate Bayesian Computation is a Monte Carlo scheme for generating samples approximately distributed according to the posterior. It is particularly useful when the likelihood is intractable, but easy to simulate.

This is actually quite a common scenario, as the Likelihood $L(\theta; y)$ is normalised over $\theta$. If the observation model is $p(y|\theta)$ proportional to $\hat{p}(y, \theta)$ (as a function of $y$) then

$$L(\theta; y) = \frac{\hat{p}(y, \theta)}{c(\theta)}$$

with

$$c(\theta) = \int \hat{p}(y, \theta) dy$$

and $c(\theta)$ may be intractable. These problems are called doubly intractable. MCMC won’t work, as we can’t calculate the acceptance probability $\alpha(\theta'|\theta)$.

6.1. Rejection, ABC-style. The basic ABC algorithm approximates the following variant of rejection. Suppose $y$ and $\theta$ are both discrete so $p(y|\theta) \leq 1$. The following algorithm simulates $\theta \sim \pi(\theta|y)$ where (as usual) $\pi(\theta|y) \propto p(y|\theta)\pi(\theta)$.

1. Simulate $\theta \sim \pi(\theta)$ and $y' \sim p(y'|\theta)$.

2. If $y' = y$ return $\theta$ and stop, otherwise goto (1).

The probability $y' = y$ is $p(y|\theta)$, so if $\Theta_{\text{rej}}$ is the value returned by this algorithm then

$$\Pr(\Theta_{\text{rej}} = \theta) = \pi(\theta)p(y|\theta) + \pi(\theta)p(y|\theta)\sum_{\theta'} \pi(\theta')(1 - p(y|\theta')) + ...$$

$$= \pi(\theta)p(y|\theta)(1 + (1 - m(y)) + (1 - m(y))^2 + ... )$$

$$= \frac{\pi(\theta)p(y|\theta)}{m(y)} \text{ with } m(y) = \sum_{\theta} \pi(\theta)p(y|\theta).$$

summing over all sequences of rejections returning $\Theta_{\text{rej}} = \theta$.

6.2. ABC. The idea here is that if $y'$ is “close” to $y$ then $\pi(\theta'|y')$ should be close to $\pi(\theta|y)$. To measure “close” we choose some summary statistics $S(y) = (S_1(y), ..., S_J(y))$ of the data, which inform $\theta$, and which we want our model to reproduce well. We define a distance measure (often Euclidean) $d(s', s)$ between $J$-component vectors $s' = S(y')$, $s = S(y)$, and a threshold $\delta$, and modify the algorithm above.

1. Simulate $\theta \sim \pi(\theta)$ and $y' \sim p(y'|\theta)$.

2. If $d(S(y'), S(y)) < \delta$ return $\theta$ and stop, otherwise goto (1).

If $\Theta_{\text{abc}}$ is the output then $\Theta_{\text{abc}} \sim \pi(\theta|d(S(Y), S(y)) < \delta)$. If $S$ is a sufficient statistic then

$$\pi(\theta|Y = y) = \pi(\theta|S(Y) = s(y)),$$

so $\Theta_{\text{abc}} \overset{D}{\to} \pi(\theta|y)$ as $\delta \to 0$ in that case.

Example: Data model $y_i \sim \text{Poisson}(\Lambda = 2)$, $i = 1, 2, ..., n$ with $n = 5$. Prior $\Lambda \sim \Gamma(\alpha = 1, \beta = 1)$. Take $S(y) = \bar{y}$, $d(\bar{y'}, \bar{y}) = |\bar{y'} - \bar{y}|$ and $\delta = 0.5, 1$.

1. Simulate $\Lambda \sim \Gamma(\alpha, \beta)$ and $y'_i \sim \text{Poisson}(\Lambda)$, $i = 1, 2, ..., n$.

2. If $|\bar{y'} - \bar{y}| < \delta$ return $\theta$ and stop, otherwise goto (1).
6.3. Regression adjustment of ABC samples. ABC generates samples \((\theta^{(t)}, y^{(t)}), t = 1, 2, ..., T\) from the joint distribution \((\theta, y') \sim \pi(\theta, y')\). We want pairs \((\theta, y) \sim \pi(\theta, y)\) with \(y\) fixed. Suppose \(s = S(y)\) is sufficient.

Assume (1) shifting the data, \(y\) to \(y'\), shifts the posterior mean \(\mu(s) = E(\theta|S(Y) = s)\) but has no other effect on the distribution of \(\theta\) and \((2)\) for \(d(s', s)\) small, the linear approximation

\[
\mu(s') \simeq \beta_1 + \beta_2(s' - s)
\]

is justified. Clearly \(\beta_1 = \mu(s)\) is the posterior mean at the data.

Now \(\epsilon \sim \theta - \mu(s)\) is a mean zero rv with a distribution not depending on \(s\). If we knew \(\mu(s)\), we could simulate \(\theta|y'\) by simulating \(\epsilon\) and setting

\[
\theta' = \mu(s') + \epsilon.
\]

where \(s' = S(y')\).

We have lots of pairs \(\theta^{(t)}, S(y^{(t)})\) and we can use them to estimate our local linear approximation to \(\theta' = \mu(s') + \epsilon\),

\[
\theta^{(t)} = \beta_1 + (S(y^{(t)}) - s)\beta_2 + \epsilon^{(t)}.
\]

Our approach is to estimate \(\hat{\beta}_1, \hat{\beta}_2\) and then set

\[
\theta_{\text{adj}}^{(t)} = \theta^{(t)} - (S(y^{(t)}) - s)\hat{\beta}_2
\]

since this is an estimate of \(\beta_1 + \epsilon^{(t)} = m(s) + \epsilon^{(t)}\), which is a sample from the posterior under our assumptions.
Example: We did exactly this for the Poisson example above. Worked well. See the figure and R-code for this lecture.

6.4. **ABC example: the Ising model.** - a doubly intractable model.

Denote by \( \Omega \) = \( \{0,1\}^{n^2} \) the set of all binary images \( Y = (Y_1, Y_2, ..., Y_{n^2}) \), \( Y_i \in \{0,1\} \), where \( i = 1, 2, ..., n^2 \) is the cell index on the square lattice of image cells. Let \( \#y \) give the number of disagreeing neighbors in the binary image \( Y = y \).

The **Ising model** is the following distribution over \( \Omega \):

\[
p(y|\theta) = \frac{\exp(-\theta \#y)}{c(\theta)}.
\]

Here \( \theta \) is a positive **smoothing parameter** and \( c(\theta) \)

\[
c(\theta) = \sum_{y \in \Omega} \exp(-\theta \#y).
\]

is a normalizing constant which we can’t compute for \( n \) at all large. However the Ising model is easy to sample for moderate \( n \)-values using MCMC. Here are 3 samples \( Y \sim p(y|\theta) \). Technically these samples are not exactly distributed according to \( p(y|\theta) \) (convergence) but we can make them as good as we need by taking long MCMC runs.

We can sample \( Y \sim p(y|\theta) \) using MCMC: Suppose \( Y^{(t)} = y \).

[Step 1] Choose an update, something simple. Choose a cell \( i \sim U \{1, 2, ..., n^2\} \). Set \( y'_i = 1 - y_i \) and \( y'_j = y_j \) for \( j \neq i \). Notice that \( q(y'|y) = q(y|y') = 1/n^2 \) for \( y', y \) differing at exactly one cell.

[Step 2] Write down the algorithm. Let \( Y^{(t)} = y \). \( Y^{(t+1)} \) is determined in the following way.

1. Simulate \( y' \sim q(y'|y) \) as above, and \( u \sim U(0, 1) \).
2. If \( u < \alpha(y'|y) \) set \( Y^{(t+1)} = y' \) and otherwise set \( Y^{(t+1)} = y \).

[Step 3] Calculate \( \alpha \). The \( q \)’s cancel as usual, so

\[
\alpha(y'|y) = \min \left\{ 1, \frac{p(y'|\theta)q(y|y')}{p(y|\theta)q(y'|y)} \right\} = \min \{1, \exp(-\theta(\#y' - \#y))\}
\]

It is clear the algorithm is irreducible (\( q \) is irreducible and \( \alpha \) is never zero) and aperiodic (rejection is possible), so it is ergodic for \( p(y) \).
An implementation of this algorithm in R is available in the code for this lecture. Some samples produced using this code are shown above.

Suppose we have data $Y$ which is an $n \times n$ binary matrix and our observation model for $Y$, $Y \sim p(y|\theta)$ is the Ising model. Our goal is to estimate $\theta$. The posterior is

$$
\pi(\theta|y) = \frac{L(\theta; y)\pi(\theta)}{m(y)}
$$

We are not immediately concerned about the intractable $m(y)$ (doesnt depend on $\theta$, cancels in MCMC) but

$$
L(\theta; y) = p(y|\theta)
$$

with

$$
p(y|\theta) = \exp(-\theta \#y)/c(\theta)
$$

so the Likelihood depends on $c(\theta)$, an intractable function of $\theta$ itself.

Our ABC algorithm sampling $\theta \sim \pi(\theta|y)$ approximately is as follows. Suppose we have a prior for $\theta$. Given the scale of $\theta$, Exp(2) is a natural generic prior. We take $S(y) = \#y$, which is actually sufficient for $\theta$.

1. Simulate $\theta \sim \text{Exp}(2)$ and $y' \sim \exp(-\theta \#y')/c(\theta)$.
2. If $|\#y' - \#y| < \delta$ return $\theta$, otherwise, goto (1).

We implemented this (see attached R, 8x8 Ising, $\theta = 0.8$) and estimated the posterior densities in the figure.

The distribution converges to something stable as $\delta \to 0$. The regression adjustment for $\delta = 0.1$ corrects its distribution to agree with that for $\delta = 0.05$. 
6.5. Population genetics and ABC. The woodmouse data $y_D$ is a set of $L = 15$ sequences of a $n = 965$-base mitochondrial gene of the woodmouse (so $y_D$ is a $L \times n$ matrix). Each mouse has a 965 character sequence made up of A,C,G and T characters (and some gaps, "-", corresponding to missing DNA).

How are these mice evolving? Each mouse has an ancestry, and as we trace back the ancestry, they are related. The "family tree" is a genealogy - a rooted tree $g$ say, with 15 leaves. The branches of their genealogy correspond to an evolving lineage of mice, and as we trace down one of these lineages, the mtDNA of generations of mice evolves through mutation.

<table>
<thead>
<tr>
<th>No</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>305</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>304</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>306</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>0906S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>0908S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>0909S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>0910S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>0912S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>0913S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>1103S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>1007S</td>
<td>&quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; ...</td>
</tr>
<tr>
<td>1114S</td>
<td>&quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; ...</td>
</tr>
<tr>
<td>1202S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>1206S</td>
<td>&quot;a&quot; &quot;t&quot; &quot;t&quot; &quot;c&quot; &quot;g&quot; &quot;a&quot; &quot;a&quot; &quot;a&quot; ...</td>
</tr>
<tr>
<td>1208S</td>
<td>&quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; &quot;n&quot; ...</td>
</tr>
</tbody>
</table>

Sample genealogy of 15 mice from woodmouse data.

6.5.1. Mutation model. Label the leaves of the tree 1,...,L and the tree nodes $L + 1,...,2L - 1$ in order of increasing age. Let $R = 2L - 1$ be the label of the root node. Each node $i \in \{1,...,R − 1\}$ has a unique parent, $\sigma(i)$ say. Let $\tau_i, i \in 1,...,R$ give the age of the $i$th node [increasing into the past, in units yet to be defined], with $\tau_i = 0, i = 1,...,L$ at the leaves. Let $t_i = \tau_R − \tau_i$ be time increasing from $t_R = 0$ at the root toward the present.

For $i \in 1,...,L$, denote by $y_i = (y_{i,1},...,y_{i,n})$, $y_{i,j} \in \{A,C,G,T, −\}$ the DNA character sequence for the $i$th mouse. The ancestral nodes on the tree correspond to mice which existed in the past. They also have (unobserved) sequences, $y_i, i \in L + 1,...,R$.

The sequences evolve down the lineages undergoing mutation. Suppose base mutations occur independently at rate $\lambda$ at each site. This process determines our observation model for the data $y_D = (y_1,...,y_L)$ and missing data $y_A = (y_{L+1},...,y_R)$. Take two nodes on the tree, node $j$ and its parent $i = \sigma(j)$ and consider the characters $y_{i,k}$ and $y_{j,k}$ at site $k$ in the parent and child respectively. Many generations pass between $t_i$ and $t_j$.

Denote by $P(t) = [P_{a,b}], a, b \in \{A,C,G,T\}$ the $4 \times 4$ transition probability matrix with entries

$$P_{a,b}(t) = \Pr(y_{j,k} = b|y_{i,k} = a, t = t_j − t_i).$$
This the probability for a mutation from base $a$ to base $b$ over $t$ units of time. We assume the mutation rate $\lambda$ is the same at all times, in all mice, at all sites, so $P$ does not depend on site $k$, or node label $i,j$, and only on the elapsed time $t_j - t_i$ not the time itself. This leads to

$$P_{a,k}(t) = \frac{1}{4} - \exp(-4\lambda t/3)/4 \quad a,b \in \{A,C,G,T\}$$

and

$$P_{a,a}(t) = \frac{1}{4} + 3 \exp(-4\lambda t/3)/4 \quad a \in \{A,C,G,T\}.$$  

If we let this process run for a long time the distribution over bases will converge to a uniform equilibrium distribution $\pi^{EBF} = (1/4, 1/4, 1/4, 1/4)$ over bases at each site,

$$\pi^{EBF}_b = \lim_{t \to \infty} \Pr(y_{j,k} = b | y_{i,k} = a, t)$$

satisfying $\pi^{EBF} = \pi^{EBF} P(t)$.

This model for how the base at each site on each lineage changes independently in time is called the Jukes-Cantor model.

6.5.2. Simulating sequence data on a fixed tree-genealogy. In order to simulate sequences $y^*_A,y^*_D$ according to the mutation model on a given tree-genealogy

1. Start at the root. Simulate $y'_{R,k} \sim \pi^{EBF}$ for $k = 1,...,n$.
2. Let $i = R$. The nearest node to the root, node $j = R - 1$, is a child node of $i$. Now for each site $k = 1,...,n$, simulate

$$y'_{j,k} \sim P_{y'_i,k}(t_j - t_i),$$
where \( P_{x,:}(t) \) is the \( i \)th row of \( P(t) \).

3. Repeat step (2) for \( i = R - 2, \ldots, L + 1 \) and \( j \) the nearest child of \( i \) to simulate sequences at all nodes.

Parents are simulated before children by the time ordering.

The sequences \( y_D' = (y_1', \ldots, y_L') \) simulated at the leaves are our simulated data. They don’t include gaps. We should allow for this when we calculate our ABC statistics \( S(y_D) \) and \( S(y_D') \).

6.5.3. Genealogy model. The sequence data were generated by simulating the mutation process down the branches of the tree representing the genealogy of the mice. This unknown tree, \( g \) say, is one of the things we don’t know so we need a prior model for the unknown tree. This is a good example of model-based prior elicitation.

Suppose there are \( N \) individuals in the population, and \( N \) is constant in time.

As we move back in time (\( \tau' \), with \( \tau' = 0 \) at the leaves, and increasing into the past), each individual chooses their parent (mtDNA is single-sex) UAR from the \( N \) individuals in the previous generation.

The number of generations \( \tau' \) back to coalescence is \( \text{Geometric}(1/N) \).

Rescale time from units of single generations to units of \( N \)-generations, \( \tau = \tau'/N \).

\[
\Pr(\tau \leq z) = 1 - (1 - N^{-1})^{Nz},
\]

so

\[
\Pr(\tau \leq z) \rightarrow 1 - \exp(-z) \quad \text{as} \quad N \rightarrow \infty.
\]

The time back to coalescence \( \tau \sim \text{Exp}(1) \) approximately, when \( N \) is large, so the ancestry of each pair of individuals coalesces independently at instantaneous rate equal one.

This process determines our prior on genealogical trees. If at any time we have \( k \) lineages, each pair of lineages “coalesces” independently at rate 1 for a total rate \( k(k-1)/2 \). The pair that coalesce is equally likely to be any pair.

6.5.4. Simulating a genealogy. This process generates a random tree \( g = (E, \tau, V) \) with edge set \( E \), vertex set \( V = \{1, \ldots, R\} \) and node ages \( \tau = (\tau_1, \ldots, \tau_R) \).

1. Start with \( k = L \) lineages, one for each leaf. Work backwards in time. Set \( \tau_1, \ldots, \tau_L = 0 \). The set of active lineages is \( S = \{1, \ldots, L\} \). Initialise the edge set \( E = \{\} \).

2. Simulate \( \delta_{k+1} \sim \text{Exp}(\lambda_k) \) where \( \lambda_k = k(k-1)/2 \) and set \( \tau_{k+1} = \tau_k + \delta_{k+1} \).

3. Pick a pair of nodes \( i, j \) uniformly at random from \( S \) to coalesce. Update the edge set \( E \rightarrow E \cup \{(i, k+1), (i, k+1)\} \) and the lineage set \( S = (S \setminus \{i, j\}) \cup \{k+1\} \).

4. Set \( k \rightarrow k + 1 \). Repeat steps 2, 3 and 4 until we reach \( k = 2L - 1 \) (ie the root index \( k = R \)) where we will have \( S = \{R\} \).

6.5.5. Bayesian inference for the scaled mutation rate. Let \( \mu = \lambda N \), so \( \mu \) is the mutation rate for time measured in units of \( N \) generations.

Consider the problem of estimating \( \mu | y_D \). The posterior, including auxiliary “parameters” \( g \) and \( y_A \), is

\[
\pi(\mu, y_A, g | y_D) \propto p(y_D, y_A | \mu, g) \pi_C(g) \pi(\mu).
\]
Here $p(y_D, y_A | \mu, g)$ is the probability to simulate the set of sequences $(y_1, ..., y_R)$ on the given tree $g$ with a given mutation rate $\mu$. $\pi_C(g)$ is the probability density our coalescent simulation realises tree $g$, and $\pi(\mu)$ is a prior on the rate $\mu$.

Notice that we don’t need to write down formulae for $p(y_D, y_A | \mu, g)$ and $\pi_C(g)$. We only need the algorithms that we wrote own above that tell us how to simulate $g$, $y_A$ and $y_D$. This frees up our modeling a great deal.

6.5.6. Prior for $\mu$. Time on the tree is measured in units of $N$ generations. A typical mutation rate per site per generation for mice is about $6 \times 10^{-7}$. Modern populations are smaller, but I will suppose the sample population size was around 10K (obviously I should allow significant uncertainty here). This gives a prior mean equal 0.006. I take for my prior distribution $\mu \sim \text{Gamma}(\text{mean} = 0.006, \text{sd} = 0.012)$.

6.5.7. ABC for $\mu \sim \pi(\mu | y)$. We will focus on estimating the mutation rate $\mu$. The summary statistic we use informs $\mu$ but is not informative of $g$, so we cannot hope to recover $g$ [estimating $g$ would be very hard - can you think why?].

To simulate the posterior distribution for $\mu | y$ I will simulate the distribution for $g, y_A, \mu | y$ and ignore $g$ and $y_A$.

We will use the number of “segregating sites” as our ABC statistic. This is the number of columns of the data $Y$ which contain at least one mutation. Let $s = S(y)$ denote this statistic computed on the data (In fact $s = 50$ of the columns show mutations).

6.5.8. ABC algorithm:

1. Simulate $\mu \sim \pi(\mu)$ and $g \sim \pi_C(g)$. Simulate $(y', y_A) \sim p(y', y_A | \mu, g)$ and calculate $s' = S(y')$.

2. If $|s - s'| < \rho$ return $\mu$ and otherwise goto 1.

Calling this lots of times generates $\mu \sim p(\mu | y)$, approximately. Let $(\mu^{(t)}, s^{(t)}), t = 1, ..., T$ denote the sampled $\mu$-values, and the associated ABC statistic $s$.

We ran this algorithm with $\rho = 50$ (large, of the same order as $s$) and $\rho = 10$. We regressed

$$E(\mu^{(t)}) = a + b(s^{(t)} - s),$$

and made the regression adjustment

$$\mu_{\text{adj}}^{(t)} = \mu^{(t)} + (s - s^{(t)})b.$$
(Left) ABC v. MCMC with low precision (large) cut-off $\rho = 50$ on $|S(y) - S(y')| < \rho$. The samples (histogram) resemble the Gamma-prior $\pi(\mu)$. The MCMC estimate of $\pi(\mu | y)$ (thick black line) is taken as the truth (though the prior is uniform in that simulation, so the posterior is expected to differ slightly). The regression adjusted estimate of $\pi(\mu | y)$ is greatly improved and resembles the MCMC “truth”. The green line is the prior mean.

(Right) ABC v. MCMC with higher precision (small) cut-off $\rho = 10$ on $|S(y) - S(y')| < \rho$. The samples (histogram) are now more representative of the posterior but the gain from the regression adjustment is slight.

6.5.9. *Mouse Genealogy case study: Conclusions.* This is a relatively complex model with substantial missing data. However simulating coalescent trees and synthetic data is straightforward and this makes ABC a good option for fitting complex models. It would be easy to change the simulation algorithm for mutations and that for trees to be much more realistic without complicating the analysis. In fact this type of problem was one of the motivating problems for the original development of ABC.
7. Reversible-Jump MCMC

7.1. What problem does RJMCMC solve? We begin the lead up to the reversible jump algorithm. What problem does RJMCMC solve? In the joint distribution of the model and the parameter

\[ \pi(\theta, m|y) \propto p(y|\theta, m)\pi(\theta|m)\pi(m), \quad \theta \in \Omega, m \in \{1, 2, ..., M\} \]

the dimension of the parameter \( \theta \) may vary depending on the model. For example, \( m=1 \) then \( Y = \alpha + \epsilon \) and \( \theta = \alpha \in \mathbb{R} \), and \( m=2 \) then \( Y = \alpha + \beta x + \epsilon \) and \( \theta = (\alpha, \beta) \in \mathbb{R}^2 \) and

\[ (\theta, m) \in \Omega, \quad \Omega = (\Omega_1 \times \{1\}) \cup (\Omega_2 \times \{2\}). \]

"The number of things we don't know is one of things we don't know". The MCMC algorithm must jump from between spaces of different dimension.

7.2. MCMC proposals restricted to a parameterised curve or subsurface of the parameter space. Suppose we are doing MCMC targeting \( \pi(\theta) \). How does the proposal step really work? Given \( \theta \) we simulate a rv \( u \sim g(u) \) say and set \( \theta' = \psi_1(\theta, u) \). We choose the function \( \psi_1 \) and the density \( g(u) \) so that we reproduce our intended target density at \( \theta' = \psi_1(\theta, u) \):

\[ q(\theta'|\theta) = g(u) \left| \frac{\partial \theta'}{\partial u} \right|^{-1} \]

Example: our proposal \( \theta' = \theta + a(2u - 1) \) with \( u \sim U(0,1) \), \( a \) fixed, is

\[ g(u) = 1_{0\leq u \leq 1}, \quad \psi_1(\theta, u) = \theta + a(2u - 1), \quad q(\theta'|\theta) = \frac{1_{0\leq u \leq 1}}{2a}. \]

In order to reverse the move we would choose

\[ u' \sim g(u') \]

and set

\[ \theta = \psi_1(\theta', u'). \]

The \( u' \) that reverses the move satisfies

\[ \theta = \psi_1(\psi_1(\theta, u), u') \quad (\star) \]

The mapping from \( \theta \) to \( \theta' \) is a mapping between pairs

\[ (\theta', u') = \psi(\theta, u) \]

with

\[ \psi = (\psi_1, \psi_2). \]

The function \( \psi_2 \) is determined from \( \psi_1 \) and equation \((\star)\). Example: if \( \psi_1(\theta, u) = \theta + a(2u - 1) \) as before then

\[ \psi_1(\psi_1(\theta, u), u') = \theta + a(2u - 1) + a(2u' - 1) \]

so we need \( \theta + a(2u - 1) + a(2u' - 1) = \theta \) for all \( \theta, \mu \) and the solution is \( u' = 1 - u \).

In our simple example, \( \psi_2 = 1 - u \) so the mapping is

\[ (\theta', u') = (\theta + a(2u - 1), 1 - u). \]

Up till now we chose \( q(\theta'|\theta) \) and found a density \( g(u) \) and a function \( \theta' = \psi_1(\theta, u) \) to simulate it. Let’s just write down \( g \) and \( \psi_1 \) and let \( q \) be whatever it is. This gives us a bit more freedom.
We are making a change of variables in a density so when we check detailed balance we check it holds over all sets \((\theta, \theta') \in A \times B\).

**Claim:** let \((\theta', u') = \psi(\theta, u)\) be an invertible, differentiable mapping with \(\theta, \theta' \in \Omega\) and \(u, u' \in \mathbb{U}\). If

\[
q(\theta'|\theta) = g(u) \left| \frac{\partial \theta'}{\partial u} \right|^{-1} \tag{1}
\]

and

\[
\alpha(\theta'|\theta) = \min \{1, \ r(\theta', u'|\theta, u)\} \tag{2.1}
\]

where

\[
r(\theta', u'|\theta, u) = \frac{\pi(\theta')q(u')}{\pi(\theta)g(u)} \left| \frac{\partial(\theta', u')}{\partial(\theta, u)} \right| \tag{2.2}
\]

then

\[
\int_B \int_A \pi(\theta')q(\theta|\theta')\alpha(\theta|\theta')d\theta d\theta' = \int_A \int_B \pi(\theta)q(\theta'|\theta)\alpha(\theta'|\theta)d\theta' d\theta
\]

for all sets \(A, B \subseteq \Omega\) for which the integrals are defined.

The final relation is detailed balance. If \(P(\theta, \theta') = q(\theta'|\theta)\alpha(\theta'|\theta)\) is the transition probability density (for \(\theta' \neq \theta\)) the DB is

\[
\int_B \int_A \pi(\theta')P(\theta', \theta)d\theta d\theta' = \int_A \int_B \pi(\theta)P(\theta, \theta')d\theta' d\theta
\]

over all sets. This is sufficient for stationarity.

\[
\int_B \pi(\theta')d\theta' = \int_B \left[ \int_\Omega \pi(\theta)P(\theta, \theta')d\theta \right] d\theta'.
\]

Proof: starting on the RHS, make a change of variables from \(\theta'\) to \(u\) (at fixed \(\theta\)). Using Eq (1),

\[
\pi(\theta)q(\theta'|\theta)\alpha(\theta'|\theta)d\theta' d\theta = \pi(\theta)g(u)\alpha(\theta'(\theta, u))\theta)du d\theta.
\]

Suppose WLOG that \(r(\theta', u'|\theta, u) \leq 1\). Then the LHS is

\[
\pi(\theta)g(u)\alpha(\theta'(\theta, u))\theta)du d\theta = \pi(\theta)\left[ \pi(\theta)g(u) \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)} \right] \left| \frac{\partial(\theta', u')}{\partial(\theta, u)} \right| dud\theta
\]

\[
= \pi(\theta')g(u') \left| \frac{\partial(\theta', u')}{\partial(\theta, u)} \right| dud\theta
\]

\[
= \pi(\theta')g(u')du' d\theta'
\]

since we have exactly the Jacobian we need for the change of variables. Considering now the LHS of detailed balance

\[
\pi(\theta')q(\theta'|\theta)\alpha(\theta|\theta')d\theta d\theta' = \pi(\theta')q(u')du' d\theta'
\]

and we have equality if the integration domains are equal. This must hold as all the mappings are invertible.

This works if the Jacobian is non-singular, so that the change of variables is well defined. The condition \(\dim(\theta', u') = \dim(\theta, u)\) is called “dimension matching”.

**Example:** “Random walk on a log scale”. Suppose we are targeting \(\theta \sim \exp(1)\) and we use the proposal

\[
u \sim U(1/2, 2), \quad \theta' = u\theta \quad \text{so that} \quad (\theta', u') = (u\theta, 1/u)
\]
Here \( g(u) = 1/(2 - 0.5), 0.5 < u < 2 \) and \( \dim(\theta', u') = \dim(\theta, u) = 2 \) so dimensions match. The Jacobian is \( 1/u \) since 
\[
\left| \frac{\partial(\theta', u')}{\partial(\theta, u)} \right| = \begin{vmatrix} u & 0 \\ \theta & -1/u^2 \end{vmatrix} = 1/u,
\]
The algorithm is as follows. If \( X_t = \theta \) then
1) simulate \( u \sim U(1/2, 2) \) and set \( \theta' = u\theta; \)
2) with probability \( \alpha(\theta'|\theta) = \min \left\{ 1, u^{-1}e^{-\theta'+\theta} \right\} \), set \( X_{t+1} = \theta' \) and otherwise \( X_{t+1} = \theta. \)
Factors of \( g(u)/g(u') \) cancel in \( \alpha. \) This proposal is useful if simulating a density which is peaked or diverges at a boundary. This whole framework is a useful generalisation of MCMC.

Exercise: show that the Jacobian for the SRW proposal
\[
u \sim U(-a, a), \quad (\theta', u') = (\theta + u, -u)
\]
is equal one.

Exercise: in the polynomial regression example in lecture 5, I gave the MCMC updates for \( z \) and \( \theta. \) The update for \( \sigma \) is
\[
u \sim U(\delta, 1/\delta), \quad \sigma' = u\sigma
\]
with \( 0 < \delta < 1 \) a constant we can choose (and adjust for efficient MCMC). Calculate the acceptance probability (answer in L5.R).

7.3. **Matched proposals.** In order to apply this to variable dimension problems we have to generalise it a little more. The proposal generator \( g(u) \) to go from \( \theta \rightarrow \theta' \) may have a different distribution from the proposal \( g'(u') \) generating \( \theta' \rightarrow \theta. \) Let \( \rho \) be the probability we chose the \( g(u)-\)update and let \( \rho' \) be the probability we choose to apply the \( g'(u')-\)update. Detailed balance is
\[
\pi(\theta')\rho'g'(u')\alpha(\theta'|\theta') \left| \frac{\partial(\theta', u')}{\partial(\theta, u)} \right| = \pi(\theta)\rho g(u)\alpha(\theta'|\theta, u)|\theta)
\]
and the acceptance probability is just
\[
\alpha(\theta'|\theta) = \min \left\{ 1, \frac{\pi(\theta')\rho'g'(u')}{\pi(\theta)\rho g(u)} \left| \frac{\partial(\theta', u')}{\partial(\theta, u)} \right| \right\}.
\]
Example: target \( \theta \sim \exp(1) \) as before but this time we take \( u \sim U(1, 2) \) wp 1/2 and otherwise \( u \sim U(0.5, 1) \) so \( \rho = \rho' = 1/2. \) The mapping \( \theta', u' = (u\theta, 1/u) \) is the same as before, so the Jacobian does not change. The algorithm becomes
1) wp 1/2 (a) set \( u \sim U(1, 2) \) and otherwise (b) \( u \sim U(0.5, 1) \). Set \( \theta' = u\theta. \)
2) if we chose (a) then
\[
g'(u')/g(u) = \frac{U(u'; 0.5, 1)}{U(u; 1, 2)} = 2
\]
and we accept \( \theta' \) wp
\[
\alpha(\theta'|\theta) = \min \{ 1, 2 \times e^{-\theta'+\theta-\log(u)} \},
\]
and if we chose (b) then \( g'(u')/g(u) = 0.5 \) and we accept \( \theta' \) wp \( \alpha(\theta'|\theta) = \min \{ 0.5 \times e^{-\theta'+\theta-\log(u)} \}. \)
7.4. Reversible Jump MCMC. Consider now a set of models \( L_m(\theta; y) \pi(\theta|m), \theta \in \Omega_m \), with model prior \( \pi(m), m = 1, ..., M \). We are targeting
\[
\pi(\theta, m|y) \propto p(y|\theta, m) \pi(\theta|m) \pi(m)
\]
The MCMC state is \( X_t = (\theta, m) \). Let \( \rho_{m,m'} \) be a matrix of proposal probabilities: the probability to propose a move to model \( m' \) given the current state is \( m \).
Suppose model \( m' \) has one more parameter than \( m \), for example, \( \theta|m = (\theta_1, ..., \theta_d) \) and \( \theta'|m' = (\theta_1', ..., \theta_d', \theta_{d+1}') \).
Proposal \( d \to d + 1 \): selected with probability \( \rho_{m,m'} \); simulate \( u \sim g(u) \) and set \( \theta'_{1:d+1} = \psi(\theta, u) \) with \( \theta'_{1:d} = \theta_{1:d} \) and \( \theta'_{d+1} = \psi_{d+1}(\theta_{1:d}, u) \) for the \( d+1 \)st component. For simplicity I will assume \( \theta'_{d+1} = \theta'_{d+1}(u) \).
Proposal \( (d + 1 \to d) \): selected with probability \( \rho_{m',m} \); set \( \theta_i = \theta'_i, i = 1, ..., d \) (ie, delete \( \theta'_{d+1} \)). In this move \( u' = \emptyset \) and \( g(u') = 1 \) and \( \psi(\theta'_{1:d+1}, u') = (\theta_{1:d}, u) \).
The Jacobian is
\[
\left| \frac{\partial \psi(\theta, u)}{\partial \psi(\theta', u')} \right| = \left| \frac{\partial \theta'}{\partial (\theta, u)} \right| = \left| \frac{\partial \theta'_{d+1}}{\partial u} \right|.
\]
Notice that the dimensions are matched, \( \theta' = \psi(\theta, u) \) and so \((\theta, u)\) and \( \theta' \) have the same dimension, namely \( d + 1 \).

The acceptance probability for the proposal \( d \to d + 1 \) is
\[
\alpha(\theta'|\theta) = \min \left\{ 1, \frac{p(y|\theta', m') \pi(\theta'|m') \pi(m') \rho_{m',m} g(u') | \partial \theta'| \partial (\theta, u) \} \right\}
\]
where
\[
q(\theta'_{d+1}) = g(u) \left| \frac{\partial \theta'_{d+1}}{\partial u} \right|^{-1}
\]
is just the proposal distribution for \( \theta'_{d+1} \) if the proposals depended on \( \theta \) then this would be \( q(\theta'_{d+1}|\theta) \).
The acceptance probability for the reverse move, \( d + 1 \to d \) from \( \theta = (\theta_1, ..., \theta_{d+1}) \) to \( \theta' = (\theta_1, ..., \theta_d) \) is just the inverse
\[
\alpha(\theta|\theta') = \min \left\{ 1, \frac{p(y|\theta', m) \pi(\theta|m) \pi(m) \rho_{m,m} g(\theta_{d+1})} {p(y|\theta, m) \pi(\theta|m) \pi(m) \rho_{m,m}} \right\}.
\]
The setup described above is a special case, jumping one dimension, with a simple proposal scheme. The framework generalises a great deal.

See for eg the Christian Robert books on Monte Carlo, and “The Bayesian Choice” or the original 1995 Biometrika paper by Peter Green.

7.4.1. RJMCMC - simple example. Let \( X = 1/2 \) with probability \( 1/3 \), otherwise \( X \sim f = 2xI_{0<x<1} \).²
In this example we have a mixture of two models of dimension zero and one: Model \( M = 1 \) has state space \( X|M = 1 \in \{1/2\} \) (ie, a point); Model \( M = 2 \) has state space \( X|M = 2 \in [0,1] \).
In terms of the joint (value,model) pair the target pmf/pdf is
\[
\pi(x,m) = \pi(m) \pi(x|m)
\]
²The CDF is \( F_X(x) = \Pr(X \leq x) \) with \( F_X(x) = \frac{2}{3}x^2 + \frac{1}{2}I_{x\geq 1/2} \)
with $\pi(m = 1) = 1/3$, $\pi(m = 2) = 2/3$, $\pi(x|m = 1) = 1_{x = 1/2}$, $\pi(x|m = 2) = 2x$.

We would like to give a RJ MCMC algorithm targeting this distribution.

**RJ MCMC algorithm targeting** $(X, M) \sim \pi(x, m)$:

MCMC state: $(X_t, M_t) = (x, m)$. First the proposal rules:

- (increase dimension) if $m = 1$ propose $m' = 2$ with probability $\rho_{1, 2} = 1$. Propose $x' \sim \text{Beta}(x'; \alpha = 1/2, \beta = 1/2)$ say;
- (decrease dimension) if $m = 2$ propose $m' = 1$ with probability $\rho_{2, 1} = 1$ and set $x' = 1/2$.

Acceptance probabilities:

If $(x, m) = (1/2, 1)$ and we propose $(x', m' = 2)$ (ie, increase dimension), the acceptance probability is

$$
\alpha(x', m'|x, m) = \min\left\{1, \frac{\pi(x'|m')\pi(m')\rho_{m', m}}{\pi(x|m)\pi(m)\rho_{m, m'}}\right\}.
$$

Substituting in our values the AP is

$$
\alpha(x', m'|x, m) = \min\left\{1, \frac{4x'/3}{\text{Beta}(x'; \alpha, \beta)/3}\right\}.
$$

If $(x, m) = (x, 2)$ and we propose $(x' = 1/2, m' = 1)$ (ie, decrease dimension), the acceptance probability is

$$
\alpha(x', m'|x, m) = \min\left\{1, \frac{\pi(x'|m')\pi(m')\rho_{m', m}}{\pi(x|m)\pi(m)\rho_{m, m'}}\right\}.
$$

Substituting in our values the AP is

$$
\alpha(x', m'|x, m) = \min\left\{1, \frac{\text{Beta}(x; \alpha, \beta)/3}{4x/3}\right\}.
$$

Iteration: we generate our chain $(X_t, M_t)$ iterating proposals and acceptance steps using the formula above. If the model is $m = 1$ (so the state is $x = 1/2$) we propose to jump to model $m' = 2$ and a new state $x' \in [0, 1]$, and vice versa.

Remark 1: the dimension of the proposal dbn matches the change in dimension in the target - in terms of our original notation one we decide to propose a switch from $m = 1$ to $m' = 2$, we simulate $u \sim g(u)$ and set $(x', u') = \psi(x, u)$. In our example $u \sim \text{Beta}(\cdot; \alpha, \beta)$, $u' = \{}$ (the move down a dimension is deterministic) and the “transformation” is just $x' = u$ so the Jacobian is one. In terms of our old notation $\dim(x', u') = \dim(x, u) = 1$ (because $x' = u$ with $u \in [0, 1]$, $u' = \{}$ and $x = 1/2$ belongs to a space of dimension zero also). We often use this sort of simple choice (where $u$ and $x'$ are the same thing).

Remark 2: we could if we wished mix in a fixed-dimension update (ie set $\rho_{2, 1} = \rho_{2, 2} = 1/2$ so we have two options if $m = 2$). In this fixed dimension update we target $\pi(x|m = 2)$ using our standard MCMC tools.

### 7.4.2. Summarising the posterior from RJ-MCMC output.

RJ MCMC is a Monte Carlo method useful for Bayesian model selection and model averaging. If we can sample the joint posterior for model $m$ and parameter $\theta$

$$
\theta^{(t)}, m^{(t)} \sim \pi(\theta, m|y)
$$

we can carry out model averaging and model selection.
Model Choice: Since \( m^{(t)} \sim \pi(m|y) \) (ie, marginally), the maximum a posteriori model (the MAP)

\[
m_{\text{MAP}} = \arg \max_{m=1,\ldots,M} \pi(m|y)
\]

can be estimated by the mode \( \hat{m} \) say, of \( \{m^{(t)}\}_{t=1}^T \). The MAP model is the choice minimising the risk for the 0-1 loss function \( L(m^*, m) = I_{m=m^*} \).

When the number of models is very large, we often summarise the uncertainty over models using an HPD credible set over models. This is just a standard level \( \alpha \) HPD credible set \( C \) for \( m \) using the posterior \( \pi(m|y) \). The HPD set minimises the risk for the loss

\[
L(m^*, C) = \alpha_m \cdot g_C + \text{card}(C),
\]

where \( \alpha \) depends on \( \alpha \) (see The Bayesian Choice Section 5.5.3).

Parameter estimation: Since \( \theta^{(t)} \sim \pi(\theta|y) \), the model averaged posterior expectation \( E_{\theta|y} \Theta (\Theta) = E_{\theta,M|y} (\Theta) \) can be estimated by the mean, \( \hat{\theta} \) say, of \( \{\theta^{(t)}\}_{t=1}^T \).

If we want to do prediction, and for goodness of fit checking, the posterior predictive distribution

\[
p(y'|y) = \sum_m \int L_m(\theta; y') \pi(\theta, m|y) d\theta
\]

(\( L_m \) is the likelihood under model \( m \)) can be simulated or compared with the data. When carrying out model selection, the posterior predictive distribution conditioned on the selected model

\[
p(y'|y, \hat{m}) = \int L_{\hat{m}}(\theta; y') \pi(\theta|y, \hat{m}) d\theta
\]

can be plotted over the data (for a quick visual check, a test needs reserved data).

7.5. RJ MCMC and fitting mixture models. The Galaxy radial velocity data are shown in the figure below. It is natural to model this via a mixture of normals. However we do not know the number of components in the mixture.

Likelihood: Suppose our data \( y_i, i = 1, 2, \ldots, n \) are jointly independent scalars sampled from a mixture model with \( m \) components \( N(\mu^{(m)}_j, \sigma^{(m)}_j^2) \), and mixture weights \( w^{(m)}_j, j = 1, 2, \ldots, m, w_j > 0, \sum_{j=1}^m w_j = 1 \).
The observation model we consider for the iid $y_i, i = 1, 2, \ldots, n$ is the mixture

$$(y_i | \mu^{(m)}, \sigma^{(m)}, w^{(m)}, m) \sim \sum_{j=1}^{m} w_j^{(m)} N(y_i; \mu_j^{(m)}, \sigma_j^{(m)}).$$

The likelihood is therefore

$$L(\mu^{(m)}, \sigma^{(m)}, w^{(m)}, m; y) = \prod_i \left[ \sum_{j=1}^{m} w_j^{(m)} N(y_i; \mu_j^{(m)}, \sigma_j^{(m)}) \right].$$

**Priors:** We take as our priors

$$w^{(m)} \sim \text{Dirichlet}(\alpha 1_m)$$

with $1_m$ a vector of $m$ ones, $\alpha = 1$ ($w^{(m)}$ uniform, sum to one),

$$\mu_j^{(m)} \sim N(20, 10), \quad \text{iid for } j = 1, 2, \ldots, m,$$

which of course covers the data (covers $[0, 40]$ at $2\sigma$ - I assume the scale of the response is known), and

$$\sigma_j^{(m)} \sim \text{Gamma}(1.5, 0.5), \quad \text{iid for } j = 1, 2, \ldots, m,$$

again informed by the scale: mean equals 3; shape 1.5 rules out very dense clusters at small $\sigma$; small rate gives heavy tail, standard deviation about 2.5. For a model prior I take $m \sim \text{Poisson}(\lambda)$ with $\lambda = 10$, which is centred at 10, and tails off above about 20 clusters.

**Posterior:** The posterior for the model and parameters

$$\theta^{(m)} = (\mu^{(m)}, \sigma^{(m)}, w^{(m)})$$
and add its weight to a randomly chosen component \( j \) is then

\[
\pi(\theta^{(m)}, m | y) \propto L(\mu^{(m)}, \sigma^{(m)}, w^{(m)}, m; y) \\
\times \text{Dirichlet}(w^{(m)}; \alpha_{1_m}) \\
\times \prod_{j=1}^{m} N(\mu_j^{(m)}; 20, 10) \text{Gamma}(\sigma_j^{(m)}; 1.5, 0.5) \\
\times \text{Poisson}(m; \lambda)
\]

Remark: I left off the prior on the model index \( m \), which is the number of mixture components (it was in my code); I take a prior \( \text{Poisson}(m; \lambda) \) with \( \lambda = 10 \). I am allowing potentially many clusters; we will see that the data favor a smaller number.

7.5.1. RJ MCMC algorithm fitting a normal mixture with an unknown number of components to the Galaxy Velocity data. Suppose the state is \( X_t = (\mu, \sigma, w, m) \) with \( \mu = (\mu_1, ..., \mu_m) \) etc. To get irreducibility we again need fixed dimension moves (3 of these) and variable dimension moves (2 of these).

Step 1. Choose a move \( move \sim U\{1, 2, ..., 5\} \).

Step 2I. If \( move = 1 \) add a component (increase state dimension by three). Set \( m' = m + 1 \).

Step 2Ia Simulate \( \mu'_{m+1}, \sigma'_{m+1} \sim g(\mu'_{m+1}, \sigma'_{m+1}) \). We will take \( g() \) to be the Normal-Gamma prior above. Set \( \mu' = (\mu, \mu'_{m+1}) \) and \( \sigma' = (\sigma, \sigma'_{m+1}) \).

Step 2Ib Now simulate \( w' \). We have to make sure \( \sum_j w'_j = 1 \) (still). Choose a weight \( j \sim U\{1, 2, ..., m\} \) to “split”. Simulate \( w'_{m+1} \sim U(0, w_j) \) and for \( k = 1, 2, ..., m + 1 \) set

\[
w'_k = \begin{cases} 
  w_k & k = 1, ..., m, k \neq j \\
  w_k - w'_{m+1} & k = j \\
  w'_{m+1} & k = m + 1
\end{cases}
\]

The probability to propose \( m' \) given \( m \) is \( \rho_{m,m'} = 1/5 \). The probability to propose \( (\mu', \sigma', w') \) given \( (\mu, \sigma, w) \) is

\[
q(\mu', \sigma', w'|\mu, \sigma, w) = g(\mu'_{m+1}, \sigma'_{m+1}) \times \frac{1}{m} \times \frac{1}{w_j}.
\]

In the reverse move we will pick a component \( i \) of the mixture at random, delete it and add its weight to a randomly chosen component \( j \) out of the remainder. The probability to propose this reverse move back from \( (\mu', \sigma', w') \) to \( (\mu, \sigma, w) \) is just

\[
p(i, j) = \frac{1}{m(m + 1)},
\]

(given \( m, m' \) already decided) since we must choose the two components involved in the update.

Step 3I. Accept the proposal \( (\mu', \sigma', w', m') \) with probability

\[
\alpha^+ = \alpha(\mu', \sigma', w', m' | \mu, \sigma, w, m)
\]

where

\[
\alpha^+ = \min \left\{ 1, \frac{\pi(\theta^{(m)}', m' | y)p(i, j)}{\pi(\theta^{(m)}, m | y)q(\theta^{(m)}$, $w'|\mu, \sigma, w)} \right\}
\]

Step 2D. If \( move = 2 \) delete a component (decrease state dimension by three). Set \( m' = m - 1 \) (if \( m' = 0 \), reject the move and set \( X_{t+1} = X_t \)).

Step 2Da Simulate \( i \sim U\{1, 2, ..., m\} \). Set \( \mu' = \mu_{-i}, \sigma' = \sigma_{-i} \).
Step 2D. To update $w$ (ensuring $w$ is still normalised), simulate $j \sim \mathcal{U}\{1, 2, ..., m\} \backslash \{i\}$ and then (i) set $w' = w$, (ii) set $w'_j = w_j + w_i$, (iii) set $w' = w'_{-i}$.

The probability to propose $m'$ given $m$ is $\rho_{m,m'} = 1/5$ again. The probability to propose $(\mu', \sigma', w')$ given $(\mu, \sigma, w)$ is just

$$p(i,j) = 1/m(m-1),$$

and to propose the move back, $(\mu', \sigma', w') \rightarrow (\mu, \sigma, w)$, is

$$q(\mu, \sigma, w|\mu', \sigma', w') = g(\mu'_m, \sigma'_m) \times \frac{1}{m-1} \times \frac{1}{w_i + w_j}.$$

Step 3D. Accept the proposal $(\mu', \sigma', w', m')$ with probability

$$\alpha^- = \alpha(\mu', \sigma', w', m'|\mu, \sigma, w, m)$$

where

$$\alpha^- = \min \left\{ 1, \frac{\pi(\mu', \sigma', w', m'|y)q(\mu, \sigma, w|\mu', \sigma', w')}{\pi(\mu, \sigma, w, m|y)p(i,j)} \right\}.$$

We have additionally moves 3-5 which act on $\mu$, $\sigma$ and $w$ respectively in fixed dimension moves.

7.5.2. Summarising the Posterior over models and parameters for the normal mixture. We illustrate the method on the Galaxy velocity distribution data. The R-code and further detail of the algorithm are available on the course website. We ran the code and generated samples $(\mu^{(t)}, \sigma^{(t)}, w^{(t)}, m^{(t)}), t = 1, 2, ..., T$ from the joint posterior distribution over the number of clusters and the cluster weights and parameters.
Figure 3. The plot above shows the posterior distribution over the number of components. 3-6 components is the number favored.

Figure 4. The plot shows traces for the log-prior, log-likelihood and number of components, (as the number of parameters vary, they are not easily plotted).
Figure 5. The bottom figure shows the joint posterior of the mean values and weights in each sampled state (i.e., the points are the pairs $\mu^{(t)}_i, w^{(t)}_i$, $i = 1, 2, ..., m^{(t)}$, $t = 1, 2, ..., T$. Points are colored by the number of clusters $m^{(t)}$ in the state. The top figure shows an estimate of posterior predictive distribution $p(y'|y)$ (black line) obtained by averaging the likelihood $L(\mu, \sigma, w, m, y')$ over the sampled states, at each point $y'$ on the x-axis), and the posterior predictive distribution $p(y'|y, m)$ conditioned on $m$ clusters (red is $m=3$, green is $m=4$, blue is $m=5$). The underlying histogram in black is a histogram of the data, $y$. We expect the distribution of the data to match the posterior predictive distribution, and the fit seems reasonable.
8. Decision Theory in the context of statistical inference

For action $\delta$ and truth $\theta \in \Omega$ our loss function is $L(\theta, \delta)$.

For data $y \sim p(y|\theta)$, $y \in \mathcal{Y}$ and estimator $\hat{\theta} = \delta(y)$, $\delta : \mathcal{Y} \to \Omega$ is an action for each $y \in \mathcal{Y}$, with risk

$$R(\theta, \delta) = \int_{\mathcal{Y}} L(\theta, \delta(y))p(y|\theta)dy.$$ 

If we have a prior $\pi(\theta)$, posterior $\pi(\theta|y)$ and marginal likelihood $m(y)$ the posterior expected loss is

$$\rho(\pi, \delta|y) = \int_{\Omega} L(\theta, \delta(y))\pi(\theta|y)d\theta.$$ 

The prior average risk $\rho(\pi, \delta) = E_{\theta}(R(\theta, \delta))$ is closely related,

$$\rho(\pi, \delta) = \int_{\Omega} \int_{\mathcal{Y}} L(\theta, \delta(y))p(y|\theta)\pi(\theta)d\theta dy,$$

since $\rho(\pi, \delta)$ is equivalently

$$\rho(\pi, \delta) = \int_{\mathcal{Y}} \rho(\pi, \delta(y)|y)m(y)dy.$$ 

$\rho(\pi, \delta)$ determines a total ordering on $\delta$, whilst $R(\theta, \delta)$ may give a different ordering for each $\theta$. The Bayes estimator $\delta^\pi$ for $\theta$,

$$\delta^\pi = \arg \min_\delta \rho(\pi, \delta)$$

is given for every $y \in \mathcal{Y}$ by

$$\delta^\pi(y) = \arg \min_\delta \rho(\pi, \delta|y).$$

8.1. Admissibility. If we accept the loss function $L(\theta, \delta)$ we would never use an estimator $\delta_0$ which was “never better and often worse”. If there exists an estimator $\delta_1$ satisfying

$$R(\theta, \delta_0) \geq R(\theta, \delta_1)$$

and for at least one $\theta_0$,

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1)$$

then we say $\delta_0$ is not admissible. Otherwise it is admissible.

Estimators that seem reasonable (recall James-Stein beats MLE) need not be admissible.

Proposition 2.4.22 (CR-TBC): If prior $\pi$ is strictly positive on $\Omega$ with finite Bayes risk, and the risk, $R(\theta, \delta)$, is a continuous function of $\theta$, then Bayes estimator $\delta^\pi$ is admissible.

Proof: Suppose the opposite. For some $\delta'$, $R(\theta, \delta^\pi) \geq R(\theta, \delta')$ for each $\theta$, and there exists $\theta'$ and an open neighborhood $C'$ of $\theta'$ such that $R(\theta, \delta^\pi) > R(\theta, \delta')$ for $\theta \in C'$. Integrating both sides of the inequality,

$$\int_{\Omega} R(\theta, \delta^\pi)\pi(\theta)d\theta > \int_{\Omega} R(\theta, \delta')\pi(\theta)d\theta,$$

that is

$$r(\pi, \delta^\pi) > r(\pi, \delta').$$

But that is impossible as

$$\delta^\pi = \arg \min_\delta r(\pi, \delta)$$

by definition. [EOP]
8.2. Actions, rewards, utility and loss. : Let \( r(\theta, \delta) \) denote the reward if our action is \( \delta \) and the truth is \( \theta \). We assume bounded rewards, \( r_{\min} \leq r \leq r_{\max} \). Let \( U(r) \) denote the utility of reward \( r \).

Utility is the opposite of loss. In terms of our previous notation,

\[
L(\theta, \delta) = -U(r(\theta, \delta))
\]

We replaced one function \( L \) by two, \( U \) and \( r \). Since \( \theta \sim \pi \) and \( y \sim p(y|\theta) \), \( R = r(\theta, \delta) \) is a random variable. Let \( P(r) \) denote the distribution over rewards determined by our model. Till now we chose the action minimising the posterior expected loss. Now we maximise the expected utility \( E(U(R)) \). Same thing. The utility function (or loss function), must be elicited.

Example: urn with 100 balls, red/black. Choose a color \( \delta \) (or loss function), must be elicited.

Example: as before but two urns. If \( \phi \) is proportion black in urn \( i \), possible rewards are \( r = 0 \) and \( r = 1 \). Suppose \( U(0) = 0 \) and \( U(1) = u \) with \( u > 0 \). Then

\[
P(r = 1) = E_\phi(E(I_{\theta=\text{black}}|\phi))
\]

so \( P(r = 1) = E(\phi) \). The expected utility of choosing black is

\[
E(U(r(\theta, \text{black}))) = P(0)U(0) + P(1)U(1) = uE(\phi)
\]

and

\[
E(U(r(\theta, \text{red}))) = uE(1 - \phi).
\]

Choose black if if \( E(\phi) > 1/2 \) (and indifferent if \( E(\phi) = 1/2 \)).

8.3. Choice of reward distributions. Given a choice between two reward distributions \( P(r) \) and \( P'(r) \), I prefer the one maximising the expected utility

\[
P \geq P' \iff E_P(U(r)) \geq E_{P'}(U(r)).
\]

Represent preference relation using \( \geq, \leq \) and =.

Example: As before but two urns. If \( \phi_i \) is proportion black in urn \( i \) and \( \pi_i(\phi_i) \) prior for \( \phi_i, i = 1, 2 \) then \( P_i(1) = E_{\pi_i}(\phi_i) \) and EU’s are \( uP_i(1) \). Choose reward distribution/urn with larger EU.

We may start with a preference over reward distributions. The utility function we elicit must reproduce that relation. The expected utility hypothesis says a utility function representing our preference over reward distributions exists.

8.4. Jensen’s inequality. : If \( f(x) \) is concave and \( E(X), E(f(X)) < \infty \)

\[
E(f(X)) \leq f(E(X)).
\]

Proof: \( f: \mathbb{R} \rightarrow \mathbb{R} \) is concave if for all \( x, x_0 \in \mathbb{R} \),

\[
f(x) \leq f(x_0) + (x - x_0)f'(x_0)
\]

ie, graph below tangent. Since this holds for all \( x, x_0 \in \mathbb{R} \), take \( x = X \) and \( x_0 = E(X) \) and expectations on both sides.

Remark: If \( f(x) = f(x_0) + (x - x_0)f'(x_0), x \neq x_0 \) (so that \( f(x) \) is strictly concave) and \( X \) is not constant then

\[
E(f(X)) < f(E(X)).
\]

Example: in a choice between an average reward \( r_0 = E(r) \) and a random reward \( r \sim P(r) \), for a concave utility we take the average reward, since \( E(U(r)) < U(E(r)) \) so \( E(U(r)) < U(r_0) \).
9. The Savage axioms

9.1. Does the prior exist? Does the utility function exist? When we elicit a prior for \( \theta \in \Omega \), we assume our knowledge is coherent, so a prior distribution representing our preferences exists.

If \( A, B \subseteq \Omega \) are sets and \( \pi \) a prior, we write \( A \geq B \) if \( \pi(A) \geq \pi(B) \). If we start with a system \( \leq, \geq \) of preferences over sets, then we seek a prior that expresses those preferences:

\[
A \geq B \iff \pi(A) \geq \pi(B)
\]

When we elicit a utility function, we assume the rule we use to decide between distributions over rewards \( P(r(\theta, \delta)) \) is coherent, that is, a utility function exists satisfying

\[
P \geq P' \iff E_P(U(r)) \geq E_{P'}(U(r))
\]

The two questions are linked because the prior \( \pi(\theta) \) contributes to the distribution \( P \) over rewards \( r(\theta, \delta(Y)) \).

The theory applies to any form of inference, Bayesian or otherwise, so we discuss it purely in terms of rewards and utility, and the existence of the distribution \( P(r) \) and function \( U(r) \).

9.2. The Ellsberg paradox. This paradox shows that peoples’ preferences are in some cases inconsistent with any prior.

Two urns, (A) has 50 red and 50 black balls, (B) has 100 balls, each either red or black, with an unknown proportion black. Bet £1000 on the color of draw. How would you bet if offered...

<table>
<thead>
<tr>
<th>Option 1</th>
<th>Option 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>color(_{\text{urn}})</td>
<td>color(_{\text{urn}})</td>
</tr>
<tr>
<td>Bet 1</td>
<td>( r_A )</td>
</tr>
<tr>
<td>Bet 2</td>
<td>( r_B )</td>
</tr>
<tr>
<td>Bet 3</td>
<td>( r_A )</td>
</tr>
<tr>
<td>Bet 4</td>
<td>( b_A )</td>
</tr>
</tbody>
</table>

You are (I expect) indifferent on Bet 1 and 2 since “red” and “black” are exchangeable labels. However many people prefer the objective uncertainty of urn (A) to the subjective uncertainty of urn (B), preferring \( r_A \) over \( b_B \) and \( b_A \) over \( r_B \).

Our preferences for each bet tell us about our unstated prior \( \pi(\phi) \) for the probability \( \phi = \Pr(b_B) \) to draw black from urn B.

Suppose rewards have utilities \( U(-1000) = 0 \) and \( U(1000) = u \), with \( u > 0 \). Indifference on Bet 1 is reasonable as

\[
E(U|\text{choose } r_A) = E(U|\text{choose } b_A) = u/2.
\]

On Bet 2, \( E(U|\text{choose } r_B) = u(1 - E_{\pi}(\phi)) \) and

\[
E(U|\text{choose } b_B) = uE_{\pi}(\phi),
\]

so indifference here implies \( E_{\pi}(\phi) = 1/2 \) in our prior.

On Bets 3 and 4, preferring \( r_A \) over \( b_B \) implies \( E_{\pi}(\phi) < 1/2 \), and preferring \( b_A \) over \( r_B \) implies \( E_{\pi}(\phi) > 1/2 \). This is a contradiction, so for people making the common choices outlined above, a prior does not exist.
9.3. The Allais paradox. This paradox shows that peoples’ preferences are in some cases inconsistent with any utility function. In this example the probabilities are all explicit, so the “prior” is given and exists.

Let \( p = (p_1, p_2, p_3) \) be the probability you win respectively

\[(r_1, r_2, r_3) = (£0, £500,000, £2,500,000). \]

In each of two rounds you have a choice between two lotteries.

1. (A) with \( p^{(A)} = (0, 1, 0) \) OR (B) with \( p^{(B)} = (0.01, 0.89, 0.1) \)

2. (C) with \( p^{(C)} = (0.89, 0.11, 0) \) OR (D) with \( p^{(D)} = (0.9, 0.1, 0.0) \)

People express a clear preference for (A) and (D). Choices (B) and (D) maximise the expected return. What utility function are they using?

Set the utilities to be \( U(r_1) = 0, U(r_2) = u \) and \( U(r_3) = 1 \).

The expected utilities are \( E(U) = (p_1, p_2, p_3)(0, u, 1)^T \), so

\[
E(U|A) = u \quad E(U|B) = 0.1 + 0.89u \\
E(U|C) = 0.11u \quad E(U|D) = 0.1
\]

Preferring (A) to (B) means \( u > 0.1 + 0.89u \Rightarrow u > 10/11 \).

On the other hand preferring (D) to (C) means \( 0.1 > 0.11u \Rightarrow u < 10/11 \),

which is a contradiction. This paradox shows that human decision making does not always maximise an expected utility. This is unsurprising. The difficulty here is that the decision nevertheless seems reasonable.

9.4. Probability distributions - reference notes. You know this! Let \( S \) be a sample space and let \( \mathcal{S} \) be a \( \sigma \)-field of sets in \( S \) satisfying \( S \in \mathcal{S}, A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S} \) and if

\[ A_1, A_2, \ldots \in \mathcal{S} \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathcal{S}. \]

For example if \( S \) is finite then \( \mathcal{S} \) can be taken as the power set. A probability distribution on \((S, \mathcal{S})\) is a map \( P : \mathcal{S} \rightarrow [0, 1] \) satisfying \( P(S) = 1 \), and if \( A_1, A_2, \ldots \in \mathcal{S} \) are disjoint then

\[ P(\cup_i A_i) = \sum_i P(A_i) \quad “\text{countably additive}” \]

The triple \((S, \mathcal{S}, P)\) is called a probability space.


The Savage Axioms are intended to define coherent belief. They are prescriptive - if our preferences do not satisfy the axioms, then they should be altered. We may not accept this, but that is the idea. DeGroot breaks the Axioms down into two sets: one set sufficient for a prior to exist; the other set sufficient for a utility function to exist.
We begin with the Savage axioms for qualitative probability. Inequalities are used to express a system of prior preferences. Write \( A \geq B \) if we mean \( P(A) \geq P(B) \). Relations \( \leq, \geq, <, > \), and \( \sim \) (no preference) are defined similarly.

Given a system of preferences over \( A, B \in S \), does a probability distribution \( P(A) \) satisfying \( A \leq B \Leftrightarrow P(A) \leq P(B) \) exist?

9.6. Axioms of qualitative probability. Here are the first few Savage axioms.

Axiom 1. For any two events \( A \) and \( B \) exactly one of the following relations must hold: \( A > B, A < B, A \sim B \).

Axiom 2. If \( A_1 \cap A_2 = B_1 \cap B_2 = \emptyset \) and \( A_i \geq B_i, i = 1, 2 \) then \( A_1 \cup A_2 \geq B_1 \cup B_2 \).

If in addition either \( A_1 > B_1 \) or \( A_2 > B_2 \) then \( A_1 \cup A_2 > B_1 \cup B_2 \).

Roughly speaking, if \( A \) can happen two exclusive ways, and each is more likely than corresponding events leading to \( B \) then \( A \) is more likely than \( B \).

Axiom 3. If \( A \in S \) then \( \emptyset \leq A \).

It follows from Axioms 1-3 that the order is transitive (if \( A \leq B \) and \( B \leq C \) then \( A \leq C ) \) and \( A \leq B \Rightarrow A^c \geq B^c \).

The preferences in the Ellsberg paradox do not satisfy Axiom 2. If we make the identifications

\[
A_1 = r_A, \quad A_2 = b_A, \quad B_1 = b_B, \quad \text{and} \quad B_2 = r_B,
\]

then the elicited preferences included \( A_1 > B_1 \) and \( A_2 > B_2 \). But \( A_2 = A_1^c \) and \( B_2 = B_1^c \) so if \( A_1 \cup A_2 > B_1 \cup B_2 \) then \( S > S \) which is a contradiction.

Axiom 4 ensures our final \( P(A) \) will be countably additive.

Axiom 4. If \( A_1 \supset A_2 \supset \ldots \) is a decreasing sequence of events and \( B \) is some fixed event satisfying \( A_i \geq B \) for each \( i = 1, 2, \ldots \) then \( \bigcap_{i=1}^{\infty} A_i \geq B \).

The last axiom assigns a probability to any set of outcomes.

Axiom 5. For each \( p \in [0, 1] \) there is \( B_p \in S \) satisfying \( P(B_p) = p \).

Remarks: (1) We don’t need to assume \( B_p \) exists for each \( p \), we can take \( X \sim U[0,1] \) and replace the outcomes \( s \in S \) with \((s,x) \in S \times [0,1] \). This adds events like \( B_p = \{0 \leq X \leq p\} \) to \( S \). A1-A4 must still hold for sets in the new \( \sigma \)-algebra. (2) I simplified A5 slightly to make it clearer - de Groot is more explicit about how the space is extended. (3) We can now determine \( P(A), A \in S \). Given \( A \), we find \( B \) satisfying \( A \sim B \) and \( P(B) = p \) and set \( P(A) = p \). The probability space \((S, S, P)\) defined in this way will satisfy the axioms of probability.
9.7. **Axioms of utility.** Let $\mathcal{P}$ denote a set of distributions $P$ over rewards $r$ ($P$ corresponds a choice of lottery distribution in the Allais paradox). Assume there is a max and min reward $r_{\min}, r_{\max}$ so that

$$P([r_{\min}, r_{\max}]) = \Pr(r_{\min} \leq r \leq r_{\max}) = 1.$$  

The utility function defines a preference relation for $P, P' \in \mathcal{P}$,

$$P \geq P' \iff E_P(U(r)) \geq E_{P'}(U(r)) \quad (*)$$

based on expected utility.

Suppose we start with the preference relation for $P \in \mathcal{P}$ which appears on the LHS of Equation (*), as we did for the Allais paradox, where we chose our preferred lottery. Is there a utility function that satisfies the relations imposed on the RHS?

The first two axioms state that the order relates all lotteries $P, P' \in \mathcal{P}$ and is transitive.

**Axiom 1:** Either $P \geq P'$ or $P \leq P'$.

**Axiom 2:** If $P \geq P'$ and $P' \geq P''$ then $P \geq P''$.

The next axiom says that the preference between two rewards (or distributions over rewards) should not change if we alter both in the same way.

**Axiom 3:** $P' \geq P''$ if and only if

$$\alpha P' + (1 - \alpha)P \geq \alpha P'' + (1 - \alpha)P$$

for all $0 < \alpha < 1$ and every $P \in \mathcal{P}$.

The choices we made in the Allais paradox violate Axiom 3. If

$$P' = (0, 1, 0), \quad P'' = (1/11, 0, 10/11)$$

$$P = (0, 1, 0), \quad \tilde{P} = (1, 0, 0)$$

then

$$p^{(A)} = (0, 1, 0) = 0.11P' + 0.89P$$

$$p^{(B)} = (0.01, 0.89, 0.1) = 0.11P'' + 0.89P$$

$$p^{(C)} = (0.89, 0.11, 0) = 0.11P'' + 0.89\tilde{P}$$

$$p^{(D)} = (0.9, 0, 0.1) = 0.11P'' + 0.89\tilde{P}$$

Since $P$ is the same for options (A) and (B) and we prefer (A) we imply $P' > P''$ by Axiom 3. Similarly, $\tilde{P}$ is the same for (C) and (D) so when we prefer (D) we imply $P' < P''$, which is a contradiction. It is surprising that seemingly reasonable preferences violate a seemingly reasonable axiom.

9.7.1. *[Skip this material at first reading].* I omit Axioms 4 and 5. See CR-TBC/deGroot for detail. Here are some notes outside the scope of the course.

In Axiom 4 a sufficiently small change in $P$ and $P'$ cant reverse a strict preference $P > P'$. 


We then write down a function \( U(r) \) that correctly orders lotteries \( P(r) \) with just one possible outcome \( r \). Since \( E(U|P) = U(r) \) for these lotteries, and A1-4 apply, this is relatively easy. The expectation \( E(U|P) \) is now well-defined for general \( P \).

Axiom 5 says (in effect) that for any lottery \( P(r) \) over rewards \( r_1 \leq r \leq r_2 \) there is an equivalent lottery \( \tilde{P} \sim P \) with just two possible outcomes, so \( \tilde{P}(r_1) + \tilde{P}(r_2) = 1 \).

Using A1-4 it can be shown that the equivalent lottery is

\[
P(r) \sim \frac{E(U|P) - U(r_1)}{U(r_2) - U(r_1)} 1_{r=r_2} + \frac{U(r_2) - E(U|P)}{U(r_2) - U(r_1)} 1_{r=r_1}
\]

and from this it follows that \( P \geq P' \iff E(U|P) \geq E(U|P') \).

9.8. Conclusions: Should we take the Savage axioms as prescriptive?
Ellsberg seems fine. Allais more of a concern, has motivated research on ideas of utility. More fundamentally the SA appear to be a careful answer to a question which should never have arisen, or is at least only relevant in “very-subjective”-Bayes settings.

We build models of the processes which determine parameters (the Radiocarbon and Coalescent examples illustrate this). In this sort of model-based subjective Bayes, existence is not an issue: the prior exists and we want to understand how it orders events and what consequences it has for inference. The same applies in objective Bayes. In this respect the GLM framework presents a particular challenge: there is in general no one-to-one correspondence between elements of the model (ie the parameters) and elements of reality (as is the case in the aforementioned examples). Derived quantities like means do have physical meaning, and so when we elicit priors for GLM parameters we have to fall back on finding priors that give sensible distributions for derived quantities, like the prior predictive distribution of the observed response. This doesn’t uniquely determine the priors for parameters.

More importantly, does the prior describe the “true” “distribution” of the parameter nature generates for us - if so then Bayesian inference has perfect coverage and we are simultaneously doing Bayesian and Frequentist inference.
10. de Finetti’s theorem and infinite exchangeable sequences

This will lead to another existence proof for the principal objects of Bayesian inference, and show that the Bayesian inference procedure is the correct procedure in the setting where we observe (part of) a sequence of variables which is exchangeable, and could in principle be extended to an infinite number of observations.

Consider an infinite sequence of random variables \( \{X_i\}_{i=1}^{\infty} \). The sequence is exchangeable if

\[
(X_1, X_2, \ldots, X_n) \sim (X_{\sigma_1}, X_{\sigma_2}, \ldots, X_{\sigma_n})
\]

for each \( n \geq 1 \) and any permutation \( \sigma \in \mathcal{P}_n \), \( \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \) of the numbers \((1, 2, \ldots, n)\).

Example: an iid sequence of binary random variables \( X_i \sim \text{Bern}(\theta) \) is clearly exchangeable since

\[
\pi(x_1, \ldots, x_n) = \prod_i \theta^{x_i} (1 - \theta)^{1 - x_i}
\]

and if we permute the indices we just shuffle the order of the indices in the product - they all appear once.

10.1. Polya urn: \( b \) black and \( w \) white balls in urn. Sample a ball. Let \( X_i = \mathbb{I}_{i\text{th} \text{ ball is black}} \). Ball returned plus \( A \) balls of same color. Variables \( X_1, X_2, \ldots \) are not independent but they are exchangeable. Consider the probability for sequences 0, 0, 1, 1 and 1, 1, 0

\[
\pi(0011) = \frac{w}{w + b} \frac{w + A}{(w + b + A)} \frac{b}{(w + b + 2A)} \frac{b + A}{(w + b + 3A)}
\]

\[
\pi(1100) = \frac{b}{(w + b)} \frac{b + A}{(w + b + A)} \frac{w}{(w + b + 2A)} \frac{w + A}{(w + b + 3A)}
\]

Permuting arrivals shuffles the numerator factors, they all still appear exactly once. Generalises to sequences of arbitrary length.

If \( \{X_i\}_{i=1}^{\infty} \) are iid they are exchangeable, but not conversely.

10.2. Hierarchical model: For \( n = 1, 2, \ldots \) let \( 0_n \) be a vector of \( n \) zeros, \( \Sigma^{(n)}_{i,i} = 1 \) and \( \Sigma^{(n)}_{i,j} = \rho \) with \( 1 > \rho \geq 0 \). The distribution

\[
\pi(x_1, \ldots, x_n) = N(x; 0_n, \Sigma^{(n)})
\]

gives an exchangeable sequence. For \( \sigma \in \mathcal{P}_n \) a permutation, we have \( E(X_{\sigma_i}) = \mu \) and \( \text{cov}(X_{\sigma_i}, X_{\sigma_j}) = \rho \), so that

\[
\pi(x_{\sigma_1}, \ldots, x_{\sigma_n}) = N(x_{\sigma}; 0_n, \Sigma^{(n)}) = N(x; 0_n, \Sigma^{(n)}).
\]

Here \( 0 \leq \rho < 1 \) is N&S for \( \Sigma^{(n)} \) positive definite for all \( n \geq 1 \).

The eigenvalues\(^3\) are

\[
\lambda_1 = 1 + (n - 1)\rho, \quad \text{and} \quad \lambda_2 = \ldots = \lambda_n = 1 - \rho
\]

\(^3\)eigenvectors \((1, 1, \ldots, 1)\) and anything orthogonal for eg \((1, -1, 0, 0, \ldots, 0), (0, 1, -1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1, -1)\)
so \(-1/(n-1) < \rho < 1\) is N&S for PD \(\Sigma^{(n)}\) at fixed \(n\), and \(0 < \rho < 1\) is N&S to make this work for all \(n \geq 1\).

10.3. **Theorem (de Finetti):** Let \(X_1, X_2, ... X_n, ...\) be an infinite sequence of binary random variables. The sequence is exchangeable if and only if there exists a distribution function \(F(\theta)\) on \([0, 1]\) such that

\[
p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^n p(x_i|\theta) \, dF(\theta) \tag{*}
\]

with

\[
F(\theta) = \Pr(\Theta \leq \theta) \quad \text{where} \quad \Theta = \lim_{N \to \infty} N^{-1} \sum_{i=1}^N X_i,
\]

and \(p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}\). It further holds that the conditioned distribution is Bernoulli,

\[
p(x_1, ..., x_n|\Theta = \theta) = \prod_{i=1}^n p(x_i|\theta).
\]

10.4. **Remarks.** In short, “an infinite exchangeable sequence is distributed as a mixture of iid random variables”. The theorem says \(F\) and \(\theta\) must exist to make this hold.

The theorem extends to cover infinite exchangeable sequences of random vectors (ie \(X_i\) continuous multivariate random variables) with a multivariate parameter \(\theta\). The expression \(dF(\theta)\) may be off-putting. If \(F(\theta)\) is the cdf of a pdf \(\pi\) then \(dF(\theta) = \pi(\theta) d\theta\), and

\[
p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^n p(x_i|\theta) \pi(\theta) d\theta.
\]

In general \(F\) is just some unknown distribution which puts probability mass on sets \(d\theta\) in parameter space \(\Omega\).

It isn’t always obvious what \(F\) and \(\theta\) are. Here is an example.

**Proof (of de Finetti):** we begin by looking at \(n\) of the first \(N\) outcomes in the sequence. We write down the joint pmf of \(X_1, ..., X_n\) and show that it converges to the RHS of Equation\((*)\) as \(N \to \infty\). Let

\[
S_n = X_1 + X_2 + ... + X_n, \quad n = 1, 2, ...
\]

and let \(r, s\) be two integers \(0 \leq r \leq s \leq N\). By exchangeability, the conditional distribution of \(S_n\) given \(S_N\) is hypergeometric,

\[
\Pr(S_n = r|S_N = s) = \binom{s}{r} \binom{N-s}{n-r} \binom{N}{n}
\]

since this is the probability to draw \(r\) 1’s in \(n\) draws without replacement from an urn containing \(s\) 1’s and \(N-s\) 0’s.

When \(S_n = r\) the urn contains at least \(n-r\) 0’s so

\[
\Pr(S_n = r) = \sum_{s=r}^{N-(n-r)} \Pr(S_n = r|S_N = s) \Pr(S_N = s)
\]

\[
= \sum_{s=r}^{N-(n-r)} \Pr(S_n = r|S_N/N = \theta(s)) \Pr(S_N/N = \theta(s))
\]
where \( \theta(s) \equiv s/N \).

The random variable \( \Theta_N \sim SN/N \) has CDF a \( F_N(\theta) \) which is

\[
F_N(\theta) = \Pr(\Theta_N \leq \theta) = \sum_{s=0}^{N} \Pr(S_N = N\theta(s))\mathbb{I}_{\theta(s) \leq \theta}.
\]

As \( \theta \) increases past \( \theta(s) \), \( F_N(\theta) \) jumps up by \( \Pr(S_N = N\theta(s)) \). It is not differentiable at discontinuities, but we can write down a density

\[
f_N(\theta) = \sum_{s=0}^{N} \Pr(S_N = N\theta)\delta(\theta - \theta(s))
\]

which assigns the correct probability to sets when integrated. In this expression

\[
\delta_{\theta(s)}(\theta) = \delta(\theta - \theta(s))
\]

is a Dirac delta function which places a unit point mass at \( \theta = \theta(s) \). This function is defined by its action in integrals: if \( g(\theta) \) is continuous at \( \theta(s) \) then

\[
\int_{0}^{1} g(\theta)\delta(\theta - \theta(s))d\theta = g(\theta(s)).
\]

Returning to our density we calculate

\[
\Pr(a \leq \Theta_N \leq b) = \int_{a}^{b} f_N(\theta)d\theta = \int_{a}^{b} \left[ \sum_{s=0}^{N} \Pr(S_N = N\theta)\delta(\theta - \theta(s)) \right]d\theta
\]

\[
= \sum_{s=0}^{N} \Pr(S_N = N\theta(s))\mathbb{I}_{a \leq \theta(s) \leq b}
\]

\[
= F_N(b) - F_N(a).
\]

Point masses \( \delta_{\theta(s)} \) in \( f_N \) are associated with the discontinuities at \( \theta(s) \) in \( F_N \). The derivative

\[
dF(\theta) = f(\theta)d\theta,
\]

is defined in terms of its action under an integral.

Exercise: show that \( \int_{0}^{1} f_N(\theta)d\theta = 1 \).

Now write \( \Pr(S_n = r) \) in terms of \( F_N \). Let

\[
g_N(\theta) = \mathbb{I}_{r \leq N\theta \leq (n-r)} \Pr(S_n = r|S_N = N\theta).
\]
In terms of $F_N, f_N$ and $g_N$,

$$\int_0^1 g_N(\theta) dF_N(\theta) = \int_0^1 g_N(\theta) f_N(\theta) d\theta$$

$$= \sum_{s=0}^N \int_0^1 g_N(\theta) \Pr(S_N = N\theta) \delta(\theta - \theta(s)) d\theta$$

$$= \sum_{s=0}^N g_N(\theta(s)) \Pr(S_N = N\theta(s))$$

$$= \sum_{s=r}^{N-(n-r)} [\Pr(S_n = r|S_N = N\theta(s))] \times \Pr(S_N = N\theta(s))]$$

$$= \Pr(S_n = r)$$

... so we can write

$$\Pr(S_n = r) = \int_0^1 g_N(\theta) dF_N(\theta)$$

$$= \int_0^1 \mathbb{1}_{r \leq N\theta \leq N-(n-r)} \Pr(S_n = r|S_N = N\theta) dF_N(\theta)$$

$$= \int_{r/N}^{1-(n-r)/N} \Pr(S_n = r|S_N = N\theta) dF_N(\theta)$$

We now assume two limiting results. The hypergeometric $\Pr(S_n = r|S_N = N\theta)$ converges uniformly to the binomial,

$$\frac{\binom{N\theta}{r} \binom{N(1-\theta)}{n-r}}{\binom{N}{n}} \to \binom{n}{r} \theta^r (1-\theta)^{n-r}$$

as $N \to \infty$ at fixed $\theta$. I leave this to you to show. It makes sense because we have $N\theta$ 1’s and $N(1-\theta)$ 0’s, and there is little difference between sampling with and without replacement when we sample a small number from a large population.

Helly’s Theorem (Feller (1966) Probability Theory and Applications II) “an infinite sequence of probability distributions $F_N$ on a finite interval contains a convergent subsequence”.

It follows that a limit $F_N(\theta) \to F(\theta)$ exists, that is

$$\lim_{N \to \infty} \Pr(N^{-1}S_N \leq \theta) = \Pr(\Theta \leq \theta)$$

where $\Theta = \lim_{N \to \infty} N^{-1}S_N$. Assuming these two convergence results hold, (and, as Jason pointed out, using Helly’s 2nd theorem, which says that if $F_N \to F$ then integrals $dF_N$ converge to integrals $dF$ - under certain conditions which hold here - see Feller again) then for the infinite exchangeable sequence,

$$\Pr(S_n = r) = \binom{n}{r} \int_0^1 \theta^r (1-\theta)^{n-r} dF(\theta) \quad (A)$$

Finally then,

$$\Pr(S_n = r) = \binom{n}{r} p(x_1, ..., x_n) \quad (B)$$
for any \( x_1, \ldots, x_n \) summing to \( r \), so subbing (B) in (A),
\[
p(x_1, \ldots, x_n) = \int_0^1 \theta^r (1 - \theta)^{n-r} dF(\theta) \quad (C)
\]
which is the result of de Finetti.

The representation in terms of \( F \) is unique. A distribution on a bounded interval is uniquely determined by its moments, and since (C) fixes all the moments \( E(\theta^n), n = 1, 2, \ldots \) of \( F \) (take \( r = n \) and \( n = 1, 2, \ldots \)) it follows that \( F \) exists and is unique.

[\text{EOP}]

10.5. **Hierarchical model:** As an example of this theorem in action, we return to the Hierarchical normal model and set it up in de Finetti form. If \( \Sigma^{(n)}_{ii} = 1 \) and \( \Sigma^{(n)}_{ij} = \rho \) with \( 0 \leq \rho < 1 \) then
\[
\pi(x_1, \ldots, x_n) = N(x; 0_n, \Sigma^{(n)})
\]
gives an exchangeable sequence. Write it in de Finetti form.

Simulate \( \theta \sim N(0, v) \) and set
\[
X_i = \theta + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, w).
\]

Marginally \( E(X_i) = 0, \var(X_i) = v + w \) and \( \text{cov}(X_i, X_j) = v \) so \( w = 1 - \rho \) and \( v = \rho \) gives \( X \sim N(0_n, \Sigma^{(n)}) \). Conditionally,
\[
\pi(x_1, \ldots, x_n | \theta) = \prod_i N(x_i; \theta, 1 - \rho)
\]
and unconditionally
\[
\pi(x_1, \ldots, x_n) = \int \prod_i N(x_i; \theta, 1 - \rho) N(\theta; 0, \rho) d\theta,
\]
the de Finetti form, and equal to \( N(x; 0_n, \Sigma^{(n)}) \).

10.6. **Bayesian inference.** If the data are \( x_1, \ldots, x_n \) then
\[
p(x_1, \ldots, x_n | \theta) = \prod_i \theta^{x_i} (1 - \theta)^{1-x_i}
\]
is the likelihood and \( F(\theta) \) is the CDF of the prior. de Finetti gives the form for the prior predictive distribution of the data
\[
p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n | \theta) dF(\theta).
\]

If the observations are an infinite exchangeable sequence, then all these objects exist.

Suppose we have seen \( x_1, \ldots, x_m \) and we wish to predict \( x_{m+1}, \ldots, x_n \). The predictive distribution is
\[
p(x_{m+1:n} | x_{1:m}) = \frac{p(x_{1:n})}{p(x_{1:m})}
\]
\[
= \int p(x_{m+1:n} | \theta) \frac{p(x_{1:m} | \theta) dF(\theta)}{p(x_{1:m})}.
\]

We see that
\[
dF(\theta | x_1, \ldots, x_m) \propto p(x_1, \ldots, x_n | \theta) dF(\theta)
\]
is our updated prior, in other words this is the posterior for $\theta$ given data $x_1, \ldots, x_m$. de Finetti tells us that in this exchangeable setting we should be doing Bayesian inference... if we can.

10.7. De Finetti and the Polya urn. As a second example of de Finetti in action, it may be shown that, for a Polya urn,

$$p(x_1, \ldots, x_n) = \int \prod_i \theta^{x_i} (1 - \theta)^{1-x_i} \text{Beta}(\theta; b/A, w/A) d\theta.$$ 

ie, without using de Finetti. 

Exercise: suppose $\sum_{i=1}^n x_i = k$. Show for the polya urn that

$$p(x_1, \ldots, x_n) = \frac{\prod_{i=1}^{k-1} (b + iA) \prod_{j=1}^{n-k-1} (w + jA)}{\prod_{i=1}^{n-1} b + w + iA}.$$ 

and write this in terms of $\Gamma$-functions using the identity $x\Gamma(x) = \Gamma(x + 1)$. Do the integral above and hence show directly the LHS and RHS are equal.
11. Bayesian non-parametrics - the Dirichlet Process and normal mixture

11.1. Motivation. We saw in the last lecture that for an infinite exchangeable sequence \( y_1, y_2, \ldots \) the prior predictive distribution \( p(y_1, \ldots, y_n) \) is given in terms of a likelihood \( p(y_1, \ldots, y_n | \theta) \) with a parameter \( \theta \) and a prior distribution \( G \) such that

\[
p(y_1, \ldots, y_n) = \int_\Omega p(y_1, \ldots, y_n | \theta) dG(\theta)
\]

These distributions and the posterior distribution

\[
dG(\theta | x_1, \ldots, x_m) \propto p(x_1, \ldots, x_n | \theta) dG(\theta)
\]

all exist. The difficulty is that we don’t know \( G \) (like we didn’t know \( \theta \)). The idea is to infer \( G \), treating it like a parameter. It is some unknown distribution that puts probability mass on sets \( A \subset \Omega \). If we specify \( G(A) \) for all sets \( A \) in the \( \sigma \)-algebra of \( \Omega \) then we specify \( G \).

In Bayesian inference the things you don’t know, the parameters, are random variables, with the variation representing the uncertainty in the parameter value. Here \( G \) will be a random measure over \( \theta \)-values. We will have a “prior” for \( G \)

\[
(\text{informally}) \quad \pi(G) dG
\]

and all the usual objects like \( E[G(A)] \), \( \text{var}(G(A)) \) and a posterior \( G(A | x_1, \ldots, x_n) \) defined for all sets \( A \).

11.2. Bayesian Non-parametrics. - with enough data you reject any parametric model

Non parametric models allow fitting with an unbounded number of parameters. Small model violations (skew, correlation, heavy tails, number of mixture components) built-in to parametric models may become glaring as the data set grows large.

NP models adapt themselves to data as more data is added - they are able to model data with much greater complexity.

Often, the “scientific model” is parametric, but the noise has some unknown complex structure - so express the science with a parametric model and model the noise non-parametrically. NP models commonly have parametric elements.

Prior elicitation and careful modeling of the observation process still matter. The NP model you specify will have built in structure, it won’t adapt to every possible form of model mispecification. Also, the model contains information which may bias the inference. This is good if the model biases towards the truth!

We will just look at one example of a NP model and BNP fitting: the Dirichelet process mixture for density estimation and clustering. The model changes but the methodological framework (model selection, averaging, marginal likelihoods...) is the same.

11.3. The Dirichlet Distribution. - reference notes.

Let \( w = (w_1, \ldots, w_K) \) with \( \sum_k w_k = 1 \).

Suppose \( w \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \). The density of \( w \) is

\[
\pi(w_1, w_2, \ldots, w_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} w_1^{\alpha_1-1} w_2^{\alpha_2-1} \cdots w_K^{\alpha_K-1}.
\]
Agglomeration: if \( w_1, \ldots, w_K \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \) then
\[
  w_1 + w_2, w_3, \ldots, w_K \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K).
\]

Conjugate prior to multinomial: if \( (n_1, \ldots, n_K) \sim \text{Multinom}(n, w) \) with \( n = \sum_k n_k \) and \( w \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \) then
\[
  \pi(w|n_1, \ldots, n_K) \propto w^\alpha_1 + n_1 - 1 \cdot w^\alpha_2 + n_2 - 1 \ldots w^\alpha_K + n_K - 1
\]
so \( w|n_1, \ldots, n_K \sim \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_K + n_K) \).

11.4. The finite Dirichlet process. We want to distribute probability randomly in \( \Omega \). Let \( H(\theta) \) be some simple parameteric distribution on \( \Omega \), for example
\[
  H(\theta) = \pi(\theta) d\theta.
\]
Fix the number of points \( K \geq 1 \) in the support of \( G \), and let \( \alpha = (\alpha_1, \ldots, \alpha_K) \) be given.
We can make a random discrete measure as follows.
1. For \( k = 1, \ldots, K \), sample \( \theta^*_k \sim H \).
2. Sample \( w_1, \ldots, w_K \sim \text{Dirichlet}(\alpha/K) \)
We put probability masses \( w_k \) at a random locations \( \theta^*_k \) in \( \Omega \).
Now given \( G \) (ie, given \( w, \theta^* \)) we have
\[
  \Pr(\theta = \theta^*_k|G) = w_k,
\]
or in standard notation, if \( \theta \sim G \) then \( \theta = \theta^*_k \) wp \( w_k \). If \( A \subseteq \Omega \) then \( \Pr(\theta \in A|G) \) is the chance to draw one of the \( \theta^* \)'s in \( A \),
\[
  \Pr(\theta \in A|G) = \sum_{k=1}^K w_k 1_{\theta \in A}.
\]
We can write this random discrete distribution using our delta function notation
\[
  dG(\theta) = \sum_{k=1}^K w_k \delta(\theta - \theta^*_k) d\theta
\]
For \( A \subseteq \Omega \),
\[
  G(A) = \int_{\Omega} 1_{\theta \in A} dG(\theta)
  = \int_{\Omega \cap A} \sum_{k=1}^K w_k \delta(\theta - \theta^*_k) d\theta
  = \sum_{k=1}^K w_k 1_{\theta \in A}.
\]
We sometimes use the notation
\[
  G = \sum_{k=1}^K w_k \delta_{\theta^*_k},
\]
for writing distributions of this sort since \( \delta_{\theta^*_k}(A) = 1_{\theta \in A} \) but \( \delta_{\theta^*_k} \) refers to the distribution.
We have a simple procedure that assigns a random probability mass to sets \( A \subseteq \Omega \).
By putting a prior on $w$ and $\theta^*$ we determine a prior on a probability distribution. The choice $w \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$, $\theta^*_k \sim H$, $k = 1, \ldots, K$ determines a finite Dirichlet process. $G$ is a random distribution, but it is “centred” on the “base distribution” $H$. Since $w$ and $\theta^*$ are independent

$$E(G(A)) = \sum_{k=1}^{K} E(w_k) E(\mathbb{1}_{\theta^*_k \in A})$$

$$= \sum_{k} \frac{\alpha_k / K}{\sum_j \alpha_j / K} H(A)$$

$$= H(A)$$

Exercise: look up the variance of a Dirichlet distribution and use it to calculate $\text{var}(G(A))$.

11.5. The Dirichlet process. The Dirichlet process is obtained from the finite Dirichlet process in the limit $K \to \infty$ (Q4 on PS4). The definition we give below “starts anew” but the two approaches are equivalent.

Definition: $G \sim DP(\alpha, H)$ if $\forall$ partitions $A_1, A_2, \ldots, A_r$ of $\Omega$

$$G(A_1), \ldots, G(A_r) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_r))$$

A number of properties follow immediately from the properties of a Dirichlet Distribution. We must have

$$G(A), G(A^c) \sim \text{Beta}(\alpha H(A), \alpha(1 - H(A)))$$

by agglomeration so $E(G(A)) = H(A)$ as before.

If $G \sim DP(\alpha, H)$ and $\theta \sim G$ then marginally $\theta \sim H$, because

$$\Pr(\theta \in A) = E(E(\mathbb{1}_{\theta \in A} | G)) = E(G(A))$$

so $\Pr(\theta \in A) = H(A)$ integrating $dG$.

11.6. Stick breaking representation. - not proven

The equivalence to the distribution obtained in the limit may be shown using the “stick breaking” construction of the DP.

If $\beta_k \sim \text{Beta}(1, \alpha)$ and for each $k = 1, 2, \ldots, K, \ldots$ we set

$$w_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \quad \theta^*_k \sim H \quad \text{and} \quad G(A) = \sum_{k=1}^{\infty} w_k \mathbb{1}_{\theta^*_k \in A}$$

then we move from the finite Dirichlet process to the Dirichlet process as the number of terms in the sum defining $G$ goes from $K$ to $\infty$. 

11.6.1. **Updating the DP.** Recall that $G$ is the “parameter”, the thing we don’t know and are trying to infer. Suppose we observe a sequence of $\theta$-values drawn from $G$. The process determining the $\theta$’s is

1. $G \sim DP(\alpha, H)$
2. $\theta_i \sim G, i = 1, 2, ..., n$

Here $DP(\alpha, H)$ is a prior for $G$ and $\theta_1, ..., \theta_n$ are data.

**Claim:** The posterior distribution of $G|\theta_1$ is

$$G|\theta_1 \sim DP(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}).$$

**Proof:** let $A_1, ..., A_r$ be a partition and suppose $\theta_1 \in A_j$. Then

$$\pi(G(A_1), ..., G(A_r)|\theta_1) \propto p(\theta_1|G(A_1), ..., G(A_r)) \pi(G(A_1), ..., G(A_r)).$$

Now $\pi(G(A_1), ..., G(A_r))$ is Dirichlet($\alpha H(A_1), ... \alpha H(A_r)$). Since $A_j, j = 1, ..., r$ is a partition, there is a unique $j$ such that $\theta_1 \in A_j$, so

$$p(\theta_1|G(A_1), ..., G(A_r)) \propto p(\theta_1|\theta_1 \in A_j)p(\theta_1 \in A_j|G(A_1), ..., G(A_r))$$

$$= h(\theta_1|\theta_1 \in A_j)G(A_j)$$
if \( H(d\theta) = h(\theta)d\theta \). Now \( h(\theta_1|\theta_1 \in A_j) = h(\theta_1)/H(A_j) \) does not depend on \( G \) so if we don’t know which of the sets \( A_j \) that \( \theta_1 \) is in we cover all the possibilities, writing

\[
\pi(G(A_1), ..., G(A_r)|\theta_1) \propto G(A_1)^{I_{\theta_1 \in A_1}, ..., G(A_r)^{I_{\theta_1 \in A_1}} \propto G(A_1)^{\alpha H(A_1)-1}...G(A_r)^{\alpha H(A_r)-1}
\]

or in other words

\[
G(A_1), ..., G(A_r) \sim \text{Dirichlet}(\alpha H(A_1) + I_{\theta_1 \in A_1}, ..., \alpha H(A_r)I_{\theta_1 \in A_r}).
\]

This is a \( DP(H', \alpha') \) if we choose the right \( \alpha' \) and \( H' \). If we take as our new base distribution

\[
H' = \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}
\]

and our new \( \alpha' = \alpha + 1 \) then for \( j = 1, ..., r, \)

\[
\alpha' H'(A_j) = \alpha H(A_j) + \delta_{\theta_1 \in A_j}
\]

and so

\[
G|\theta_1 \sim DP(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}) \quad \text{[EOP]}
\]

In terms of \( \delta \)-functions the new density is

\[
h'(\theta) = \frac{\alpha}{\alpha + 1} h(\theta) + \frac{1}{\alpha + 1} \delta(\theta - \theta_1)
\]

so that \( H'(A) = \int_A h'(\theta)d\theta \) gives

\[
H'(A) = \frac{\alpha}{\alpha + 1} H(A) + \frac{1}{\alpha + 1} \sum_{\theta_1 \in A} \delta_{\theta_1}
\]

as above. This is a mixture: to simulate \( \theta \sim G(\theta_1) \) we would simulate \( \theta \sim h \) wp \( \alpha/(\alpha + 1) \) and otherwise (ie wp \( 1/(\alpha + 1) \)) we set \( \theta = \theta_1 \). Notice the atom at \( \theta = \theta_1 \).

Exercise (Q3 PS4): show that if \( \theta_1, ..., \theta_n \sim G \) with \( G \sim DP \) then

\[
G|\theta_1:n \sim DP(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}).
\]

[Hint: induction replacing \( G \) by \( G|\theta_1:(n-1) \)]

The predictive distribution of \( \theta_{n+1} | \theta_1:n, \alpha, H \) is

\[
(\theta_{n+1} | \theta_1:n, \alpha, H) \sim \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}
\]

because if \( G \sim DP(\alpha, H) \) and \( \theta \sim G \) then marginally \( \theta \sim H \).

Again, this is a mixture of the continuous distribution \( H \) and the atoms at \( \theta_i, i = 1, ..., n \). To simulate this distribution we simulate \( \theta \sim H \) wp \( \alpha/(\alpha + n) \) and otherwise (ie wp \( 1/(\alpha + n) \)) we sample \( \theta \sim U\{\theta_1, ..., \theta_n\} \).

11.7. The Chinese Restaurant process. We don’t expect \( \theta_1, ..., \theta_n \) to contain \( n \) distinct \( \theta \)-values. Recall that the Dirichlet process puts discrete probability masses in \( \Omega \), so when we sample it we may get the same parameter values many times. We can rewrite the sum

\[
\sum_{i=1}^{n} \delta_{\theta_i} = \sum_{k=1}^{K} n_k \delta_{\theta_k^*}
\]

where \( \theta^* = (\theta_1^*, ..., \theta_K^*) \) are the unique \( \theta \)-values in \( (\theta_1, ..., \theta_n) \), \( K \) is random, and \( n_k \) is the number of times \( \theta_k^* \) appears in \( (\theta_1, ..., \theta_n) \). We have two representations of the simulation outcome, \( (\theta_1, ..., \theta_n) \) and equivalently \( (n_k, \theta_k^*), k = 1, ..., K \).
11.7.1. **Sequential simulation.** If for \( i = 1, \ldots, n, \theta_i \sim G \) then the \( \theta \)'s are exchangeable. However (all conditioned on \( \alpha, H \))

\[
p(\theta_1, \ldots, \theta_n) = p(\theta_n | \theta_{1:n-1}) p(\theta_{n-1} | \theta_{1:n-2}) \ldots p(\theta_1)
\]

so we can sample sequentially using predictive distributions,

\[
\theta_1 \sim p(\theta_1), \quad \theta_1 | \theta_2 \sim p(\theta_2 | \theta_1) \quad \ldots \quad \theta_n \sim p(\theta_n | \theta_{1:n-1}).
\]

when we do this we just simulate

\[
\theta_{j+1} \sim \frac{\alpha H + \sum_{k=1}^{K_j} n_k^j \delta_{\theta_k^j}}{\alpha + j}, \quad j = 1, 2, \ldots, n - 1.
\]

\( K_j \) is the number of distinct \( \theta \)'s “so far”, \( \theta_k^j \) are the distinct values, and \( n_k^j \) is the number of times \( \theta_k^j \) appears in \( (\theta_1, \ldots, \theta_j) \).

11.7.2. **Partition notation.** We have an iterative simulation scheme.

\[
\theta_{j+1} \sim \frac{\alpha H + \sum_{k=1}^{K_j} n_k^j \delta_{\theta_k^j}}{\alpha + j}, \quad j = 1, 2, \ldots, n - 1.
\]

The \( j+1 \)st \( \theta \)-value is a “new” \( \theta \)-value, \( \theta_{j+1} \sim H \), with probability \( \alpha/(\alpha + j) \) and otherwise it is one of the old ones. The probability we choose \( \theta_{j+1} = \theta_k^j \) is \( n_k^j/(\alpha + j) \).

The simulated \( \theta \)-values are partitioned. If \( S = (S_1, \ldots, S_K) \) is a partition of \( 1, 2, \ldots, n \), then \( S_k \) contains the indices of all the \( \theta \)'s we set equal \( \theta_k^j \).

If we know the partition \( S = (S_1, \ldots, S_K) \) and we know the corresponding \( \theta^* = (\theta_1^*, \ldots, \theta_K^*) \) values then we know \( (\theta_1, \ldots, \theta_n) \) so this is a third representation of the state.

11.7.3. **The Chinese Restaurant Process itself.** The sequential simulation of parameters according to a \( DP(\alpha, H) \) is analogous to restaurant seating!
1) There is \( j = 1 \) one customer in the restaurant. They are seated at table \( k = 1 \).
2) Suppose that after the customer \( j = 2, 3, \ldots, n \) arrives, there are \( n_k^j \) people seated at table \( k \), and \( K_j \) tables in all are occupied.
3) The \( j+1 \)'st arrival chooses a new table wp \( \alpha/(\alpha + j) \) and table \( k \) with probability proportional to \( n_k^j \).

After the \( n \)'th customer arrives we have a partition of customers across \( K = K_n \) tables, with \( S_k \) a set listing the customers at the \( k \)th table. If we associate an independent draw \( \theta_k^j \sim H \) with each table \( k = 1, 2, \ldots K \) then \( (S, \theta^*) \sim G \) with \( G \sim DP(\alpha, H) \).

11.7.4. **The CRP distribution over partitions.** Let \( S = (S_1, \ldots, S_K) \) be a partition of \( \{1, 2, \ldots, n\} \), and let \( n_k = card(S_k) \). The probability we get partition \( S \) after \( n \) arrivals is

\[
P(S) = \frac{\Gamma(\alpha) \alpha^K \prod_{k=1}^{K} \Gamma(n_k)}{\Gamma(\alpha + n)}.
\]

Proof: Suppose the sequence of table assignments is

\[
T = (1, 1, 2, 1, 2, 3, 3, 2, 2, 4)
\]
for \( n = 10 \) customers so \( S = \{\{1, 2, 4\}, \{3, 5, 8, 9\}, \{6, 7\}, \{10\}\}. \) Now \( P(S) = P(T) \)
if \( T = T(S) \) as \( T \) and \( S \) are 1 to 1, so
\[
P(S) = \frac{1}{\alpha + 1} \times \frac{\alpha}{\alpha + 2} \times \frac{2}{\alpha + 3} \times \frac{1}{\alpha + 4} \times \frac{\alpha}{\alpha + 5} \\
\times \frac{1}{\alpha + 6} \times \frac{2}{\alpha + 7} \times \frac{3}{\alpha + 8} \times \frac{\alpha}{\alpha + 9}
\]
\[
= \alpha^{K-1} \prod_{i=2}^{10} (\alpha + i - 1)^{-1}.
\]
In the general case, the \( i \)th arrival brings a denominator factor \( (\alpha + i - 1)^{-1} \), so the denominator is \( \prod_{i=2}^{n} (\alpha + i - 1) \).
Now look at the numerator. Suppose the customers seated at table \( k \) are \( S_k = \{i_1, i_2, ..., i_{n_k}\} \). When \( i_1 \) arrived there was no-one sitting at table \( k \), and \( k - 1 \) tables were occupied, so \( i_1 \) chooses table \( k \) wp \( \alpha/((\alpha + i_1 - 1) \)
After that, for \( j = 2, ..., n_k \), there were \( j - 1 \) seated at table \( k \) when \( i_j \) arrived, so \( i_j \) chose table \( k \) wp \( (j - 1)/(\alpha + i_j - 1) \) so the numerator factor is \( \alpha(n_k - 1)! \).
If we end up with \( K \) tables then there are \( K - 1 \) events in which a new table is chosen so
\[
P(S) = \frac{\alpha^{K-1} \prod_{k=1}^{K} (n_k - 1)!}{\prod_{i=2}^{n} (\alpha + i - 1)}
\]
\[
= \frac{\alpha^{K} \prod_{k=1}^{K} (n_k - 1)!}{\prod_{i=1}^{n} (\alpha + i - 1)}
\]
For \( n \) integer \( \Gamma(n) = (n - 1)! \) and for \( x > 0, x\Gamma(x) = \Gamma(x + 1) \) so we can write this in terms of \( \Gamma \) functions,
\[
P(S) = \frac{\Gamma(\alpha)\alpha^{K} \prod_{k=1}^{K} \Gamma(n_k)}{\Gamma(\alpha + n)}
\]
\[\text{[EOP]}\]

11.8. **Summary of Dirichlet process and CRP calculations to this point.**
So where does that get us? We can now write the joint marginal distribution of \( \theta_1, ..., \theta_n \) (ie given \( \alpha \) and \( H \) and integrating over \( G \)) in terms of \( \theta^* \) and \( S \).
If \( G \sim DP(\alpha, H) \) and \( \theta_1, ..., \theta_n \sim G \) then in terms of the equivalent representation \( (\theta^*, S) \) with \( \theta^* = (\theta^*_1, ..., \theta^*_K) \) and \( S = (S_1, ..., S_K) \) a partition of \( \{1, ..., n\}, \)
\[
P(d\theta^*, S) = P(S)P(d\theta^*|S).
\]
Now
\[
P(d\theta^*|S) = \prod_{k=1}^{K} H(d\theta^*_k),
\]
since the \( \theta^* \)-value for the \( k \)'th group was simulated independently from \( H \) at the time of the first arrival into that group. It follows that the marginal distribution of the \( \theta \)'s is
\[
P(d\theta^*, S) = P(S) \prod_{k=1}^{K} H(d\theta^*_k)
\]
\[\text{[VUE]}\]
The steps leading to this expression were (1) to show that marginally, for \( j = 1, 2, ..., n - 1 \)
\[
\theta_{j+1} | \theta_1, ..., \theta_j \sim \frac{\alpha H + \sum_{k=1}^{K} n_k^j \delta_{\theta_k^*}}{\alpha + j},
\]
and then (2) observe that this corresponds to a CRP with appropriate parameters, and then (3) compute the probability distribution over the partitions \( S \) of the indices of \( \theta_1, ..., \theta_n \). Finally, we recalled that the \( \theta^* \)-values were iid from \( H \).

11.9. Inference for a mixture. We will now take \( n \) samples \( y_i \sim f(y_i | \theta_i), i = 1, ..., n \) drawn from a normal mixture with an unknown number of components. We will use a DP model for the prior distribution over clusters and parameter values within clusters.

In a DP model for a general mixture \( y_i \sim f(y_i | \theta_i) \), and our prior for \( \theta_i \) is
\[
\theta \sim G \quad \text{with} \quad G \sim DP(\alpha, H)
\]
In \( S, \theta^* \) notation, Bayes rule for conditional probabilities is
\[
P(d\theta^*, S | y) = f(y | \theta^*, S)P(d\theta^*, S)
\]
so from (VUE) above
\[
P(d\theta^*, S | y) = f(y | \theta^*, S)P(S) \prod_{k=1}^{K} H(d\theta_k^*).
\]
If \( L(\theta^*, S; y) \propto f(y | \theta^*, S) \) and
\[
\pi(\theta^*)d\theta^* = \prod_{k=1}^{K} H(d\theta_k^*)
\]
then in terms of densities
\[
\pi(S, \theta^* | y) \propto L(\theta^*, S; y)\pi(\theta^*)P(S)
\]
with \( P(S) \) given above. We are back in familiar territory. Notice that the number of components \( K \) in \( \theta^* \) is a random variable controlled (in the prior) by the distribution \( P(S) \).

11.10. Normal mixture for the Galaxy data. Each component of the normal mixture has an unknown mean and variance, \( \theta_k^* = (\mu_k^*, \sigma_k^2) \). Our base distribution will be a simple product,
\[
H(d\theta_k^*) = H_\mu(d\mu_k^*)H_\sigma(d\sigma_k^2),
\]
If \( S = (S_1, ..., S_K) \) is a partition of \( \{1, 2, ..., n\} \) with \( n = 82 \), and \( i \in S_k \) then
\[
y_i | S, \mu^*, \sigma^* \sim N(\mu_k^*, \sigma_k^2).
\]
This determines the likelihood. Our priors are
\[
H_\mu(d\mu_k^*) = N(d\mu_k^*; \mu_0, \sigma_0^2)
\]
and
\[
H_\sigma(d\sigma_k^2) = \Gamma(d\sigma_k^2; \alpha_0, \beta_0).
\]
with $\mu_0 = 20, \sigma_0 = 10, \alpha_0 = 2$ and $\beta_0 = 1/9$ (priors) and I will later take $\alpha = 1$ in the DP-prior, so the posterior is

$$
\pi(S, \mu^*, \sigma^* | y) \propto L(\theta^*; y) \pi(\mu^*, \theta^*) P(S)
$$

$$
\propto \prod_{k=1}^{K} \prod_{i \in S_k} N(y_i; \mu_k^*, \sigma_k^*)
$$

$$
\times \prod_{k=1}^{K} N(\mu_k^*; \mu_0, \sigma_0^2) \Gamma(\sigma_k^2; \alpha_0, \beta_0)
$$

$$
\times \alpha^K \prod_{k=1}^{K} \Gamma(n_k).
$$

We dropped the denominator in the expression for $P(S)$ as it does not depend on $S$. Notice that $\alpha^K$ does depend on $S$ as it depends on $K$, the number of clusters.

11.10.1. Remarks on conjugate priors and “Collapsed Gibbs”. I chose Normal/inv-Gamma for the $\mu^*/\sigma^*$-prior to keep things simple and conjugate, so I could Gibbs sample.

If we just did straightforward MH-MCMC on $\mu^*$ and $\theta^*$ that wouldn’t be necessary. Conjugate priors are popular in this field as it allows us to integrate out $\theta^* = (\mu^*, \sigma^*)$ and sample the discrete dbn $\pi(S|y)$. This is called the “collapsed Gibbs sampler” and it is efficient. If our purpose is clustering, $S$ is all we need anyway. The down side is we can’t model $\mu^*$ and $\sigma^*$ with freedom.

11.10.2. Remarks on the choice of $\alpha$. I chose $\alpha = 1$. This controls the prior dbn number of clusters. I used simulation (of the CRP) to check this distribution was sensible. The prior mean (PS4) is

$$
E(K) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}.
$$

which is about $E(K) = 5$ here.

A slightly larger value might make sense. It is often straightforward to impose a hyper-prior on $\alpha$ and infer it along with everything else.

11.10.3. Gibbs sampler for $\mu^*, \sigma^*$. Iterate through the parameters sampling them conditionally. The conditional distribution for $\mu_k^*$ is

$$
\pi(\mu_k^* | \mu_{-k}^*, \sigma^*, y) \propto N(\mu_k^*; \mu_0, \sigma_0^2) \prod_{i \in S_k} N(y_i; \mu_k^*, \sigma_k^*)
$$

We can complete the square and find $\mu_k^* | \sigma_k^*, y \sim N(a, b)$ with

$$
a = b \left( \frac{n_k \bar{y}_k}{\sigma_k^2} + \frac{\mu_0}{\sigma_0^2} \right), \quad b = \left( \frac{n_k}{\sigma_k^2} + \frac{1}{\sigma_0^2} \right)^{-1},
$$

where $n_k = |S_k|$ and $\bar{y}_k = n_k^{-1} \sum_{i \in S_k} y_i$.

A similar calculation gives $\sigma_k^2 | \mu_k^*, y \sim \Gamma(c, d)$ with

$$
c = \alpha_0 + n_k/2, \quad d = \beta_0 + \frac{1}{2} \sum_{i \in S_k} (y_i - \mu_k^*)^2.
$$
11.10.4. **Gibbs sampler for the partition.** Consider the sequence of pairs \((y_i, \theta_i), i = 1, 2, \ldots, n\). This sequence is exchangeable so we may make \(j\) the last arrival. It either takes a new table, or joins an existing table.

Let \(S^{-j} = (S_1^{-j}, \ldots, S_{K-1}^{-j})\) denote the partition with \(j\) removed. Before \(j\) arrived the number of tables was \(K^{-j}\) and the number sitting at table \(k\) was \(n_k^{-j}\). Let \(S'\) denote the partition we get on adding \(j\) to one of the sets in \(S^{-j}\).

Denote by \(\mu^*_j, \sigma^*_j\) the parameters after \(j\) is removed from \(S\). These may differ from \(\mu^*, \sigma^*\) if \(j\) is in a cluster of size one (we would delete that cluster).

The probability \(j\) joins existing table \(k\), for \(k \in \{1, \ldots, K^{-j}\}\),

\[
\Pr(j \in S'_k|S^{-j}, \mu^*_j, \sigma^*_j, y) \propto \frac{n_k^{-j}}{\alpha + n - 1}
\]

by Bayes rule, splitting of \(y_j\) from \(y_{-j}\). Now

\[
\Pr(j \in S'_k|S^{-j}, \mu^*_j, \sigma^*_j, y_{-j}) = n_k^{-j}/(\alpha + n - 1)
\]

since this is the CRP probability to join table \(k\) - there is no data \(y_j\) on the RHS to change this.

Since \(S'\) is the partition we get by adding \(j\) to some set in \(S^{-j}\), the cluster labels of \(S'\) and \(S^{-j}\) are the same. It follows that

\[
p(y_j|S^{-j}, \mu^*_j, \sigma^*_j, y_{-j}, j \in S'_k) = N(y_j; \mu^*_j, k, \sigma^*_j)^2
\]

as \(y_j\) has this distribution, independently given \(j \in S'_k\). We have

\[
\Pr(j \in S'_k|S^{-j}, \mu^*_j, \sigma^*_j, y) \propto n_k^{-j}N(y_j; \mu^*_j, k, \sigma^*_j)^2.
\]

On the other hand \(j\) may join a new table so \(k = K^{-j} + 1\). Let \(S'\) denote the partition formed by removing \(j\) from \(S\) and putting in a new set with a new \(\mu^*_k\) and \(\sigma^*_k\). That table has values

\[
\mu^*_k \sim N(\mu_0, \sigma_0^2)
\]

and

\[
\sigma^*_k \sim \Gamma(\alpha_0, \beta_0)
\]

In the Bayes rule calculation this time

\[
P(j \in S'_k, \mu^*_k, \sigma^*_k|S^{-j}, \mu^*_j, \sigma^*_j, y) \propto p(y_j|S^{-j}, \mu^*_k, \sigma^*_k, y_{-j}, j \in S'_k)
\]

\[
\times P(j \in S'_k, \mu^*_k, \sigma^*_k|S^{-j}, \mu^*_j, \sigma^*_j, y_{-j}).
\]

Now

\[
P(j \in S'_k, \mu^*_k, \sigma^*_k|S^{-j}, \mu^*_j, \sigma^*_j, y_{-j}) = \frac{\alpha}{(\alpha + n - 1)} \times \pi(\mu^*_k, \sigma^*_k)
\]

since the parameter assigned to a new table is an independent draw from the prior.

The contribution from the likelihood term is the same as before,

\[
p(y_j|S^{-j}, \mu^*_k, \sigma^*_k, y_{-j}, j \in S'_k) = N(y_j; \mu^*_k, \sigma^*_k)^2
\]

but uses our “new” \(\mu^*_k\) and \(\sigma^*_k\) values.

We conclude that the conditional probability for \(j\) to be in cluster \(k\) given everything else is

\[
\Pr(j \in S'_k|S^{-j}, \mu^*_j, \sigma^*_j) \propto \left\{ \begin{array}{ll}
\frac{n_k^{-j}N(y_j; \mu^*_k, \sigma^*_k)}{\alpha N(y_j; \mu^*_k, \sigma^*_k)}, & k = 1, \ldots, K^{-j} \\
\alpha N(y_j; \mu^*_k, \sigma^*_k), & k = K^{-j} + 1.
\end{array} \right.
\]

where in the second case (of adding a cluster), \(\mu^*_k, \sigma^*_k \sim \pi(\mu^*_k, \sigma^*_k)\) independently of everything else.
We estimate the posterior predictive distribution (black) and the predictive distribution conditioned on 3, 4 and 5 components (red, green and blue) and plot these PPD’s over the data, as we did in lecture 11 for the RJ-MCMC fit. See associated R-code for this lecture for implementation.
12. Laplace Approximation

We didn’t have much time for this, just half a lecture. But there are a few useful points to be made, which lead to new insights on familiar objects. See CR-TBC for more on this. Suppose we have an integral \( I_n \) to do of the form

\[
I_n = \int e^{-nh(\theta)} d\theta
\]

Let \( \hat{\theta} \) satisfy \( h(\hat{\theta}) = 0 \), \( h^{(k)} \) be the \( k \)'th derivative, \( \hat{h} = h(\hat{\theta}) \) and \( \sigma^2 = 1/\hat{h}^{(2)} \). It may be shown that

\[
I_n = e^{-nh(\hat{\theta})} \sqrt{\frac{2\pi}{n\sigma}} \left( 1 + \frac{5[\hat{h}^{(3)}]^2\sigma^3 - 3\hat{h}^{(4)}\sigma^2}{24n} \right) + O(n^{-2})
\]

Exercise: verify this; expand \( h \) as a Taylor series to fourth order about \( \hat{\theta} \), pull out the constant and quadratic terms, then further expand the exponential \( \exp(f(\theta)) = 1 + f + f^2/2 + ... \) for the terms involving \( \hat{h}^{(3)} \) and \( \hat{h}^{(4)} \). This leaves normal moments \( E(θ - \hat{θ})^k \) for \( k = 4, 6, 8 \) which you can look up!

The multivariate version \( \theta = (θ_1, ..., θ_p) \) of this is

\[
\int e^{-nh(\theta)} d\theta = e^{nh(\hat{\theta})}(2\pi/n)^{p/2}|Σ|^{1/2} + O(n^{-1})
\]

where

\[
Σ^{-1} = \left. \frac{∂^2 h}{∂θ∂θ^T} \right|_{\hat{θ}}
\]

is the Hessian of \( h(θ) \) at \( \hat{θ} \).

12.1. Marginal Likelihood. For a model \( m \) and \( n \)-component data \( y \)

\[
p(y|m) = \int \exp(-n[\bar{ℓ}(θ) - \log(π(θ))/n]) dθ.
\]

where

\[
\bar{ℓ}(θ) = \frac{1}{n} \ell(θ; y).
\]

If for example \( y_1, ..., y_n \) are jointly independent,

\[
\bar{ℓ}(θ) = \frac{1}{n} \sum_{i=1}^{n} \ell(θ; y_i)
\]

in which case \( \bar{ℓ} \to E_y(\log(p(y|θ))) \), the (negative) entropy of \( p(y|θ) \).

Using the Laplace approximation, with \( h(θ) = -\bar{ℓ}(θ) - \log(π(θ))/n \),

\[
\int e^{-nh(\theta)} d\theta = e^{nh(\hat{\theta})}(2\pi/n)^{p/2}|Σ|^{1/2} + O(n^{-1})
\]

\[
p(y|m) \simeq L(\hat{θ})π(\hat{θ})(2π/n)^{p/2}|Σ|^{1/2}
\]

where \( \hat{θ} \) is the MAP estimate for \( θ \), and \( Σ \simeq j^{-1}(\hat{θ}) \), the observed Fisher information per unit observation, \( j = J/n \). At large \( n \) the contribution from \( π/n \) is \( O(1/n) \) so

\[
\frac{∂^2 h}{∂θ∂θ^T} \simeq -\frac{∂^2 \bar{ℓ}}{∂θ∂θ^T}.
\]

This can be used to estimate the marginal likelihood: find the MAP, calculate \( Σ^{-1} \), evaluate \( p(y|m) \).
There is also a direct normal approximation to the posterior, expanding $h(\theta) = -\bar{\ell}(\theta) - \log(\pi(\theta))/n$ about the MLE $\hat{\theta}_{ML}$

$$\pi(\theta | y) \simeq \pi(\hat{\theta}_{ML}) \exp(-n\bar{\ell}) \exp(-n/2(\theta - \hat{\theta}_{ML})^T j(\hat{\theta})(\theta - \hat{\theta}_{ML}))$$

giving

$$\theta | y \sim N(0, j(\hat{\theta}_{ML})^{-1}/n)$$

approximately at large $n$. A more accurate approximation (using $h$ above) expands about the MAP $\hat{\theta}$ but gives the same asymptotic form.

Notice the similarity to the frequentist results for the MLE $\hat{\theta}_{ML} \sim N(\theta, i(\theta)^{-1}/n)$ and $\hat{\theta}_{ML} \sim N(\theta, j(\hat{\theta}_{ML})^{-1}/n)$ approximately at large $n$.

12.2. The Bayesian information criterion, BIC. Replacing $\hat{\theta}$ with $\hat{\theta}_{ML}$ in the Laplace approx. to $p(y|m)$,

$$\log(p(y|m)) \simeq \ell(\hat{\theta}_{ML}) + \log(\pi(\hat{\theta}_{ML})) + \log |\Sigma|^{1/2} + \frac{p}{2} \log(2\pi/n)$$

Now $\ell$ is $O(n)$, so dropping terms which do not grow with $n$,

$$-2\log(p(y|m)) \simeq -2\ell(\hat{\theta}_{ML}) + p \log(n).$$

This is the BIC. The model with the least BIC score has the largest marginal likelihood, approximately. CR-TBC notes it isn’t Bayesian, as the prior doesn’t enter. Under a unit information prior on $\theta$ the terms $O(1)$ also cancel and it becomes $O(1/\sqrt{n})$ approximation, which mitigates. Like all estimators of this sort “it works when it works” and cross comparison with other estimators makes sense.

12.3. Remarks on the AIC (not examinable). The AIC, $AIC = -2\ell(\hat{\theta}_{ML}) + 2p$, approximates an unbiased estimator for

$$-2E_{Y,Y'}(\ell(Y'; \hat{\theta}_{ML}(Y))).$$

The idea here is that $Y$ is the data, $Y'$ is new data, and we like models that are good at predicting new data. See Davison Statistical Models (2003) Sect 4.7 for a derivation and insight.

Suppose we had a prior over models $\pi(m)$. Then $\pi(m|y) \propto \pi(m)p(m|y)$, so $\pi(m|y)$ is approximately

$$\hat{\pi}_B(m|y) \propto e^{-BIC_m/2} \pi(m).$$

If $\pi(m)$ is flat, maximising this is the same as using BIC. Now for model $m$

$$AIC_m = BIC_m + 2p_m - p_m \log(n),$$

so using the AIC is like maximising

$$\hat{\pi}_A(m|y) \propto e^{-BIC_m/2 - p_m + (p_m/2) \log(n)} \pi(m).$$

This is like weighting the maximisation using a prior $\hat{\pi}(m) \propto \pi(m)e^{-p_m + (p_m/2) \log(n)}$ over models. The prior weights in favor of more complex models as data size grows! You may have noticed that the AIC favors more complex models than the BIC (or ANOVA). The BIC is consistent (ie, asymptotically correct model, under mild conditions); not AIC.

E-mail address: nicholls@stats.ox.ac.uk