Bayes Methods - SC7 lecture notes, HT23

Geoff K. Nicholls, nicholls@stats.ox.ac.uk

Department of Statistics, University of Oxford

These notes will be updated through the term.

There are no doubt many typos in these notes, for which, apologies. Let me know if you spot any.

Problem sheet 1 - Lectures 1-3 cover standard background, Chapters 1-2.3
Problem sheet 2 - Lectures 4-7 cover Sections 2.4-4.2
Problem sheet 3 - Lectures 8-11 cover Sections 4.3-7.2
Problem sheet 4 - Lectures 12-16 cover Sections 7.3-8

This version: 12-03-23. Revisions pages 114 and 115 (in blue). These notes are complete.

Contents

1 The Bayesian inferential pipeline

1.1 Preamble

This chapter is an overview introducing some of the main ideas we will revisit. I assume that to some extent the ideas in lectures 1-4 (up to the end of Section 2) are familiar, so this is a whistle-stop tour, but I would like to get us on the same page and offer a few insights along the way.

The course aims to get you to a point where you have some idea of how to carry out subjective Bayesian inference in practice. One of the appealing things about Bayesian inference is that, at a certain level of abstraction, it is "always the same". The sequence of operations, forming the prior, observation model and loss function, expressing the posterior, and then computing summary statistics representing parameter estimates and credible sets, follow an "inferential pipeline" which can to some extent be automated. This allows the analyst to focus on building a generative model for the data. Software tools implement this process in a generic way - STAN is an outstanding example, but there are many others with a focus on making it easy to carry out inference for specialised data types and model classes. In order to illustrate the methods and key algorithms I give R-implementations and make the code available. This may help you understand what is happening, and help you if you need to make your own applications, but the R and coding is not examined in the written exams.

The topics below are chosen to represent certain important themes in modern Bayesian research, chosen to some extent for their immediate potential for use in application, but also to illustrate the subject as a whole: fundamentals (Chapters 4 and 5), computational methods (Chapters 2, 3 and 7) and different varieties of model (Chapter 8). There is a whole branch of Bayesian inference called objective Bayes which we don't really discuss. Objective Bayes methods put more weight on frequentist or other abstract properties of the statistical inference. In subjective methods the model summarises our physical knowledge of the system we are trying to understand through the data and so there is more emphasis on statistical modelling.

The material is presented in the same order as the lectures. We do "hop about" a bit. The more "applied" material presented in the first two weeks becomes a kind of extended example for some of the more "theoretical" work in the second half.

1.2 Lecture 1: Bayesian inference

Statistical inference has a "pipeline". For example for a hypothesis test, it would be exploratory data analysis then data modeling then parameter estimation and model selection using MLE's and Likelihood Ratio Tests then goodness of fit checking and then reporting.

In a Bayesian setting for the same sort of task we might start with exploratory data analysis and data modeling, followed by prior elicitation. Parameter estimation and model selection using the posterior mean and Bayes Factors follows and then goodness of fit and reporting. If the goal is prediction rather than testing then we often replace model selection with model averaging for a better quantification of uncertainty. More on that below and in Chapter 7.

1.2.1 Prior, Observation model, posterior

Some unknown real-world quantity Θ takes values in a parameter space Ω. Typically Ω is a p-dimensional subset of R^p below. Let $S \subseteq \Omega$ be a subset of possible values for Θ. Is Θ in S?

We have some knowledge that may help us answer this question, and we additionally gather data. If our knowledge of the world is *coherent* (in the sense of Chapter 4) then it can be represented by a prior distribution with density $\pi(\theta)$ on Ω which exists and is unique. We have a straightforward measure of the strength of our belief that $\Theta \in S$ holds,

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

The prior $\pi(\theta),\ \theta \in \Omega$ represents our state of knowledge before we see the data. Notice that the thing we dont know, the unknown true value Θ, is a random variable here, representing the uncertainty associated with the fact that it is unknown. We write $\Theta = \theta$ if $\theta \in \Omega$ is a possible realisation of the random variable.

The probability density or mass function $\pi(\theta)$ determines a probability distribution when taken with a measure and a suitable σ-algebra - for brevity I refer to "the distribution $\pi(\theta)$". We will occasionally need to use measure theory notation, though all the measures which appear are built from counting measure (*ie* sums) or volume measure on $\mathbb{R}^p$ (*ie* Lebesque integrals) and so I generally omit this layer. As an example if $\Omega = \mathbb{R}^p$, $\mathcal{B}_\Omega$ is the Borel σ-algebra of subsets of Ω, $d\pi(\theta)$ is a general probability measure on Ω, $d\theta$ is Lebesque volume measure in Ω, then $d\pi(\theta) = \pi(d\theta)$ because we have volume measure and $\pi(d\theta) = \pi(\theta)d\theta$ as we have a density. If $S \in \mathcal{B}_\Omega$ then $\pi(S) = \int_S \pi(d\theta)$ is a well defined probability, so $\pi : \mathcal{B}_\Omega \to [0,1]$ and we use the same symbol for the distribution and the density and know which is which by the argument, set $S \in \mathcal{B}_\Omega$ or value $\theta \in \Omega$ respectively.

Observations $Y = (Y_1, ..., Y_n), Y \in \mathcal{Y}$ are distributed according to an observation model with probability density $p(y|\theta)$ for realisation $Y = y$ with $y = (y_1, ..., y_n)$ given $\Theta = \theta$. The observations $Y_i,\ i = 1, ..., n$ are themselves vectors in $\mathbb{R}^d$ so to be clear, $p(y|\theta)$ is the joint density of *all* the data. For example, if the observations are iid then

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta),$$

and $p(\cdot|\theta)$ is a density on $\mathcal{Y}$ on the left and on $\mathbb{R}^d$ in the product on the right. The likelihood $L(\theta; y) = p(y|\theta)$ is a function of θ at fixed y. These notes generally refer to $p(y|\theta)$ rather than $L(\theta; y)$ even when we are thinking of $p(y|\theta)$ as a function of θ.

Suppose we observe $Y = y$. How do our beliefs about $\Theta \in S$ change? If $\pi(\theta|y)$ is the posterior density of the *posterior distribution* $\pi(S|y) = \Pr(\Theta \in S|Y = y)$ then by Bayes rule

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

with

$$p(y) = \int_\Omega p(y|\theta)\pi(\theta)d\theta$$

the normalising *marginal likelihood*. Here $p(y)$ is also the *prior predictive distribution* of the data - if $\theta \sim \pi(\cdot)$ and $y' \sim p(\cdot|\theta)$ then marginally $y' \sim p(\cdot)$.[1]

[1]It is good practice to use capitals for RV and lower case for realisations. I will stop doing this in subsequent

Answers to questions about Θ 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 unknown true Θ lies in S, given the data and the prior knowledge expressed in $\pi(\theta)$.

1.3 The Generative model

When we come to do model selection in Bayesian inference, the model we are selecting is the generative model for the data, with joint density

$$p(y, \theta) = p(y|\theta)\pi(\theta).$$

It is helpful to consider how nature realised the data (if we have nature's generative model).

Algorithm 1.1. *Simulation of Θ, Y given a prior $\pi(\cdot)$ and observation model $p(\cdot|\theta)$.*

1. *simulate $\theta \sim \pi(\cdot)$*
2. *simulate $y \sim p(y|\theta)$*
3. *return $(\Theta = \theta, Y = y)$.*

This is what nature does to realise the observed data y_{obs} - it generates a realisation of Θ then realises observations $Y|\Theta$ and returns $Y = y_{obs}$. The conditional distribution of $\Theta|Y = y_{obs}$ is proportional to the joint model $p(y|\theta)\pi(\theta)$. We can turn this round and use it to simulate the conditional.

Exercise 1.2. Suppose the data Y and Θ are discrete random variables and the observed data are $Y = y_{obs}$. Consider the simulation algorithm

1. Simulate $\Theta' \sim \pi(\cdot)$ and suppose $\Theta' = \theta$ is realised. Simulate $Y' \sim p(\cdot|\theta)$ realising $Y' = y$.
2. If $y = y_{obs}$ stop and return $\Theta = \theta$ and otherwise goto Step 1,

Show that the returned θ-values are distributed like the posterior, so $\Theta \sim \pi(\theta|y_{obs})$.

ANS: the pairs $(\Theta', Y') \sim p(\theta, y)$ have the joint distribution $p(y, \theta) = p(y|\theta)\pi(\theta)$ and selecting a pair with $Y' = y_{obs}$ is just conditioning on $Y' = y_{obs}$. The conditional distribution $p(\theta|y)$ of $\Theta'|Y' = y_{obs}$ is then equal to the posterior $\pi(\theta|y)$. We select the "first" such pair. This is fine. If you would like to check, recognise this as a rejection algorithm, or just compute the probability to return $\Theta = \theta$ by summing over all sequences rejections that return $\Theta = \theta$.

1.3.1 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.

chapters. The context will tell us if the quantity is random. For example in $E_\theta(f(\theta, y))$ y is fixed and $\theta \sim \pi(\cdot)$ has the prior distribution, $E_{\theta|y}(f(\theta, y))$ is the same but $\theta \sim \pi(\cdot|y)$ has the posterior distribution, and in $E_{\theta,y}(f(\theta, y))$ the pair $(\theta, y) \sim \pi(\theta)p(y|\theta)$ are both random and have the generative model as their joint distribution.

Talk to the scientists or use your own common sense knowledge about the world. There is often an attempt to present statistics as a logically closed subject. It cannot be. It is a language for formalising knowledge about the world. The scientists usually have insights about the quantities they are estimating and it is often possible to express this in the full generative model in a neat way.

Prior elicitation checklist

1. Is the parameter θ generated by some process we can model? If so then the the distribution over θ determined by the process *is* the prior.
2. Do the model elements correspond to "elements of reality" - if the parameters correspond to real world quantities it will be easier to identify prior knowledge. If you introduce these parameters as latent variables you may make modelling easier.
3. 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.
4. 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 doesnt overwhelm data-information for parameters which are poorly informed by the data [this leads to objective Bayes].
5. 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. For example if we have a parameter $\theta \in [0, 1]$ and we are interested in whether it is greater than 0.99 then the uniform prior $\theta \sim U(0, 1)$ is strongly informative. If we are using the posterior as a summary then it will reflect this information. Non-informative does not in general equal uniform. Ask, non-informative with respect to what function of the parameter?
6. Is the number of things you dont know one of the things you dont know? In this case you may need to put a prior on the number of unknowns!
7. 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.
8. It isnt 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. We are asking what conclusions another analyst would reach if they started with a different state of knowledge.

The fact that the prior represents a particular state of knowledge makes the analysis subjective. However, many of the remarks above could be made about the observation model (ie, the likelihood) or indeed any essentially parametric statistical model constructed for any purpose. In fact it is often arbitrary what we designate as prior and what is observation model. This is typical in latent variable or hierarchical models. When we have latent variables $\psi \sim \pi(\psi|\theta)$ and a generative model $p(y|\psi)\pi(\psi|\theta)\pi(\theta)$ the status of $\pi(\psi|\theta)$ is in question.

The challenge of building a prior is quite daunting - or should be - one is faced with the problem of building a mathematical model of some aspect of the world, and this must require knowledge which is not simply "statistical". Think of it as an opportunity to add information that comes with an obligation to check it is representative of the information you have.

1.3.2 Decision Theory for Bayesian estimation

If we are in interested in the value or "location" of the true parameter then we may wish to report an estimate $\delta \in \Omega$ for Θ. This will be informed by the data and prior so we must have $\delta = \delta(y)$. In the language of decision theory, δ is the action and we assume the action space is $\Omega = \mathbb{R}^p$. We pay a penalty or "loss" $L(\Theta, \delta)$ for getting our estimate wrong. We have to choose a suitable loss that represents the actual cost to us of error, so like the prior, the loss must be elicited - ie we gather the information externally and represent the loss mathematically. The loss we choose depends on the downstream use we plan to make of the estimate.

For data with observation model $Y \sim p(\cdot|\theta)$ $Y \in \mathcal{Y}$, the Θ-estimator $\delta(Y)$, $\delta : \mathcal{Y} \to \mathbb{R}^p$, is an action for each $y \in \mathcal{Y}$, with *risk* $\mathcal{R}(\theta, \delta)$ at $\Theta = \theta$ given by

$$\begin{aligned}\mathcal{R}(\theta, \delta) &= E_{Y|\Theta=\theta}(L(\theta, \delta(Y))) \\ &= \int_{\mathcal{Y}} L(\theta, \delta(y))p(y|\theta)dy.\end{aligned}$$

If we have a prior $\pi(\theta)$, posterior $\pi(\theta|y)$ and marginal likelihood $p(y)$ the *Bayes risk*, $\rho(\pi, \delta)$, is the risk averaged over the prior,

$$\begin{aligned}\rho(\pi, \delta) &= E_\Theta(\mathcal{R}(\theta, \delta)) \\ &= E_{\Theta,Y}(L(\Theta, \delta(Y))) \\ &= \int_\Omega \int_{\mathcal{Y}} L(\theta, \delta(y))p(y|\theta)\pi(\theta)dyd\theta.\end{aligned}$$

A Bayes estimator δ^π for θ minimises the Bayes risk

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

This is not straightforward as an estimator δ is a function, so we are minimising over all *functions* $\delta : \mathcal{Y} \to \mathbb{R}^p$. However the problem can be re-expressed in a simpler way. If the *Expected Posterior Loss* is defined to be

$$\begin{aligned}\rho(\pi, \delta|y) &= E_{\Theta|Y=y}(L(\Theta, \delta(y))) \\ &= \int_\Omega L(\theta, \delta(y))\pi(\theta|y)d\theta \qquad (1.1)\end{aligned}$$

then the Bayes risk can be written in the convenient form

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

If we have an estimator minimising the expected posterior loss at every $y \in \mathcal{Y}$, that is

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

then is must define the function minimising the Bayes risk: it minimises the integrand at every y so it minimises the integral.

Suppose for example our loss for estimating δ when the truth is Θ is given by the square error, $L(\Theta, \delta(Y)) = (\Theta - \delta(Y))^2$, expressing "closer is better and far off is very bad". In this case the expected posterior loss $E_{\Theta|y}(L(\Theta, \delta))$ is minimised over actions by the posterior mean, $\delta^* = E_{\Theta|y}(\Theta)$ (just differentiate wrt $\delta(y)$ at fixed y). Since the action minimising the expected posterior loss minimises the Bayes risk, this is the Bayes estimator.

1.3.3 Admissiblity

I suggest the reader skips this section at first reading and returns to it prior to reading Section 4. It belongs with the section above in terms of topic, and is included for motivation of Bayes methods, but interrupts the presentation of the Bayesian inferential pipeline which is our purpose in Chapter 1.

If we accept the loss function $L(\theta, \delta)$ we would never use an estimator δ_0 which was "never better and sometimes worse". If there exists an estimator δ_1 satisfying

$$\mathcal{R}(\theta, \delta_0) \geq \mathcal{R}(\theta, \delta_1)$$

and for at least one θ_0,

$$\mathcal{R}(\theta_0, \delta_0) > \mathcal{R}(\theta_0, \delta_1)$$

then we say δ_0 is not admissible. Otherwise it is admissible.

Estimators that seem reasonable (recall James-Stein beats MLE if risk is MSE) need not be admissible. In fact every admissible estimator is either a Bayes estimator or can be expressed as the limit of Bayes estimators. This feature is often pointed to as an advantage of Bayesian inference and is sometimes used to derive good estimators for frequentist inference (for example the derivation of the James-Stein estimator) where the method is often called "empirical Bayes".

Proposition 1.3. *(Proposition 2.4.22 in the textbook CR-TBC): If prior π is strictly positive on Ω with finite Bayes risk, and the risk, $\mathcal{R}(\theta, \delta)$, is a continuous function of θ, then Bayes estimator δ^π is admissible.*

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

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

that is

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

But that is impossible as

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

by definition. [EOP]

1.3.4 Forming summaries and Monte Carlo

Suppose we wish to estimate the expectation in the posterior of some function $f(\theta)$. If the loss is the square error then we estimate $f(\Theta)$ with the posterior mean $E_{\Theta|Y=y}(f(\Theta))$ which we commonly estimate in turn using Monte Carlo. We simulate $\theta^{(t)} \sim \pi(\cdot|y),\ t = 1, ..., T$ and compute

$$\hat{f} = \frac{1}{T}\sum_{t=1}^{T} f(\theta^{(t)}).$$

For example, if $S \in \mathcal{B}_\Omega$ and $f(\theta) = \mathbb{I}_{\theta \in S}$ then $\hat{f}$ estimates $\pi(S|y)$.

We also commonly report posterior credible sets in order to quantify uncertainty. A level-α Highest Posterior Density (HPD) credible set C_α satisfies

$$\int_{\Omega \cap C_\alpha} \pi(\theta|y) d\theta = 1 - \alpha,$$

with the additional constraint

$$\text{if } \theta \in C_\alpha \text{ and } \theta' \in \Omega \setminus C_\alpha \;\Rightarrow\; \pi(\theta|y) \geq \pi(\theta'|y).$$

The HPD set can be estimated from Monte Carlo samples $\theta^{(t)} \sim \pi(\cdot|y),\ t = 1, ..., T$ (see Exercise 1.5 below). An HPD set (or general credible set with fixed posterior probability mass) is qualitatively different in meaning from a frequentist confidence interval. The probability a CI covers the true parameter *under replication of the data* is $1 - \alpha$ so a CI for a real parameter can be the whole of $\mathcal{R}$, or the empty set (see Q8 Section 8.9 of Davison "Statistical Models" (2003)). For an HPD set we can say "given the model and data, the probability that the true parameter is in the credible set is $1 - \alpha$" which would be incorrect for a CI.

The posterior predictive distribution of the data

$$p(y'|y) = \int_\Omega p(y'|\theta)\pi(\theta|y) d\theta$$

is useful in comparing models and carrying out goodness of fit: if the model is good then the data should be "typical"; our real data should predict new data that resembles that real data and hence simulated data $y' \sim p(\cdot|y)$ should resemble the real data y. This can be measured using summary statistics on the data and looking to see if the summary computed on the real data lies in the tail of the posterior predictive distribution.

Exercise 1.4. The HPD set is a Bayes estimator. Since it is a subset of Ω, the action δ must be a subset of Ω, so suppose the action space is $\delta \in \Delta,\ \Delta = \{A \in \mathcal{B}_\Omega : \pi(A|y) = 1 - \alpha\}$. Consider the loss $L(\Theta, \delta) = \mathbb{I}_{\Theta \notin \delta} + |\delta|$ where $|\delta| = \int_\delta d\theta$ is the volume of the set δ. Verify that the expected posterior loss is minimised over the action space by $\delta^* = C_\alpha$ an HPD set.

ANS: the EPL is $1 - \pi(\delta|y) + |\delta| = \alpha + |\delta|$ which is minimised over Δ by any set with (equal) least volume. This is achieved by an HPD set C_α. If $A \in \Delta$ is some other set satisfying $A \cap C_\alpha = \emptyset$ then $1 - \alpha \leq |A| \max_{\theta \in A} \pi(\theta|y)$ and $1 - \alpha \geq |C_\alpha| \min_{\theta \in C_\alpha} \pi(\theta|y)$ and then $\min_{\theta \in C_\alpha} \pi(\theta|y) \geq \max_{\theta \in A} \pi(\theta|y)$ gives $|C_\alpha| \leq |A|$. If the sets have non-empty intersection then remove the shared set and the volume of $C_\alpha \setminus A$ is smaller than $A \setminus C_\alpha$ and they have the same probability mass.

Exercise 1.5. Practical Monte-Carlo estimation of an HPD set from samples is not so easy especially when $\Omega = \mathbb{R}^p$ with p at all large. However there is the following trick: suppose $\Theta \sim \pi(\cdot|y)$ is a continuous random variable and suppose $Q = \pi(\Theta)p(y|\Theta)$ is a continuous random variable with cdf F which we assume is strictly increasing. Let $\theta^{(t)} \sim \pi(\cdot|y),\ t = 1, ..., T$ be iid posterior samples; let $q^{(t)} = \pi(\theta^{(t)})p(y|\theta^{(t)})$ and let $q^{\{t\}},\ t = 1, ..., T$ denote the sorted values of q from smallest to largest. Explain why the HPD set is in general of the form

$$C_\alpha = \{\theta \in \Omega : \pi(\theta)p(y|\theta) > c_\alpha\},$$

where $c_\alpha = F^{-1}(\alpha)$. Hence show that $\hat{q} = q^{\{\lfloor \alpha T \rfloor\}}$ is a consistent estimate for the threshold c_α and comment briefly on how this might be used to estimate an HPD set in $p = 1$ dimension. Hint: you may assume that a suitably scaled order statistic is a consistent estimator for its quantile.

ANS: the set C_α satisfies the second part of the definition of an HPD so we need to check $\pi(\Theta \in C_\alpha|y) = 1 - \alpha$. Since $\Theta \in C_\alpha$ if and only if $Q > c_\alpha$ we have $\pi(\Theta \in C_\alpha|y) = \Pr(Q > c_\alpha) = 1 - F(c_\alpha)$ and we need this equal $1 - \alpha$, so C_α is an HPD set if $c_\alpha = F^{-1}(\alpha)$. Now $q^{\{\lfloor \alpha T \rfloor\}}$ is the empirical α-quantile of Q so by the hint it is a consistent estimator for c_α. If we estimate $\hat{c}_\alpha = q^{\{\lfloor \alpha T \rfloor\}}$ then all the samples $\{\theta^{(t)} : q^{(t)} > \hat{c}_\alpha, t = 1, ..., T\}$ are inside the HPD set and the intervals in $\mathcal{R}$ covering these samples and no others converge to the HPD set.

1.3.5 Model selection

Suppose we are considering a discrete set $\mathcal{M}$ of models indexed by integers $m = 0, 1, 2, \ldots$. What do we mean by a model? In Bayesian inference we have, for model m, a parameter prior $\Theta \sim \pi(\theta|m)$, $\Theta \in \Omega_m$ and an observation model $Y \sim p(y|\theta, m), Y \in \mathcal{Y}$. The parameter space may vary from model to model. The "model" is the joint model for the "generative process" for Θ, Y, with joint density $\pi(\theta|m)p(y|\theta, m)$ and state space $\Omega_m \times \mathcal{Y}$. We made this explicit in Section 1.3. All aspect of this model are up for selection.

In this context it is natural to treat the model index m as parameter. There is an unknown true model $M \in \mathcal{M}$ say. Given $M = m$ the true generative model is $\pi(\theta|m)p(y|\theta, m)$. Conditioning on $Y = y$ we get the posterior under model $M = m$,

$$\pi(\theta|y, m) = \frac{p(y|\theta, m)\pi(\theta|m)}{p(y|m)}$$

with

$$p(y|m) = \int_{\Omega_m} p(y|\theta, m)\pi(\theta|m)d\theta$$

the marginal likelihood under model m.

We can now shift the discussion up a level to model space. If we have reason to favor some models over others then, since M is a discrete parameter, we express this prior preference in terms of a probability mass function $\pi_M(m)$ over $m \in \mathcal{M}$. The posterior model probability is

$$\pi(m|y) = \frac{p(y|m)\pi_M(m)}{p(y)},$$

where $\pi_M(m)$ is our prior probability that m is the correct model and

$$p(y) = \sum_{m \in \mathcal{M}} p(y|m)\pi_M(m)$$

is the marginal likelihood, now averaged over models.

Under the $0-1$ loss function with truth M and action $\delta \in \mathcal{M}$ the loss is $L(M, \delta) = \mathbb{I}_{M \neq \delta}$, so we loose 1 unit if we get the model wrong and 0 if we get it right. The expected posterior loss $E_{M|y}(L(M, \delta)) = 1 - \pi(\delta|y)$ and this is minimised by the choice $\delta = m^*$ with m^* the mode, the most probable model *a posteriori*. It follows that the Bayes estimator (ie, the action minimising the Bayes risk for this loss, which equals the action minimising the expected posterior loss) is the maximum a posteriori (MAP) model. If we accept the $0-1$ loss as representative of the price we pay for error then we select the MAP model.

I mention briefly here the *model averaged posterior* which allows for uncertainty in the model

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

which appears at the end of question 1 in Problem Sheet 1. We return to it in detail in Section 6.

1.3.6 Why do model selection?

The problem of model selection and hypothesis testing are formally the same thing in Bayesian inference. If the hypothesis is an explicit statement about the value of the parameter then this can

be expressed by choosing priors that express the belief represented in the hypotheses. However the setup allows us to compare any two generative models that model the same data. The data must be the same.

It isn't actually all that common that model selection is a sensible thing to do. Commonly we should aim to do model averaging. This is discussed at the start of Chapter 7. In particular it is generally sub-optimal to select a model and then estimate a parameter. This is called estimation after model selection and although it is common in practice we should be aware that we are making a kind of an approximation. This will become more obvious from a decision theory point of view when we discuss model averaging in Chapter 7, but intuitively, when we choose a model and then estimate, we condition on the model choice. This choice is subject to uncertainty, and may be wrong. We do nevertheless in practice do estimation after model selection out of necessity - we will see that carrying out a full analysis averaging over models is computationally very challenging. Sometimes, once we see the model that has been selected and reflect on it we become very confident we have the right model and so we go forward to see what the model implies for the parameter.

When the number of models is large, the fragility of a $0-1$ loss is exposed. Our chances of getting the right model are very small. The problem is that this is not typically our real loss - models in a large family are often more or less similar to one another, so a "closer is better" loss usually makes sense. This can have dramatic consequences. See for example Figure 2 in this paper

`https://aip.scitation.org/doi/abs/10.1063/1.1381890`

Generally speaking model selection and the $0-1$ loss make sense when the number of models is small.

Some situations where model selection may make sense.

1. Model Construction/Improvement - we have a small number of generative models and we want to find the one that best describes the data. Rather similar to goodness of fit. An example might be that we want to check for sensitivity to the choice of link function in a GLM by comparing against an alternative.
2. Model comparison - two scientists have different beliefs about θ (so, different priors but agree on the observation model) and want to decide which is more in line with reality. You can see an example of this in Section 1.4.
3. 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 process generating the data. Which one?
4. Goodness of fit/Model expansion - we want to check a model M0 is adequate. We define a model M1 incorporating likely model extensions and compare M0 and M1. We are hoping we will reject the more complex model. This is similar in spirit to residual deviance tests against the saturated model in GLM's.

There is some overlap between the situations listed above.

1.3.7 How to do model selection

Suppose we choose the model with the largest posterior probability as above. If m and m' are two models we favor m if

$$A_{m,m'} = \pi(m|y)/\pi(m'|y) > 1.$$

The posterior odds $A_{m,m'}$ have a simple meaning. Model m is $A_{m,m'}$ 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 $A_{m,m'}$ if $\pi_M(m) = \pi_M(m')$, ie, if the prior weighting is equal. The Bayes factor measures the relative support for the whole generative model coming from the data. It has the same straightforward meaning as the posterior odds, if the model-prior weights $\pi_M(m) = \pi_M(m')$ are equal.

Marginal likelihoods, Posterior odds and Bayes factors have a built in penalty on model complexity. As the prior becomes more diffuse, or high dimensional, the probability mass it puts on the support of the likelihood goes down, so, unless there is a compensating increase in the likelihood,

$$p(y|m) = E_{\Theta|M=m}(p(y|\Theta, m))$$

tends to decrease with increasing model complexity. We see this formally in Chapter **??** where we compute the Bayesian Information Criterion (BIC). The BIC for a model is monotone in an approximation to the marginal likelihood for that model, and we will see that it has a built in penalty on model complexity.

Because they are averages ($p(y|\theta, m)$ averaged over $\pi(\theta|m)$), ML's depend on $p(y|\theta, m)$ and $\pi(\theta)$ as functions of θ *everywhere* in Ω_m (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 θ) then $B_{m,m'}$ may be distorted. We may have a generative model $\pi(\theta|m)p(y|\theta, m)$ that is good around the true Θ but poor elsewhere and hence is rejected. If we are using priors to represent hypotheses we have to be careful the tail behavior of the generative model is good.[2]

Notice that $p(y|m)$ is the prior predictive distribution (under model m) for the data y we observed, so the Bayes factor is measuring how good the models are at predicting the data. If the models are mispecified we can't interpret $\pi(m|y)$ straightforwardly (it should be zero for all models!). If we use the Bayes factor to choose a model then we get the one that best predicts the data we saw, so this is still meaningful.

Marginal likelihoods are often very hard to estimate (as a function of θ, $p(y|\theta)$ is typically concentrated in Ω, π is diffuse) and this simple computational obstacle has been one of the principal obstacles to more widespread use of Bayes methods.

1.3.8 Multiple testing

(see D. Cox, *Principles of Statistical Inference*)

This material is optional and should be omitted at first reading. It was covered more carefully in previous years.

Suppose we have a baseline model M_0 and K alternatives $M_k, k = 1...K$, which are tested using K different independent data sets (one each) and we report the model with the smallest P-value, P_{min} say, in a frequentist setup in which M_0 is the null model. What level $\tilde{\alpha}$ should we take to get

[2]This leads to the Lindley paradox.

a false positive rate equal α? If P_{min} is uniform under M_0 then

$$\Pr(P_{min} < \tilde{\alpha}) = 1 - (1 - \tilde{\alpha})^K,$$

so if we take $\tilde{\alpha} = \alpha/K$ we get $\Pr(P_{min} < \alpha/K) \simeq \alpha$, a good approximation if α is not too large. This is called a Bonferroni correction. It is rather crude as it assumes independence of P-values whilst in practice the tests will use a common data set. We require a higher level of evidence to allow for the hazard of multiple testing.

If we have K Bayes factors $B_{m,0}, m = 1...K$ and report the largest there is a similar issue. For example, suppose we have K genes and we think one of them might be associated with some disease. The null model is that none of them are associated with the disease. We might estimate the model/gene $m = m^*$ with highest value of $p(y|m)$ and report its Bayes factor $B_{m,0}$ as a measure of the strength of evidence. Our intuition tells us there must be some equivalent to a hazard for multiple testing here. Careful consideration of the prior $\pi_M(m)$ over models does indeed lead to some down-weighting of the evidence.

For example, consider the model prior

$$\pi_M(0) = 1/2, \quad \text{and} \quad \pi_M(k) = 1/2K, \quad \text{for} \quad k = 1..K.$$

This says that the event that none of the genes have an effect, and the event that one of them does, are equally likely a priori. It is non-informative with respect to the hypothesis that there is an effect, and then non-informative with respect to which gene gives the effect, given there is one. The ratio of posterior probabilities for model k over model 0 is

$$\begin{aligned} \frac{\pi(M = k|y)}{\pi(M = 0|y)} &= \frac{p(y|M = k)\pi_M(k)}{p(y|M = 0)\pi_M(0)} \\ &= B_{k,0}/K. \end{aligned}$$

Notice the down-weighting of the evidence by a factor K and the similarity to Bonferroni. We get this if we measure the evidence using the posterior ratio rather than the BF/marginal likelihood ratio. The BF corresponds to a uniform prior over models so in this example it expresses a prior probability $K/(K + 1)$ that the disease is associated with one of the genes, and this is close to one when K is large. When many models are compared against a baseline, and care is taken over the choice of π_M, the posterior probabilities themselves may make a better measure of the strength of the evidence than Bayes Factors.

1.4 Case study: Radiocarbon dating

The data come from the site of an ancient settlement in NZ. The researchers are very confident, before seeing the radiocarbon dates (RCD's), that there was no settlement on the site prior to $U = 1000$ years BP (Before the Present, where by convention the present is taken to be the year 1950), and that the settlement had been abandoned prior to $L = 500$ BP.

For how long was the site settled? There is a suggestion that this camp was settled for just weeks or months rather than years. This is a question about the occupation span.

As the archaeologists dug down they dug through a habitation layer. Above and below this layer there is no evidence for dwellings on the site. They obtained "uncalibrated" radiocarbon dates for $n = 7$ charcoal samples from the habitation layer.

```
#Date id - i - y_i - sigma_i
#NZ 7758 - 1 - 580 - 47
#NZ 7761 - 2 - 600 - 50
#NZ 7757 - 3 - 537 - 44
#NZ 7756 - 4 - 670 - 47
#NZ 7755 - 5 - 646 - 47
#WK 2589 - 6 - 630 - 35
#NZ 7771 - 7 - 660 - 46
```

Each date $y_i,\ i = 1, ..., n$ comes with an associated measurement uncertainty $\sigma_i,\ i = 1, ..., n$. These uncalibrated radiocarbon dates are *roughly speaking* dates in years before the present. However the observation model for these data is non-linear as we will see.

1.4.1 Observation model

An uncalibrated radiocarbon age $y_i \in \mathbb{R}$ is, for $i = 1, 2, ..., n$, a noisy biased measurement of the unknown true age $\Theta_i \in [L, U]$ of the i'th dated specimen. Here age $\theta_i,\ i = 1, ..., n$ is a quantity which increases into the past. The observation model for the data is

$$y_i = \mu(\theta_i) + \epsilon_i$$

with

$$\epsilon_i \sim N(0, \sigma_c(\theta_i)^2 + \sigma_i^2).$$

There are two sources of noise for observation i: measurement error σ_i (which is given as part of the data) and standard deviation $\sigma_c(\theta_i)^2$ in the calibration map μ which is to some extent uncertain. This latter variance depends on time, but the dependence is known. The likelihood for a single observation is

$$p(y_i|\theta_i) = \frac{\exp(-(y_i - \mu(\theta_i))^2/2(\sigma_c(\theta_i)^2 + \sigma_i^2))}{\sqrt{2\pi(\sigma_c(\theta_i)^2 + \sigma_i^2)}}.$$

This model has been derived by the radiocarbon-dating community through extensive experimentation and measurements and is reasonably reliable.

The functions $\mu(\theta)$ and $\sigma_c(\theta)$ are available from radiocarbon dating labs (Eg OxCal, a widely used Bayesian package for RCD) and are reported in a table with 5-year intervals (over this time-period). In our implementation we interpolated them to one-year intervals and treated these functions as piecewise constant within each year. The likelihood plot in Figure 1 shows $p(y_6|\theta_6)$ on the y-axis as a function of θ_6 on the x axis.

Let $y = (y_1, ..., y_n)$ and $\theta = (\theta_1, ...\theta_n)$. The overall likelihood is

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta_i).$$

1.4.2 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:

$$\pi_u(\theta) = (U - L)^{-n} \prod_{i=1}^{n} \mathbb{I}(L \leq \theta_i \leq U).$$

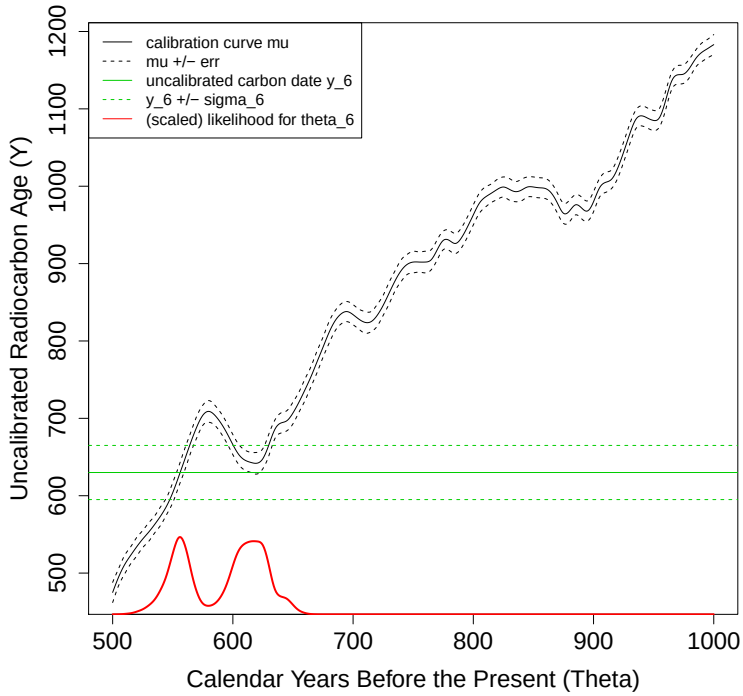


Figure 1: Calibration curve μ with likelihood $p(y_6|\theta_6)$ (in red) for sixth parameter. The solid green line is the observed value y_6 (with dashed lines at $y_6 \pm \sigma_6$). The solid black line gives the calibration curve $\mu(x)$ (with dashed lines at $\mu(x) \pm \sigma_c(x)$). The red curve, representing the likelihood, gives the probability density to realise y_6 by projecting any x-value through the calibration curve onto the y-axis, and adding noise with variance $\sigma_c(\theta_6)^2 + \sigma_6^2$.

The parameters space is $\Omega_u = [L, U]^n$.

Recall that the occupation span is of particular interest. We could estimate the start and end of settlement using $\min(\theta)$ and $\max(\theta)$ the most recent and earliest sample 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 $\theta_i, i = 1, 2, ..., n$ as generated by some "physical" process?

(B) the span is linked to a key scientific question - is the prior non-informative on this variable?

Does π_u meet these criteria? Certainly not (A). How do we answer (B)? The prior weights the span. To compute this weighting we need the marginal distribution $\pi_{S_u}(s_u)$ of the span S_u in the prior. It is very easy to get this numerically by simulating $\theta \sim \pi_u(\cdot)$ and making a histogram of $\max(\theta) - \min(\theta)$-values. We can calculate it.

Exercise 1.6. Show that the joint distribution of $\theta^- = \min(\theta),\ \theta^+ = \max(\theta)$ is

$$\pi_{u,\pm}(\theta^-, \theta^+) = \frac{n(n-1)}{(U-L)^n}(\theta^+ - \theta^-)^{n-2}.$$

ANS: informally, $\pi_u(\theta) = \pi(\theta \backslash \theta^{\pm} | \theta^{\pm}) \pi_{u,\pm}(\theta^{\pm})$. The LHS is $(U-L)^{-n}$ and $\pi(\theta \setminus \theta^{\pm} | \theta^{\pm}) \propto (\theta^+ - \theta^-)^{n-2}$ as $n-2$ parameters are uniformly distributed between max and min. Solve for $\pi_{u,\pm}$ and normalise.

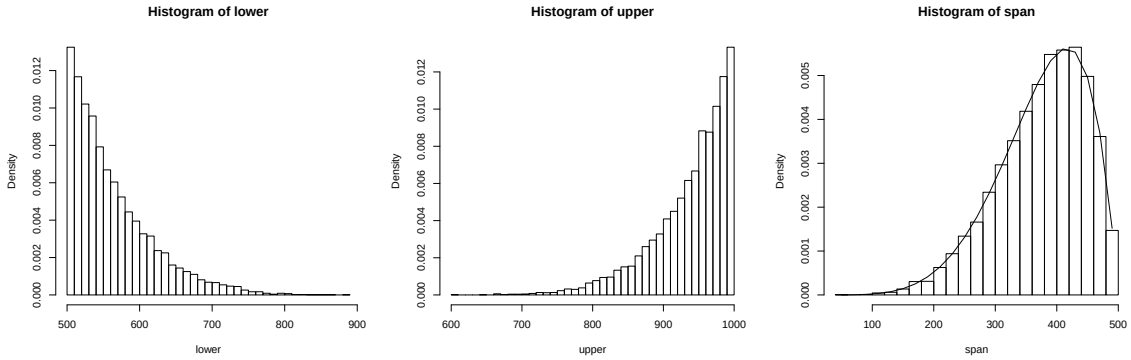


Figure 2: Marginal priors $\pi_u(\theta_{(1)})$ (left), $\pi_u(\theta_{(n)})$ (mid), $\pi_u(S_u)$ (right). We have no basis in fact for favoring large S_u values over smaller values.

Exercise 1.7. Make a change of variables from (θ^-, θ^+) to (θ^-, s_u) with $s_u = \theta^+ - \theta^-$ and Jacobian equal one and integrate over θ^- from L to $U - s_u$ to obtain

$$\pi_{S_u}(s_u) = \frac{n(n-1)}{(U-L)^n} s_u^{n-2}(U - L - s_u) \quad \text{for } 0 \le s_u \le U - L.$$

ANS: the Jacobian is equal one so $\pi_{\theta^-, s_u}(\theta^-, s_u) \propto s_u^{n-2}$. If s_u is fixed then the least value of θ^- is L and the largest value it takes is $U - s_u$, so the marginal $\pi_{S_u}(s_u) \propto \int_L^{U-s_u} s_u^{n-2} d\theta^-$ which doesnt depend on θ^- and gives $\pi_{S_u}(s_u)$ above.

We can check the distribution of S_u (and my calculation) using histograms (see Figure 2) from prior simulations. Let's interpret these graphs: in the uniform prior π_u, the span is clearly weighted towards larger values. If n is large then this effect will be strong, but it is clearly undesirable at any n. It is an unintended consequence of the choice of a uniform prior. The prior π_u was a bad choice as it does not represent prior knowledge.

This is a common problem with uniform priors. They weight by "Metric factors" - the volume of space and this grows exponentially with parameter dimension p. There are simply far more points θ in $[L, U]^p$ (with $p = n$ here) with the property that $\max(\theta) - \min(\theta)$ is large. For small s_u (away from the constraint that $s_u \le U - L$) the probability $\pi_{S_u}(s_u) ds_u$ grows like the volume of the shell of an $n-2$-dimensional sphere, so like s_u^{n-2}.

1.4.3 A prior from a process generating θ

Let us see what happens if we just think in a very simple way about the process that generated the data and introduce variables corresponding to the real world events which shaped the data.

Model assumptions

1. Settlement starts at time $\psi_2 < U$ and ends at ψ_1 with $L < \psi_1 < \psi_2$ (notice that the evolution forward in time is backward in age) so the span will be $S_s = \psi_2 - \psi_1$.
2. The probability that an interval of time (actually age) dt contains a dated specimen is λdt - that is, the dates θ are a realisation of a Poisson process with constant rate λ over the interval $[\psi_1, \psi_2]$.
3. any value *of the span* $0 \le S_s \le U - L$ is equally likely.

You may be skeptical about the uniform assumption in item 3 given that we criticised it for the previous prior π_u. However it is qualitatively different to take a uniform density in 1-D where weighting is much weaker. Also it expresses explicitly stated prior knowledge on a function of the parameters (ignorance of span - we are interested in learning the span, which is consistent with the idea that we are ignorant of its value within limits) and at the same time avoids biasing a sensitive statistic. Arguably the limits L and U are conservative, so ψ-values values close to the boundary could be penalised. We could, on discussion with the archaeologists who gave us those values of L and U, down-weight larger values of the span with something like a Beta(1,2) prior on $S_s/(U-L)$, still only weakly informative.

Consequences

1. We choose the number of dates n so we condition on n. We therefore have $\theta_i \sim U(\psi_1, \psi_2)$ (conditioning a Poisson process on the number of events gives a uniform distribution for the event times).
2. Our prior for $\theta|\psi$ is therefore

$$\pi_s(\theta|\psi) = \frac{1}{(\psi_2 - \psi_1)^n}\mathbb{I}(\psi_1 < \theta_1, ..., \theta_n < \psi_2).$$

We have to specify the prior for $\psi = (\psi_1, \psi_2)$. 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.

Exercise 1.8. Show that if $(\psi_1, \psi_2) \sim \pi_s(\psi)$ given above then the marginal distribution of the span $S_s = \psi_2 - \psi_1$ is $\psi_{S_s}(s) \propto \mathbb{I}_{0 \le s \le U-L}$.

Ans: make a change of variables from (ψ_1, ψ_2) to (ψ_1, s) in the density and integrate ψ_1 from L to $U - s$ similar to last problem.

Since $\pi_s(\theta, \psi) = \pi_s(\theta|\psi)\pi_s(\psi)$,

$$\pi_s(\theta, \psi) \propto \frac{1}{(\psi_2 - \psi_1)^n} \frac{1}{(U - L - (\psi_2 - \psi_1))}$$

models the prior information we actually have. The parameter space is

$$\Omega_s = \{(\theta, \psi) \in [L, U]^{n+2} : \psi_1 < \theta_i < \psi_2, i = 1, ..., n\}$$

We can carry out the same exercise as before, simulating the prior and checking the distribution of key summary statistics are representative of prior information we actually have. We simulate $\theta, \psi \sim \pi_s(\theta, \psi)$ and plot histograms of ψ_1, ψ_2 and $S_s = \psi_2 - \psi_1$ in Figure 3. These marginal priors on ψ_1, ψ_2 and S_s in Figure 3 better represent the prior information we have. The prior on span is uniform, as we would hope from the math, and the priors $\pi_s(\psi_1)$ and $\pi_s(\psi_2)$ distribute probability mass more evenly over the parameter domain.

1.4.4 Radiocarbon example continued ... Sampling the posterior

We have two posterior distributions,

$$\pi_u(\theta|y) \propto p(y|\theta)\pi_u(\theta)$$

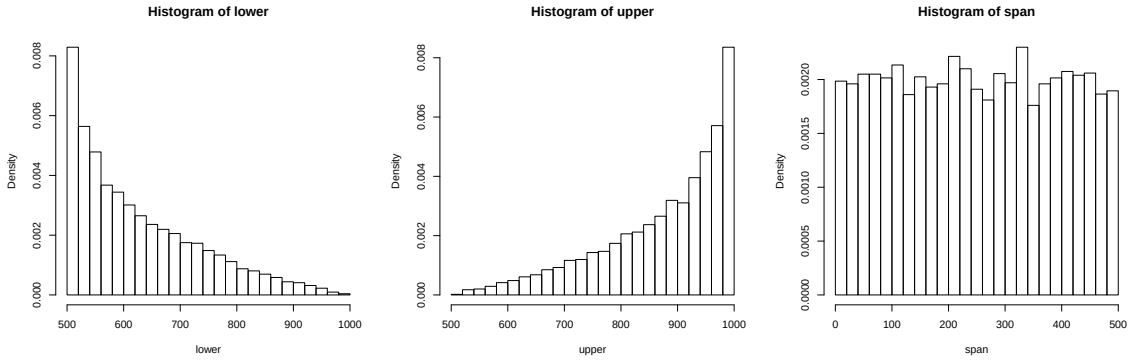


Figure 3: Marginal priors $\pi_s(\psi_1)$ (left), $\pi_s(\psi_2)$ (mid), $\pi_{S_s}(s)$ (right).

and

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

The two posterior distributions have the same likelihood, but different priors π_u and π_s, and parameters spaces Ω_u and Ω_s. We would like to summarise these distributions, form histograms and HPD credible sets.

I used simple random walk Metropolis Hastings MCMC updating one variable at a time. We review MCMC shortly. See online R-code for details. I sampled $\theta^{(t)} \sim \pi_u(\theta|y), t = 1, ..., T$, computed the order statistics $\theta^{(t)}_{(1)}, \theta^{(t)}_{(n)}$ and span $S^{(t)}_u = \theta^{(t)}_{(n)} - \theta^{(t)}_{(1)}$ for each MCMC sample and used them to plot posterior histograms and compute an HPD set for $S_u|y$. I also sampled $(\theta^{(t)}, \psi^{(t)}) \sim \pi_s(\theta, \psi|y), t = 1, ..., T$, in the second model, and computed the span $S^{(t)}_s = \psi^{(t)}_2 - \psi^{(t)}_1$ for each MCMC sample. I plotted histograms of these quantities and computed an HPD set for $S_s|y$.

1.4.5 Summarising the results

We are interested in the span, and comparing two models with and without shrinkage.

Model 1 (unif/π_u): HPD set [70,160] Model 2 (shrink/π_s): HPD set [0,160]

The marginal posterior histograms differ in shape and support.

A Bayes Factor compares models. This is a goodness of fit check on the prior, not testing a pre-defined scientific hypothesis. Under model one the marginal likelihood $p_u(y)$ say, is

$$p_u(y) = \int_{\Omega_u} p(y|\theta)\pi_u(\theta)d\theta,$$

(this is $p(y|m = 1)$ if π_u is model 1) and under model two (so $p(y|m = 2)$) we have

$$p_s(y) = \int_{\Omega_s} p(y|\theta)\pi_s(\theta, \psi)d\psi d\theta.$$

We estimate these using Monte Carlo and bridge sampling.[3] We find $\hat{p}_u \simeq 4 \times 10^{-21}$ and $\hat{p}_s \simeq 8 \times 10^{-19}$, so the Bayes factor $B_{s,u} = p_s(y)/p_u(y)$ for shrinking over uniform priors is about $\hat{B}_{s,u} \simeq 200$. The shrinkage prior π_s is clearly favoured.

[3]We will see how this is done in Chapter 2 towards the end.

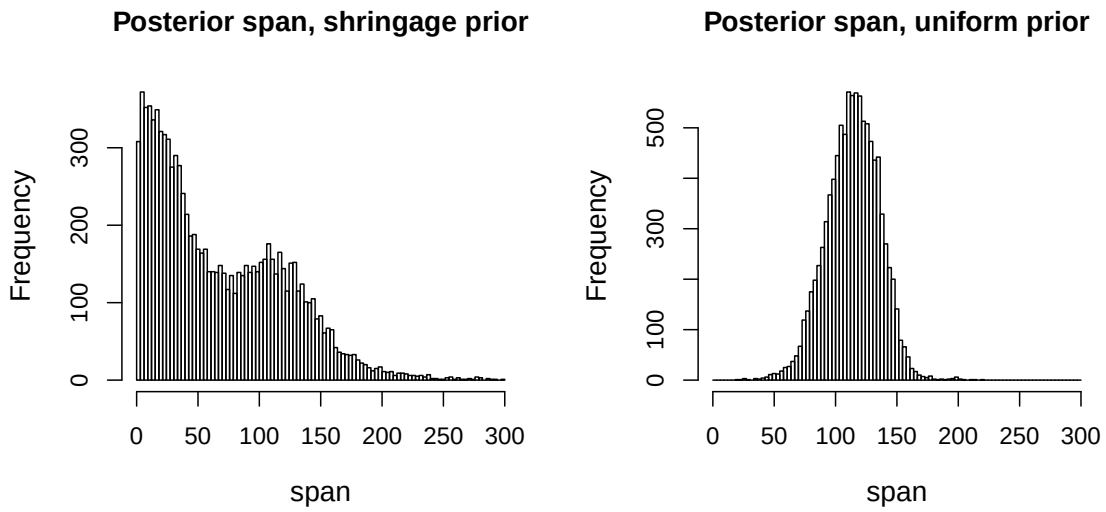


Figure 4: (Left) Posterior for the occupation span under model 2, the uniform span prior π_s. (right) posterior for the occupation span under model 1 the uniform age prior π_u.

Model 2 is overwhelmingly favored. We see from the HPD sets and posterior distributions under model 2 that a very short occupation span is plausible.

1.4.6 Conclusions

We conclude that a brief settlement time is not ruled out (see histogram of S_s) in our favored analysis. It is ruled out by an analysis using a prior which weights against brief settlement times.

We dont have much data (7 noisy numbers) so the conclusions show some sensitivity to the choice of prior. However the prior π_s better represented actual prior knowledge than π_u and was strongly favored by the data ($B_{s,u} \simeq 67$).

The process-model based prior π_s allows very small spans close to zero, while the uniform prior rules them out. This is clearly a case where we dont want the prior to impose structure we cant support on prior grounds.

2 An introduction to Markov chain Monte Carlo Methods

2.1 MCMC

2.1.1 Introduction

MCMC is a family of algorithms for simulating $X_0, X_1, X_2, ...$ so that $X_t \xrightarrow{D} p$ (X_t converges to p in distribution) for a user-defined probability distribution p. When we come to use this p will be π, the posterior, and the distribution of X_t will converge to the distribution of $\Theta|Y = y$. The sequence of samples in the chain will be (approximately) a sequence of correlated samples from the posterior.

MCMC methods are among of the most versatile classes of Monte Carlo algorithms we have, and are in routine use across statistics. It is striking that many papers developing deterministic approximation schemes, offered as scalable alternatives to MCMC, continue to use MCMC as a baseline to establish the true distribution in test cases where this cannot be simply computed. The fact that they are asymptotically exact is very appealing.

I will set out theory for the case that Ω, the space of states of X_t, $t = 0, 1, 2, ...$, 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 θ, $\pi(\theta)$ and $\pi(\theta|y)$ using finite precision arithmetic so we are really working with finite Ω anyway.

2.1.2 Markov chains

Let $\{X_t\}_{t=0}^{\infty}$ be a homogeneous Markov chain of random variables on Ω with starting distribution $X_0 \sim p^{(0)}$ and transition probability matrix $P = (P_{i,j})_{i,j\in\Omega}$ with

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

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 $p_i^{(n)} = \mathbb{P}(X_n = i)$, with $p^{(n)}$ a row vector.

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.

2.1.3 The Stationary Distribution and Detailed Balance

In discussing Markov chains we will work with a generic "target" distribution $p = (p_i)_{i\in\Omega}$ (taken as a row vector in the discrete setting). 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 $p(\theta) = \pi(\theta|y)$.

I assume familiarity with classification of Markov chains on a countable space. See the book by James Norris, or the Part A Probability lecture notes if you would like more detail. The following

is primarily notation-setting. The probability mass function (PMF) $p_i, i \in \Omega$, $\sum_{i\in\Omega} p_i = 1$ is a *stationary distribution* of P if $pP = p$ in the following sense. If $p^{(0)} = p$ then

$$p_j^{(1)} = \sum_{i\in\Omega} p_i^{(0)} P_{i,j},$$

so $p_j^{(1)} = p_j$ also. Iterating, $p^{(t)} = p$ for each $t = 1, 2, ...$ in the chain, so the distribution of $X_t \sim p^{(t)}$ doesn't change with t, it is stationary.

We have a given target distribution p and a transition matrix P and want to check that $X_t \stackrel{D}{\to} p$. The convergence theorem for finite irreducible Markov chains tells us that if Ω is finite, if $pP = p$, and if P is irreducible and aperiodic, then indeed $X_t \stackrel{D}{\to} p$. Checking $pP = p$ is hard, as we have to sum over all Ω to evaluate pP.

Definition 2.1. *Detailed balance (discrete case)*: If there is a probability mass function $p_i,\ i \in \Omega$ satisfying $\sum_{i\in\Omega} p_i = 1$ and

$$p_i P_{i,j} = p_j P_{j,i} \quad \text{holds for all } i, j \in \Omega,$$

then P and p satisfy detailed balance.

Exercise 2.2. Show that if p and P satisfy detailed balance then p is stationary for P.
ANS: sum both sides of DB over $i \in \Omega$ and use $\sum_i P_{j,i} = 1$ to establish $p_j = [pP]_j,\ j \in \Omega$.

Detailed balance is sufficient for stationarity, and it is much easier to check than $pP = p$ as it is a simple algebraic relation. A Markov chain satisfying DB is *reversible*.

2.1.4 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, ie $X_t \stackrel{D}{\to} p$, then $X_t \sim p^{(t)}$ with $p^{(t)} \simeq p$ at large t, so when we look at our Markov chain $X_0, X_1, ..., X_T$, "most" of the samples are "nearly" distributed according to p.

Let $f : \Omega \to R$. Let $\hat{f}_T = T^{-1} \sum_t f(X_t)$ estimate $E_{X\sim p}(f(X))$. If we form an average over states in the chain then we might expect it to converge to an expectation over the target, as the random variables we are averaging are converging.

Theorem 2.3. *(Ergodic Theorem) If $\{X_t\}_{t=0}^{\infty}$ is an irreducible and aperiodic Markov chain on a finite space of states Ω satisfying detailed balance with respect to the probability distribution p, then as $T \to \infty$*

$$\hat{f}_T \stackrel{a.s.}{\to} E_{X\sim p}(f(X))$$

for any bounded function $f : \Omega \to R$. The convergence is almost surely (a.s.). [For proof see eg Norris Markov Chains, CUP, (1997)]. Such a chain is ergodic with target distribution $p = (p_i)_{i\in\Omega}$.

The more general statement covering continuous target distributions 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 (not just countable).

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*]

2.1.5 The Metropolis-Hastings Algorithm

Suppose we need samples from a pmf $p_i, i \in \Omega$ with Ω a finite set (for example we may wish to form Monte-Carlo summaries of the kind described in Section 1.3.4). We give an algorithm simulating X_{t+1} given X_t. The algorithm determines the transition probabilities $P(X_{t+1} = j|X_t = i)$ and hence the transition matrix P. The algorithm is constructed to ensure the chain targets p.

We simulate a random walk $X_0, X_1, X_2, ...$ in Ω by accepting or rejecting proposals from a simple irreducible transition matrix $Q_{i,j}, i, j \in \Omega$ which we get to choose. There is then a rejection step according to a rule which "corrects" proposals drawn from Q to get a new effective transition matrix P. We will see that this transition matrix satisfies detailed balance for p. If the Markov Chain is irreducible and aperiodic then we have satisfied the ergodic theorem and have a chain targeting p.

Definition 2.4. *Metropolis Hastings MCMC*: the following algorithm simulates a Markov chain. Let $q(j|i) = Q_{i,j}$ be a *proposal probability distribution* with transition probability Q satisfying

$$q(j|i) > 0 \Leftrightarrow q(i|j) > 0.$$

Let $X_t = i$. The next state X_{t+1} is realised in the following way.

1. Draw $j \sim q(\cdot|i)$ and $u \sim U[0, 1]$.

2. If $u \leq \alpha(j|i)$ where
$$\alpha(j|i) = \min\left\{1, \frac{p_j\, q(i|j)}{p_i\, q(j|i)}\right\}$$
then set $X_{t+1} = j$, otherwise set $X_{t+1} = i$.

We initialise this with some $X_0 = i_0$ satisfying $p_{i_0} > 0$ and iterate for $t = 1, 2, 3, ...T$ to simulate the samples we need. [End of Definition]

Lemma 2.5. *If the Markov chain realised by the algorithm given in Definition 2.4 is irreducible and aperiodic then it is ergodic with target p.*

Proof: We assume the chain is irreducible and aperiodic - this has to be checked separately for each MH MCMC algorithm and depends on our choice of $q(j|i)$ and the acceptance probabilities $\alpha(j|i),\ i, j \in \Omega$. Since we are assuming Ω is finite, it is then sufficient to show that the Markov chain determined by the random MCMC update has p as a stationary distribution, so compute the transition matrix P and verify detailed balance,

$$P_{i,j}\ p_i = P_{j,i}\ p_j.$$

Detailed balance for $i = j$ is trivial so 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 j 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:

$$\begin{aligned} p_i P_{i,j} &= p_i\, q(j|i)\alpha(j|i) \\ &= p_i\, q(j|i) \min\left\{1, \frac{p_j\, q(i|j)}{p_i\, q(j|i)}\right\} \\ &= \min\{p_i q(j|i), p_j q(i|j)\} \\ &= p_j\, q(i|j) \min\left\{\frac{p_i\, q(j|i)}{p_j\, q(i|j)}, 1)\right\} \\ &= p_j\, q(i|j)\alpha(i|j) \\ &= p_j P_{j,i} \end{aligned}$$

and we are done. [End of Proof]

2.1.6 Example: Simulating the hypergeometric distribution

The hypergeometric distribution 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 for $k \in \Omega$,

$$\Omega = \{k \in \mathbb{Z} : \max\{0, n + K - N\} \leq k \leq \min\{n, K\}\}$$

Let $B^- = \max\{0, n + K - N\}$ and $B^+ = \min\{n, K\}$. The PMF is

$$p_k = \binom{K}{k}\binom{N-K}{n-k} / \binom{N}{n} \quad B^- \leq k \leq B^+.$$

Give a MH MCMC algorithm ergodic for p.

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 we can leave the state space $\Omega = \{B^-, B^- + 1, ..., B+\}$ given above. If $i = B^+$ and we propose $j = i + 1$ then j has 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 $j \notin \Omega$.

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

$$\begin{aligned} u &\leq \min\left\{1, \frac{p_j\, q(i|j)}{p_i\, q(j|i)}\right\} \\ &= \min\left\{1, \frac{\binom{K}{j}\binom{N-K}{n-j}}{\binom{K}{i}\binom{N-K}{n-i}}\right\} \end{aligned}$$

then set $X_{t+1} = j$, else (if either condition fails) set $X_{t+1} = i$.

Notice that if we propose $j < B^-$ or $j > B^+$ then we reject. This is the same as taking $p_j = 0$ for these states so $\alpha = 0$ and we would reject and stay in Ω for any u from the first step.

Step 3: check chain is irreducible and aperiodic. This can be seen since q allows us to visit any state in Ω and $\alpha(j|i)$ is never zero for any pair $i, i+1 \in \Omega$ so $P_{i,i+1} > 0$ and $P_{i-1,i} > 0$ for such states. It is aperiodic because it is irreducible and can reject, so $P_{i,i} > 0$ for some $i \in \Omega$. We dont check aperiodicity unless perhaps the acceptance probability is always one. The algorithm is implemented in the online code. Output is illustrated in Figure 5.

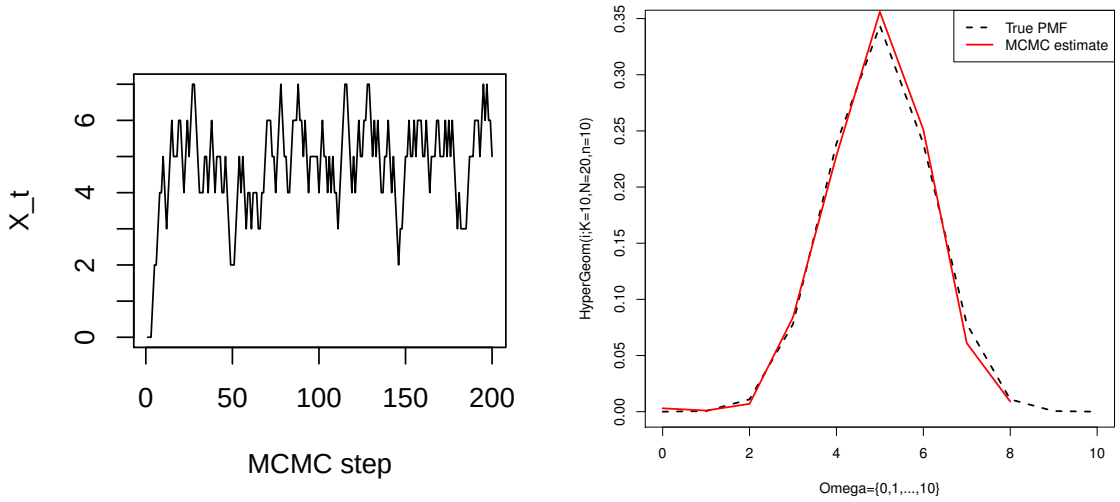


Figure 5: Summary output of the R-implementation given in the text. Left: x-axis is step counter $t = 1, 2, 3...200$. The y-axis is Markov chain state X_t targeting HyperGeom$(K = 10, N = 20, n = 10)$. Right: histogram of $X_1, X_2, ..., X_n$ for $T = 1000$.

2.1.7 Notation for the continuous case

In this section we give notation for the continuous case and write down detailed balance. We give the extension (essentially always used in practice) to the case where we choose a MH-update at random from a set of candidates.

When the random variable $\theta \in \Omega$ is continuous with density $p(\theta)$ then the distribution $\mathbb{P}(d\theta) = p(\theta)d\theta$ is defined for sets $\mathcal{B}_\Omega$ so that $P : \mathcal{B}_\Omega \to [0, 1]$ and the probability space is $(\Omega, \mathcal{B}, \mathbb{P})$. In a simple setup we have a proposal density $q(\theta'|\theta)$ and an acceptance probability

$$\alpha(\theta'|\theta) = \min\left\{1, \frac{p(\theta')q(\theta|\theta')}{p(\theta')q(\theta'|\theta)}\right\}. \tag{2.1}$$

With these substitutions, the Metropolis-Hastings algorithm is otherwise unchanged. Write $q(d\theta'|\theta)$ for proposal *distribution* so that $q(A|\theta) = \int_A q(d\theta'|\theta)$. Let

$$c(\theta) = 1 - \int_\Omega \alpha(\theta'|\theta)q(d\theta'|\theta)$$

give the probability for a proposal $\theta' \sim q(\cdot|\theta)$ to be rejected when the acceptance probability in our MCMC is $\alpha(\theta'|\theta)$.

The transition probability $P_{i,j}$ becomes a transition kernel (a conditional probability distribution) which we denote $K(\theta, d\theta')$ with $K(\theta, A) = \mathbb{P}(X_{t+1} \in A | X_t = \theta)$ for $A \in \mathcal{B}_\Omega$. This gives the probability the chain is in A at the next step given it is at θ at the current step.

Proposition 2.6. *The transition kernel for Metropolis-Hasting MCMC is*

$$K(\theta, d\theta') = \alpha(\theta'|\theta) q(d\theta'|\theta) + c(\theta)\delta_\theta(d\theta'), \tag{2.2}$$

where $\delta_\theta(d\theta')$ is the Dirac delta-function, so that $\int_A \delta_\theta(d\theta') = \mathbb{I}_{\theta' \in A}$.

Proof: The update $\theta \to \theta$ occurs if any θ' is proposed and then rejected. Suppose $\theta \neq \theta'$. The update $\theta \to \theta'$ occurs iff θ' is proposed and then accepted. Partitioning on the event that the update is accepted, we have, for $A \in \mathcal{B}_\Omega$,

$$\begin{aligned} \Pr(X_{t+1} \in A | X_t = \theta) &= c(\theta)\mathbb{I}_{\theta \in A} \; + \; \int_A \alpha(\theta'|\theta) q(d\theta'|\theta) \\ &= \int_A K(\theta, d\theta'), \end{aligned}$$

where the second line follows by substituting the proposed form for K given in Eqn. 2.2 and verifying it matches the first line. Since we have shown the proposed kernel matches $K(\theta, A) = \Pr(X_{t+1} \in A | X_t = \theta)$ for every $A \in \mathcal{B}_\Omega$ it follows that it is the transition distribution. [EOP]

Remark 2.7. The term involving $c(\theta)$ is the probability for rejection and ensures that when we integrate $K(\theta, d\theta')$ over Ω we get one. We previously only considered the case where $\theta' \neq \theta$.

Definition 2.8. The MCMC transition kernel satisfies detailed balance with respect to the target distribution $p(d\theta)$ if

$$p(d\theta')K(\theta', d\theta) = p(d\theta)K(\theta, d\theta').$$

Explicitly, for $\theta \in A$ and $\theta' \in B$,

$$\int_B p(d\theta') \int_A K(\theta', d\theta) = \int_A p(d\theta) \int_B K(\theta, d\theta'). \tag{2.3}$$

must hold for every pair of sets $A, B \in \mathcal{B}_\Omega$.

Remark 2.9. If this holds then the process is stationary. Take $A = \Omega$ to obtain

$$p(B) = \int_\Omega K(\theta, B) p(d\theta), \quad \text{for all } B \in \mathcal{B}.$$

Proposition 2.10. *Detailed balance holds for K and p in the MH algorithm if and only if*

$$\int_B p(d\theta') \int_A q(d\theta|\theta')\alpha(\theta|\theta') = \int_A p(d\theta) \int_B q(d\theta'|\theta)\alpha(\theta'|\theta). \tag{2.4}$$

for every pair of sets $A, B \in \mathcal{B}_\Omega$.

Proof: Substitute Equation 2.2 into Equation 2.3. We get Equation 2.4 because the terms involving c will cancel: on the RHS we get

$$\begin{aligned} \int_A p(d\theta) \int_B c(\theta)\delta_\theta(d\theta') &= \int_A p(d\theta)c(\theta)\mathbb{I}_{\theta \in B}, \\ &= \int_{A \cap B} p(d\theta)c(\theta) \end{aligned}$$

from this term and on the LHS we get $\int_{B \cap A} p(d\theta')c(\theta')$ on the LHS. These are equal, so the contribution cancels and we are left with Equation 2.4. We can reverse this. [EOP]

Remark 2.11. In terms of probability densities, detailed balance is then

$$p(\theta')q(\theta|\theta')\alpha(\theta|\theta') = p(\theta)q(\theta'|\theta)\alpha(\theta'|\theta),$$

and this may readily be verified for the acceptance probability in Eqn. 2.1

2.1.8 Mixing updates for continuous multivariate targets

We conclude this development of the theory by remarking on a useful generalisation to multiple proposals. When the parameter $\theta = (\theta_1, ..., \theta_p)$ has more than one dimension, it is often desirable to update different components using different proposal distributions. Suppose we have N different proposal densities $q_k(\theta'|\theta)$ and we select the k'th one to make a proposal with probability ξ_k and then accept or reject. The algorithm is given below.

Definition 2.12. *Metropolis Hastings MCMC (continuous mixture)*: the following algorithm simulates a Markov chain targeting p. For $k = 1, ..., N$ let $q_k(\theta'|\theta)$ be a *proposal probability density* satisfying

$$q_k(\theta'|\theta) > 0 \Leftrightarrow q_k(\theta|\theta') > 0.$$

Let $X_t = \theta$. The next state X_{t+1} is realised in the following way.

1. Draw $k \sim \text{Multinon}(\xi_1, ..., \xi_N)$ and $\theta' \sim q_k(\cdot|\theta)$ and $u \sim U[0, 1]$.

2. If $u \leq \alpha_k(\theta'|\theta)$ where
$$\alpha_k(\theta'|\theta) = \min\left\{1, \frac{p(\theta')\, q_k(\theta|\theta')}{p(\theta)\, q_k(\theta'|\theta)}\right\}$$
then set $X_{t+1} = \theta'$, otherwise set $X_{t+1} = \theta$.

We initialise this with some $X_0 = \theta_0$ satisfying $p(\theta_0) > 0$ and iterate for $t = 1, 2, 3, ...T$ to simulate the samples we need. [End of Definition]

This targets p. We are choosing a kernel at random from $K_1, ..., K_N$ and using it to update the state. The individual kernels are, from Eqn. 2.2,

$$K_k(\theta, d\theta') = \alpha_k(\theta'|\theta) q_k(d\theta'|\theta) + c_k(\theta)\delta_\theta(d\theta'),$$

with $c_k(\theta) = `1 - \int_\Omega \alpha_k(\theta'|\theta) q_k(d\theta'|\theta)$. The overall kernel is

$$K(\theta, d\theta') = \sum_{k=1}^{N} \xi_k K_k(\theta, d\theta').$$

Exercise 2.13. Show that the overall transition kernel K satisfies detailed balance WRT p.

ANS: If K_k satisfies DB for $k = 1, ..., N$ then this linear combination does: substituting $K(\theta, d\theta')$ above into the expression for DB in Definition 2.8, the terms on the left and right sides cancel in pairs. However, the K_k do each satisfy DB WRT p as they are MH kernels.

If $\theta = (\theta_1, ..., \theta_p)$ and we target $p(\theta)$, $\theta \in \Omega$, we commonly update one variable at a time. If we have p components, we can take $N = p$ update types. If we choose to update component k then we propose $\theta'_{-k} = \theta_{-k}$ and simulate $\theta'_k \sim q_k(\cdot|\theta)$. Smaller changes to the state will typically have a higher acceptance probability (if the target $p(\theta)$ is a smooth function of θ, then $p(\theta) \simeq p(\theta)$ when $|\theta - \theta'|$ is small, so the acceptance probability will typically be closer to one). The chain takes more steps to explore the support of the target density in Ω if the change to the state is always small, so there is a trade off.

There are some subtleties here. In the above setup, for $\theta \neq \theta'$, the kernel $K(\theta, d\theta') = 0$ unless the two states θ and θ' differ at a single component (i say) and the density for the proposal, $q_i(\theta'_i|\theta) > 0$ is non zero. We often add many different proposals (for example, we might cycle through each

component as above, and then additionally have a transition that updates all components at once, or randomly chosen subsets etc). So there may be many proposals connecting two states θ, θ'. The mixture argument above shows that it is only necessary for each proposal, separately, to satisfy detailed balance. This kind of thing is discussed in the Advanced Simulation course.

2.1.9 MH example: an equal mixture of bivariate normals

As an example of MCMC targeting a density, consider a mixture of two bivariate normals, with target density defined in $\mathbb{R}^2$,

$$\pi(\theta) = (2\pi)^{-1}\left(0.5e^{-(\theta-\mu_1)\Sigma_1^{-1}(\theta-\mu_1)/2} + 0.5e^{-(\theta-\mu_2)\Sigma_2^{-1}(\theta-\mu_2)/2}\right)$$

and $\theta = (\theta_1, \theta_2)$. Take $\mu_1 = (1,1)^T$, $\mu_2 = (4,4)^T$ and $\Sigma_1 = \Sigma_2 = I_2$ for illustration.

Step 1. For a proposal distribution q we want something simple to sample and easy to evaluate. A simple choice that "will do" is

$$\theta_i' \sim U(\theta_i - a, \theta_i + a),$$

with $a > 0$ a fixed constant. We jump uniformly in a box of side $2a$. This is easy to sample, and we can evaluate the conditional density

$$q(\theta'|\theta) = q(\theta|\theta') = 1/4a^2$$

up to a constant (in fact, including the constant).

Step 2. The algorithm: given $\theta^{(n)} = (\theta_1, \theta_2)$,

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 $\theta^{(n+1)} = \theta$.

Step 3. This algorithm is ergodic for any $a > 0$ (clearly irreducible in the computer measure) but we will see that the choice of a makes a difference to efficiency.

Here is an implementation.

```
#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)
```

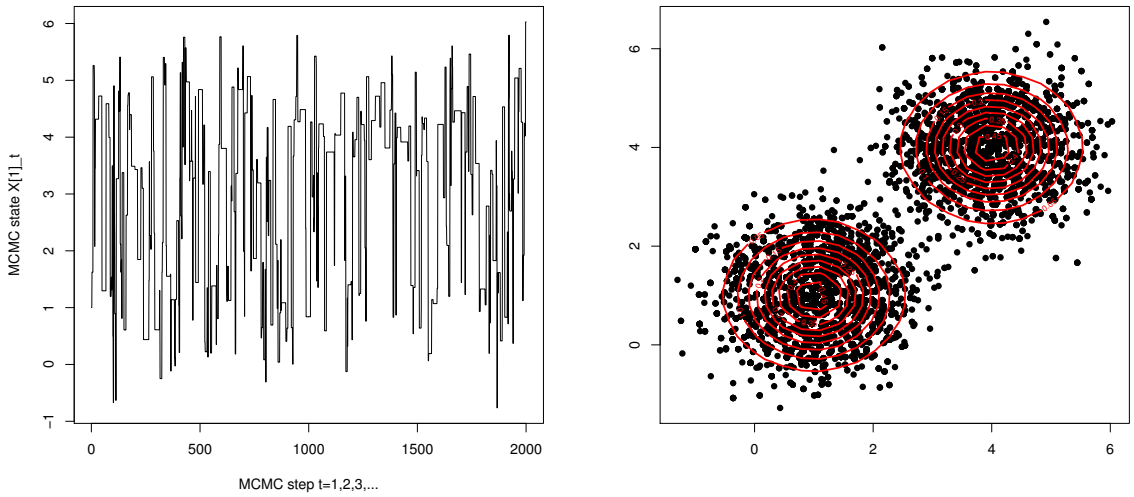


Figure 6: MCMC targeting a mixture of bivariate normals: (Left) MCMC trace of $\theta_1^{(t)}$ plotted against t for $t = 1, ..., 2000$; (Right) scatter plot of sampled parameter vectors $(\theta_1^{(t)}, \theta_2^{(t)})$, $t = 1, ..., 2000$.

```
   if (runif(1)<MHR) x<-y
   X[,t+1]<-x
}
```

Try this implementing this yourself as an exercise, experimenting with different values of the "jumpsize" $a > 0$. The key range of values is $a \simeq 3$, so slightly smaller or larger than the separation between modes.

Figure 6 shows some sample output. If $X_t = \theta^{(t)}$, $t = 1, ..., T$ with $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$ then the plot on the left shows $\theta_1^{(t)}$ plotted against t for $t = 1, ..., 2000$. The plot on the right in Figure 6 is a scatter plot of the points $(\theta_1^{(t)}, \theta_2^{(t)})$, $t = 1, ..., 2000$. If the jump size a is chosen to be too small then the chain cant move easily between modes as the path between modes must include a pair of states $\theta^{(t)}, \theta^{(t+1)}$ with $\pi(\theta^{(t)}) \ll \pi(\theta^{(t+1)})$ (as the state moves across the saddle between modes). The acceptance probability for the proposal $\theta^{(t+1)}$ will be small and so the sequence of acceptance events in a path crossing the saddle will be a rare sequence. The chain is still ergodic, but as we will see the sampler in inefficient. If a is chosen to be too big (like $a = 100$ say) then the sampler can cross from one mode to another in a single step, but it will also make many proposals into very low density states in the tail of the density, which will be rejected, so again this will be inefficient in a sense we define bellow.

2.2 Output analysis

See CJ Geyer "Practical Markov Chain Monte Carlo" Statistical Science 1992 and Sokal Cargèse summer school lecture notes "Monte Carlo Methods in Statistical Mechanics" 1996 for much useful insight.

An ergodic MCMC algorithm gives us converging estimates of expectations in the target. However, we have in general little idea of how biased any estimate formed at a fixed finite run length might

be. If the chain is not initialised with a sample from the target then there will be an initialisation bias. We cant in general sample the target exactly (that's why we are doing MCMC) so in general we wont have such initialisation. The bias can be very large if we stop the run before it has reached an entire mode or any other region of substantial probability mass in the target.

We have ultimately (in general) no way of knowing if the MCMC has converged and the samples we have are representative of the target distribution as a whole. We can only make consistency checks (ie, look for evidence that the chain hasnt converged) so we can only check necessary conditions for convergence, not sufficient conditions. Outside "perfect sampling" algorithms such as rejection sampling, which are not generally applicable in practice, this kind of problem always appears. For example, in importance sampling and particle filters, which are unbiased, one has the matching problem of rare high weight states.

Some of the most effective checks are extremely obvious and apply in all Monte-Carlo settings: run the simulation several times from different start states and check you get the same answer (to the precision desired). Plot the sequence of samples and look for trends.

2.2.1 Convergence and mixing

We want to estimate $E_{X\sim p}(f(X))$ using our MCMC samples

$$X_0, X_1, X_2, ..., X_n$$

targeting $p(x)$ and calculate the estimate $\bar{f}_n = n^{-1}\sum_t f(X_t)$. 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, bias and variance, respectively "convergence" and "mixing".

First, we dont start the chain in equilibrium. Samples from the first part of the chain are biased by initialization. We drop the first part of the MCMC run (called "burn-in") to reduce the initialization bias. We know $p^{(t)} \to p$ as $t \to \infty$ so choose a cut-off T such that $p^{(t)} \simeq p$ for $t \geq T$ is a good approximation. Take $n \gg T$ so that retained samples are representative of p.

If $n \gg T$ then the bias in $\bar{f}_n$ due to burn-in will be slight. If you need to drop alot of states from the start of the chain to reduce this bias, you may not have run the chain long enough anyway.

Second, suppose $p^{(0)}(x) = p(x)$, so we start the chain in equilibrium and we can forget about initialisation bias. The variance, $\text{var}(\bar{f}_n)$, of $\bar{f}_n$ will decrease 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.

Figure 7 shows output from two MCMC runs for the normal mixture with jump size $a = 2, 4$. We can see from the ACF plots that serial correlation between states in the run at $a = 4$ falls off more rapidly than when we take $a = 2$. When we take $a = 2$ we see the chain gets "stuck" in one mode for many updates before jumping to the other.

2.2.2 MCMC variance in equilibrium

We now give an approximation to $\text{var}(\bar{f}_n)$. Suppose any burn-in samples have been dropped so we have samples which we consider to be representative of the target. The MCMC output samples

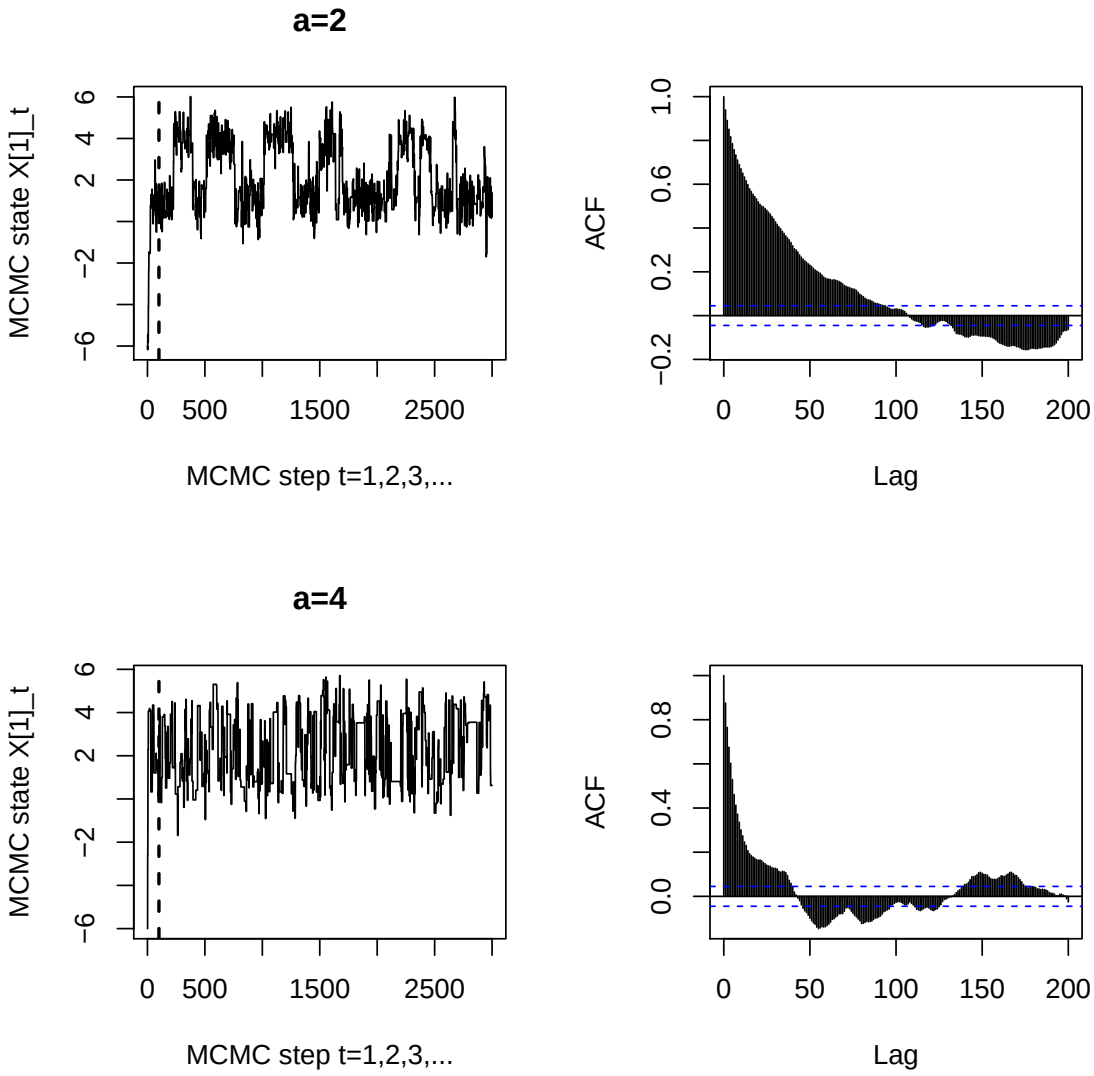


Figure 7: MCMC targeting the normal mixture in Section 2.1.9: (Left column) run traces for $\theta_1^{(t)}$, $t = 1, ..., 3000$ taking the jump size $a = 2$ (top) and $a = 4$ (bottom); (Right column) autocorrelation plots for the MCMC output sequences at left. Dotted lines in the ACF plots show asymptotic standard deviation estimates for the ACF estimates.

$X_1 = \theta^{(1)}, X_2 = \theta^{(2)}, ...$ targeting p are correlated so we define the Effective Sample Size (ESS) so that

$$\text{var}(\bar{f}_n) = \frac{\text{var}(f(X))}{ESS}$$

with $X \sim p$. The ESS is the number of independent samples which would give the same variance reduction as our n correlated samples. Typically $ESS \ll n$. If the MCMC samples were independent we would have $ESS = n$.

We begin by giving a straightforward but inefficient way to estimate $\text{var}(\overline{f}_n)$ and the ESS in order to show what is happening. Let $\sigma_f^2 = \text{var}(f(X))$ with estimate $\hat{\sigma}_f^2$. Let $\sigma_{f,n}^2 = \text{var}(\bar{f}_n)$ so that $ESS = \sigma_f^2/\hat{\sigma}_{f,n}^2$. We can simply make K runs each of length n, realising K sets of samples $\theta^{(k,t)},\ k = 1, ..., K,\ t = 1, ..., n$. These samples are correlated within a run but independent across runs and can be used to estimate $\text{var}(\bar{f}_n)$. We estimate $\overline{f}_{k,n} = n^{-1}\sum_t f(\theta^{(k,t)})$ as the average from each run, and then estimate $\text{var}(\overline{f}_n) \simeq \hat{\sigma}_{f,n}^2$ with

$$\hat{\sigma}_{f,n}^2 = \frac{1}{K-1}\sum_{k=1}^{K}\left(\overline{f}_{k,n} - K^{-1}\sum_{j=1}^{K}\overline{f}_{j,n}\right)^2,$$

from the standard error across independent runs. Now $\hat{\sigma}_{f,n}^2$ is computed from K runs each of length n and is an estimate of the variance of $\overline{f}_n$ if $\overline{f}_n$ is computed from a single run of length n. Now $\hat{\sigma}_{f,n}^2 \simeq \hat{\sigma}_f^2/ESS$ so the ESS for a single run of length n is approximately

$$ESS \simeq \hat{\sigma}_f^2/\hat{\sigma}_{f,n}^2.$$

The ESS is a measure of the precision gain afforded by our n correlated samples.

If we have two MCMC algorithms which we want to compare, with perhaps different proposal distributions, it is natural to prefer the one yielding the larger ESS at fixed run length n (so ESS/n is "effective independent samples per MCMC sample"). This is a measure of statistical efficiency (variance reduction per sample). In practice we often find that computationally expensive methods with a high statistical efficiency per sample are slow to compute, and a less statistically efficient MCMC algorithm may actually return a larger effective sample size in a given wall-clock time. For this reason we often make comparisons between Monte-Carlo estimation methods based on computational efficiency. If S is the time in CPU seconds to make n steps of the MCMC, we report ESS/S, the effective independent samples per CPU second. If the implementations and hardware are comparable then this is typically a more useful basis for comparison.

The simple approach to estimating $\text{var}(\bar{f}_n)$ given above wouldnt be sensible, as we incur a burn-in cost for each run. This leads to alot of discard samples across multiple runs and is inefficient. A method that estimates the variance of an estimate from a single long run is preferred. An approach *like* the one above but based on dividing a single long run into blocks is often used. This is sometimes called "binning". It assumes the blocks are long so that most of the samples in a block are effectively independent of the samples in other blocks. Here is an alternative approach.

Suppose the chain was initialised in the target distribution or otherwise has reached equilibrium, so we dropped burn-in (for this single run) as before. For $s, t \geq 0$ define the *correlation of f at lag s* to be

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

(so $\rho_0 = 1$ and we immediately drop the f but recall this is all just for a single function). Let $\sigma^2 = \text{var}(f(X_t)),\ X_t \sim p$. This doesn't depend on t because the chain is stationary. Express

$\text{var}(\bar{f}_n)$ in terms of ρ_s. This gives insight and leads to an estimator for $\text{var}(\bar{f}_n)$, since we can estimate ρ_s from the MCMC samples. We have,

$$\begin{aligned}
\text{var}(\overline{f}_n) &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}(f(X_i), f(X_j)) \\
&= \sigma^2 n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{|i-j|} \\
&= \sigma^2 n^{-1}\left[1 + 2\sum_{s=1}^{n-1}\left(1-\frac{s}{n}\right)\rho_s\right] \\
&\simeq \sigma^2 n^{-1}\left[1 + 2\sum_{s=1}^{n-1}\rho_s\right] \\
&= \sigma^2\tau_f/n,
\end{aligned}$$

where the approximation is good if $\rho_s \to 0$ sufficiently rapidly at large s. Here τ_f is the *Integrated Autocorrelation Time* (IACT) and

$$ESS = n/\tau_f$$

is the Effective Sample Size - the number of independent samples giving the same precision for $\overline{f}_n$ as the n correlated samples we have. This second definition of the ESS is only approximately equal to the last, due to the approximation made in the derivation.

We can estimate[4] $\gamma_s = \text{cov}(f(X_i), f(X_{i+s}))$ 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 $\hat{\rho}_s = \hat{\gamma}_s/\hat{\gamma}_0$. This leads to an estimate of τ_f,

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

with M a cut-off on the sum.

This cut-off is needed as $\hat{\rho}_s$ goes to zero with s and is dominated by estimation noise at large s. If we added terms at large s where ρ_s is very close to zero we are in effect just adding noise to our estimate, so we truncate the sum over s at $s = M$. The resulting estimate for $\hat{\tau}_f$ is consistent if M is chosen according to suitable criteria. Geyer (cited above) shows that for a Markov Chain, $\rho_t + \rho_{t+1}$ is positive, monotone and convex, so we might hope to use a violation of these conditions as evidence that noise is dominating signal. We can for example choose M equal to the least s such that $\rho_s + \rho_{s+1} > 0$ and $\hat{\rho}_s + \hat{\rho}_{s+1} > \hat{\rho}_{s-1} + \hat{\rho}_s$. As n grows, the variance of $\hat{\rho}_s$ decreases, and it converges to a positive monotone function of t, so these conditions are violated at increasingly large values of t (random, but this is the trend) and so the truncation bound grows with n. Geyer shows that this leads to a consistent estimate for τ_f and hence the ESS.

[4] See Priestly 1981 "Spectral Analysis and Timeseries" Academic London, pp323 for the factor $1/n$ where we expect $1/(n-s)$.

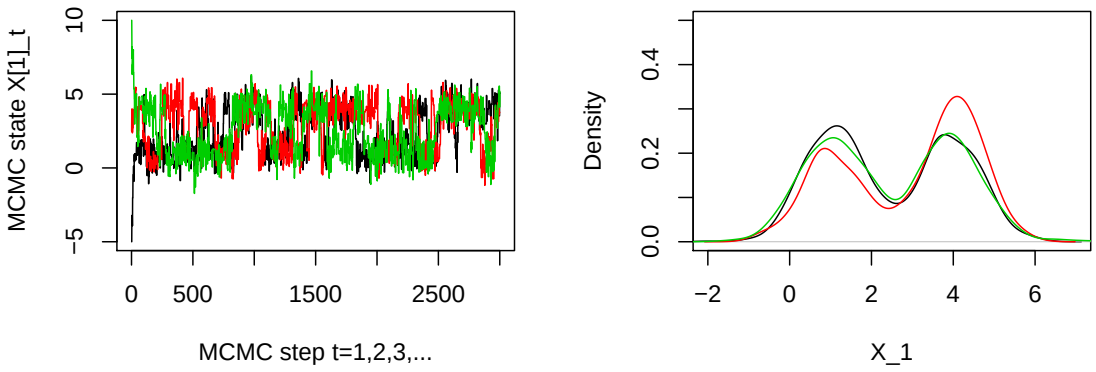


Figure 8: Two diagnostic plots for convergence checking: (Left) MCMC traces for selected statistics - in this case one of the components of the parameter vector - the log-likelihood is often a good choice; (Right) Histograms - or density plots - of a parameter of interest, which should be near identical. In this example mixing is poor. The plot at left shows the state is moving only slowly over its range and the densities at right would in general differ too greatly to be acceptable.

2.2.3 MCMC convergence

There is no simple generic sufficient condition we can test for convergence. Here some checks we should 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.

Figure 8 shows an example of some of the plots I would use for convergence checking on the normal mixture MCMC sampler. I have not included an auto-correlation plot which would be a natural addition though in this example it already clear that the convergence is quite poor: the three histograms on the right differ by more than the precision I would typically be aiming for. See associated R-file for further examples and illustrative experiments.

2.3 The Gibbs sampler and data augmentation

2.3.1 The Gibbs sampler

Consider a p-dimension target distribution $\pi(\theta)$, $\theta \in \mathbb{R}^p$, with $\theta = (\theta_1, ..., \theta_p)$ as in Section 2.1.8. *Random scan Gibbs* is a multi-component Metropolis Hastings sampler of the form given in Section 2.1.8 which selects components $i = 1, ..., p$ for update with probability $\xi_i = 1/p$ and takes as proposal density the conditional density, $q_i(\theta_i|\theta) = \pi(\theta_i|\theta_{-i})$ where

$$\pi(\theta_i|\theta_{-i}) = \frac{\pi(\theta)}{\pi(\theta_{-i})}$$

is the conditional density. If we drop this proposal into the MH-MCMC algorithm we find the acceptance probability is equal one.

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

1. Simulate $i \sim U\{1, ..., p\}$ and $\theta'_i \sim \pi(\cdot|\theta_{-i})$. Set $\theta'_{-i} = \theta_{-i}$.
2. Set $X_{t+1} = \theta'$.

Exercise 2.14. Write down $\alpha(\theta'|\theta)$ for this case and show it equals one, so the random-scan Gibbs sampler is a special case of Metropolis-Hastings.

We still need to check irreducibility as transition probability densities based on sequences of conditionals need not be irreducible. On the other hand if the chain is irreducible and aperiodic (in the discrete sense at least) the the chain is an ergodic chain targeting π.

The description above assumes the parameters are taken one at a time, so the conditionals are univariate. If it is possible to group parameters and sample the joint distribution of the parameters in each group conditional on all the others then we would generally do this for an improved mixing rate.

The *sequential-scan Gibbs sampler* visits each variable θ_i in turn from $i = 1...p$. This is also stationary. It is not reversible as the direction of simulation in the chain can be determined from the sequence of updates (for $p > 2$). It is sometimes possible to achieve better mixing rates (ie, larger ESS per update) but choosing the order in which the components are visited in a good way. As always irreducibility must be checked.

Example: Bivariate density $\pi(\theta_1, \theta_2)$ and $X_t = (\theta_1, \theta_2)$.

Algorithm:

1. Simulate $\theta'_1 \sim \pi(\theta'_1|\theta_2)$ then $\theta'_2 \sim \pi(\theta'_2|\theta'_1)$.
2. Set $X_{t+1} = (\theta'_1, \theta'_2)$.

Proposition 2.15. *If $X_t \sim \pi$ then after these two steps we have a new correlated sample $X_{t+1} \sim \pi$, so the process is stationary wrt π.*

Proof: The distribution of $X_{t+1} = (\theta'_1, \theta'_2)$ is

$$\begin{aligned} p(\theta'_1, \theta'_2) &= \int \pi(\theta_1, \theta_2)\pi(\theta'_1|\theta_2)\pi(\theta'_2|\theta'_1)d\theta_1 d\theta_2 \\ &= \int \pi(\theta_1, \theta_2)\frac{\pi(\theta'_1, \theta_2)}{\pi(\theta_2)}\frac{\pi(\theta'_1, \theta'_2)}{\pi(\theta'_1)}d\theta_1 d\theta_2 \\ &= \int \pi(\theta_1|\theta_2)\pi(\theta_2|\theta'_1)\pi(\theta'_1, \theta'_2)d\theta_1 d\theta_2 \\ &= \pi(\theta'_1, \theta'_2) \end{aligned}$$

as $\pi(\theta_1|\theta_2)$ and $\pi(\theta_2|\theta'_1)$ are normalised in the last step. 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 same approach works for $p > 2$ variables. [End of proof]

2.3.2 Data Augmentation

Some important early applications of the Gibbs sampler arise for missing data. This is also called "data augmentation" (DA). 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

$$z \sim p(z|\theta), \qquad y \sim p(y|z,\theta)$$

and we observe y. The posterior $\pi(\theta|y)$ is awkward as the likelihood function is an integral,

$$\pi(\theta|y) \propto \pi(\theta) \int p(y|z,\theta)p(z|\theta)dz.$$

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

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

This idea - of treating missing data z as simply another parameter like θ, but one for which the "prior" is the observation model - works for more literal "missing data" as well - the case where some components or entries in the data table $y = (y_1, ..., y_n)$ are missing. We can simultaneously infer the missing data, and use it along with the observed data to learn about the parameters. The extra uncertainty due to missing data is fed through into the uncertainty in the parameters in the overall posterior. The next example is not really in this class and uses the idea in a different way. In the next example the "missing data" are auxiliary variables introduced as something of a mathematical artifact to make the MCMC easier.

2.3.3 A Gibbs sampler for Probit regression

Probit regression is similar to logistic regression. It fits a GLM in which we have covariates $x_i = (x_{i,1}, ..., x_{i,p})$ for the ith observation $y_i \in \{0, 1\}, i = 1, ..., n$, parameters $\theta = (\theta_1, \theta_2, ..., \theta_p)$, a linear predictor $\eta_i = \sum_j \theta_j x_{i,j}$, observation model

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

and inverse link function $E(Y_i) = \Phi(\eta_i)$, where Φ is the cdf of $N(0, 1)$. If the prior for θ is $\pi(\theta)$ then the posterior for $\theta|y$ is

$$\pi(\theta|y) \propto \pi(\theta) \prod_{i=1}^{n} \Phi(\eta_i)^{y_i}(1 - \Phi(\eta_i))^{1-y_i}$$

with $\eta_i = \eta_i(\theta), i = 1, ..., n$. We cant calculate conditionals $\pi(\theta_j|\theta_{-j}, y)$ as θ appears inside Φ so we cant Gibbs sample. We could just target $\pi(\theta|y)$ using MH-MCMC, but Gibbs sampling is particularly neat, so here is a way to recover it by introducing latent variables.

There is another way to represent this model. If

$$z_i \sim N(\eta_i, 1), i = 1, ..., n$$

then $z = \eta + W$ with $W \sim N(0, 1)$ so $\Pr(z > 0) = \Pr(W > -\eta) = \Phi(\eta)$ by symmetry. It follows that if we set $y_i = 1$ if $z_i > 0$ and $y_i = 0$ if $z_i \leq 0$ then

$$\Pr(y_i = 1|\theta) = \Phi(\eta_i).$$

If we have z then the value of y is known so $p(y_i|z_i) = \mathbb{I}_{y_i=\mathbb{I}_{z_i>0}}$. The joint posterior augmented with z is

$$\begin{aligned}\pi(\theta, z|y) &\propto p(y|z)\pi(z|\theta)\pi(\theta) \\ &= \pi(z|\theta)\pi(\theta)\prod_i \mathbb{I}_{y_i=\mathbb{I}_{z_i>0}}.\end{aligned}$$

The marginal for θ is just $\pi(\theta|y)$, since

$$\begin{aligned}\pi(\theta|y) &\propto \pi(\theta)\prod_i \int \pi(z_i|\theta)\mathbb{I}_{y_i=\mathbb{I}_{z_i>0}}dz_i \\ &\propto \pi(\theta)p(y|\theta),\end{aligned}$$

so nothing has changed. In this representation we have a latent 'propensity' score z for each observation y, and we effectively observe the sign of z.

Consider Bayesian inference with a normal prior $\theta \sim N(0, \Sigma)$. To get a Gibbs sampler we need the conditionals $\pi(\theta|y, z)$ and $\pi(z_i|\theta)$. Looking at the θ-dependence in $\pi(\theta, z|y) \propto p(y|z)\pi(z|\theta)\pi(\theta)$ we see $\pi(\theta|y, z) \propto \pi(z|\theta)\pi(\theta)$ and this is jointly normal[5] in θ as the prior is conjugate. The z-dependence is in $p(y|z)\pi(z|\theta)$ with the first term a product of indicators so

$$\pi(z_i|\theta, y_i) \propto N(z_i; \eta_i, 1)\mathbb{I}_{y_i=\mathbb{I}_{z_i>0}}.$$

Here then is the Gibbs sampler targeting $\pi(\theta, z|y)$ by alternating between θ-updates and z-updates. Suppose $X_t = (\theta, z)$ and $\eta = \eta(\theta)$.

1. For $i = 1, ..., n$, simulate $z_i' \sim N(\eta_i, 1)\mathbb{I}_{y_i=\mathbb{I}_{z_i'>0}}$.
2. Simulate $\theta' \sim \pi(\theta|z')$ (multivariate normal, no y given z').
3. Set $X_{t+1} = (\theta', z')$.

This is implemented in some code available online with the course material, and applied to a simple example.

2.4 Some further applications of Monte Carlo methods in Bayesian inference

2.4.1 Estimating marginal likelihoods

We have Monte-Carlo tools summarising

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

the posterior under model m with $\theta \in \Omega_m$. How do we use the MCMC output to do model selection? Let $\hat{p}_m$ estimate the Marginal Likelihood (ML) $p(y|m)$ and $\hat{B}_{m',m}$ estimate the Bayes factor $B_{m,m'} = p(y|m)/p(y|m')$ for comparison of models $M = m$ and $M = m'$.

Here are some consistent ML-estimators, of increasing stability.[6]

The naive estimate: Since $p(y|m) = E_\theta(p(y|\theta, m))$ we could simply average the likelihood in the prior. Simulate $\theta^{(t)} \sim \pi(\theta|m), t = 1...T$ and form the estimate $\hat{p}_m = T^{-1}\sum_t p(y|\theta^{(t)}, m)$.

[5] See Problem Sheet 1 and R code example.

[6] The Laplace estimator for ML appears later in this course. For now see "Core Statistics", Simon Wood, page 153.

The failure of this estimator in practice reflects the fundamental problem of estimating a marginal likelihood. The prior is typically diffuse over the parameter space, while the function we are averaging, $p(y|\theta, m)$ is typically very close to zero except on a relatively small set of θ-values. Most of the mass of the function is concentrated in this small set. If we simply simulate the prior, the proportion of samples actually hitting this set may be small (or zero).

The Harmonic Mean estimate: this is importance sampling using the posterior.

Simulate $\theta^{(t)} \sim \pi(\theta|y, m), t = 1...T$, perhaps using MCMC. If

$$w_t = \pi(\theta^{(t)}|m)/\pi(\theta^{(t)}|y, m)$$

then

$$\hat{p}'_m = \frac{1}{T}\sum_t w_t\, p(y|\theta^{(t)}, m)$$

is a consistent and unbiased estimate for $p(y|m)$. This is standard importance sampling: since the samples are identically distributed (not necessarily independent),

$$\begin{aligned} E_{\theta^{(1:T)}|y,m}(\hat{p}'_m) &= T^{-1}\sum_t \int_\Omega w_t\, p(y|\theta^{(t)}, m)\pi(\theta^{(t)}|y, m)\, d\theta^{(t)} \\ &= \int_\Omega p(y|\theta, m)\pi(\theta|m)\, d\theta \end{aligned}$$

where $\theta^{(1:T)} = (\theta^{(1)}, ..., \theta^{(T)})$ and we substituted in the weights and canceled the posterior. We cant compute normalised weights w_t as we dont know the marginal likelihood which appears in the posterior $\pi(\theta^{(t)}|y, m)$, so we use $\tilde{w}_t \propto w_t$ (ie, dropping the marginal likelihood factor, which is constant in $\theta^{(t)}$). The prior factors cancel and we have

$$\tilde{w}_t = 1/p(y|\theta^{(t)}, m).$$

Now

$$\begin{aligned} E_{\theta^{(t)}|y,m}(\tilde{w}_t) &= \int_\Omega \frac{\pi(\theta^{(t)}|y, m)}{p(y|\theta^{(t)}, m)}\, d\theta^{(t)} \\ &= \int_\Omega \frac{\pi(\theta|m)}{p(y|m)}\, d\theta \\ &= p(y|m)^{-1}. \end{aligned}$$

It follows that $T^{-1}\sum_t \tilde{w}_t$ converges in probability to $p(y|m)^{-1}$. The "self-normalised", biased but still consistent, IS-estimator for the marginal likelihood $p(y|m)$ is then the inverse of this,

$$\hat{p}_m = \left[\frac{1}{T}\sum_t \frac{1}{p(y|\theta^{(t)}, m)}\right]^{-1}.$$

This estimator is not to be trusted (though widely used). The problem is that it is exposed to rare very large weights which arise when $\theta^{(t)}$ is in the tail of the posterior, so $p(y|\theta^{(t)}, m)$ is very small, and its inverse large. Essentially we need a "better" importance sampling distribution. Bridge estimators generalise importance sampling in this setting and typically yield much more stable estimates, as they directly minimise the mean squared error.

Bridge Estimate: In the following presentation we temporarily drop the model indicator m, so $p(y)$ is $p(y|m)$ etc.

Proposition 2.16. *Let $h : \Omega \to R$ be a given function with the property that following expectations are finite and non-zero. The identity*

$$p(y) = \frac{E_\theta(\pi(\theta)p(y|\theta)h(\theta))}{E_{\theta|y}(\pi(\theta)h(\theta))}$$

holds. Expectation in the numerator is over the prior and in the denominator expectation is over the posterior.

Exercise 2.17. Verify this. Replace the expectations with integrals, substitute $\pi(\theta|y) = \pi(\theta)p(y|\theta)/p(y)$ and cancel.

Exercise 2.18. The Harmonic mean estimator is based on the identity $p(y) = 1/E_{\theta|y}(1/p(y|\theta))$. What choice of h gives this identity?

We can estimate the RHS of the identity in Proposition 2.16 straightforwardly. Let $\theta^{(1,t)} \sim \pi(\theta)$ be a set of T samples from the prior and $\theta^{(2,t)} \sim \pi(\theta|y), t = 1...T$ be a set of T samples from the posterior. Plug in the natural estimates for the numerator and denominator and get

$$\hat{p} = \frac{\sum_t \pi(\theta^{(1,t)})p(y|\theta^{(1,t)})h(\theta^{(1,t)})}{\sum_t \pi(\theta^{(2,t)})h(\theta^{(2,t)})}.$$

This is consistent for $p(y)$ if the two sets of samples are each iid (or if they are suitable MCMC output, as for the harmonic mean estimate).

The power of this setup is that the identity holds for a very large class of functions h. We can choose $h(\theta)$ to make the (RMSE) Relative Mean Square Error $E((\hat{p} - p(y))^2)/p(y)^2$ small (expectation is over Monte-Carlo sampling variation of $\hat{p}$).

Exercise 2.19. (PS2!) Show that the choice of h minimising the RMSE for iid samples is

$$h(\theta) \propto \frac{1}{\pi(\theta) + \pi(\theta|y)}.$$

This can be shown using calculus of variations or see Meng and Wong (1996).

The problem here is that the optimal h depends on the normalised posterior, and we cant calculate that without knowing the thing we are trying to estimate. Meng and Wong (1996) give an iterative algorithm which works very well. However they also remark that the simple choice $h \propto 1/\sqrt{\tilde{p}_1\tilde{p}_2}$ is often near optimal for bridging densities $\tilde{p}_1/Z_1$ and $\tilde{p}_2/Z_2$. In our setting with densities $\tilde{p}_1 = \pi(\theta)$ and $\tilde{p}_2 = \pi(\theta)p(y|\theta)$ this gives $h(\theta) = \pi(\theta)^{-1}p(y|\theta)^{-1/2}$ and

$$\hat{p} = \frac{\sum_t p(y|\theta^{(1,t)})^{1/2}}{\sum_t p(y|\theta^{(2,t)})^{-1/2}}.$$

This has much lower RMSE than the harmonic mean. The optimal bridge aside, this is one of the best generic and reasonably straightforward estimators currently known.

Returning our model index, so $p(y) \to p(y|m)$ and $\hat{p} \to \hat{p}_m$, the estimator above gives $\hat{p}_m$. In order to estimate the Bayes factor we estimate the marginal likelihood for each model and form $\hat{B}_{m',m} = \hat{p}_{m'}/\hat{p}_m$.

If you refer to the code I used to carry out model selection for the radiocarbon dating example in Section 1.4 you will see both the harmonic mean and bridge sampling estimators are computed there. If you run the code several times (quite time-consuming) you will see that the HM estimator fluctuates quite a bit, the bridge estimate not so much.

In special case where $\Omega_1 = \Omega_2 = \Omega$ (so the two models have the same parameter space), we can directly estimate the Bayes factor in a single estimate using samples from the two posterior distributions.

Proposition 2.20. *Let $h : \Omega \to R$ be a given function with the property that the following expectations are finite and non-zero. The identity*

$$\frac{p(y|m)}{p(y|m')} = \frac{E_{\theta|y,m'}(\pi(\theta|m)p(y|\theta,m)h(\theta))}{E_{\theta|y,m}(\pi(\theta|m')p(y|\theta,m')h(\theta))}$$

holds. Expectation in the numerator/denominator is over the posterior in model m'/m.

Exercise 2.21. Verify this identity using the same procedure as Exercise 2.17.

Taking $m = 1$ and $m' = 2$ as the model indices, $\theta^{(1,t)} \sim \pi(\theta|y, m = 1)$ and $\theta^{(2,t)} \sim \pi(\theta|y, m = 2)$, $t = 1...T$, and

$$h(\theta) = (\pi(\theta|m)p(y|\theta,m)\pi(\theta|m')p(y|\theta,m'))^{-1/2},$$

gives

$$\hat{B}_{m,m'} = \frac{\sum_t \left(\frac{\pi(\theta^{(2,t)}|m)p(y|\theta^{(2,t)},m)}{\pi(\theta^{(2,t)}|m')p(y|\theta^{(2,t)},m')}\right)^{1/2}}{\sum_t \left(\frac{\pi(\theta^{(1,t)}|m')p(y|\theta^{(1,t)},m')}{\pi(\theta^{(1,t)}|m)p(y|\theta^{(1,t)},m)}\right)^{1/2}},$$

fairly stable and close to the state of the art for generic estimators. It is very convenient to have a model selection estimator given in terms of samples from the two posteriors, as we typically have these available anyway. Also, estimating a single ratio rather than the ratio of two estimates will in this bridge-setting typically give a more stable estimate.

2.4.2 Example: Selecting a link function in a model for O-ring data

Consider regression in a Bernoulli GLM with p covariates $x = (x_1, ..., x_p)$, p-component vector $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ of effects, linear predictor $\eta = x\beta$, and binary response $Y \sim Bern(\mu_m(\eta))$ in model $m \in \mathcal{M} = \{1, 2\}$.

Consider the two inverse link functions:

1. logistic $\mu_1(\eta) = \exp(\eta)/(1 + \exp(\eta))$;
2. probit $\mu_2(\eta) = \Phi(\eta)$.

Illustrate this on the Challenger O-ring data, data $(y_i, x_i), i = 1, 2, ..., n$ with

$$y_i \sim Bern(\mu_m(\beta_1 + \beta_2 x_i)),\ i = 1, ..., n$$

under link-function model m and $x_i \in R$ centred and scaled temperature `temp[i]` from data.

The observation models are, for $m = 1, 2$,

$$p(y|\theta, m) = \prod_{i=1}^{n} \mu_m(\beta_1 + \beta_2 x_i)^{y_i}(1 - \mu_m(\beta_1 + \beta_2 x_i))^{1-y_i}.$$

Consider the prior $\beta_1, \beta_2 \sim N(0, 3^2)$. The marginal likelihoods are

$$p(y|m) \propto \int \prod_{i=1}^{n} p(y_i|\beta, m) \exp(-\beta^T\beta/18) d\beta_1 d\beta_2.$$

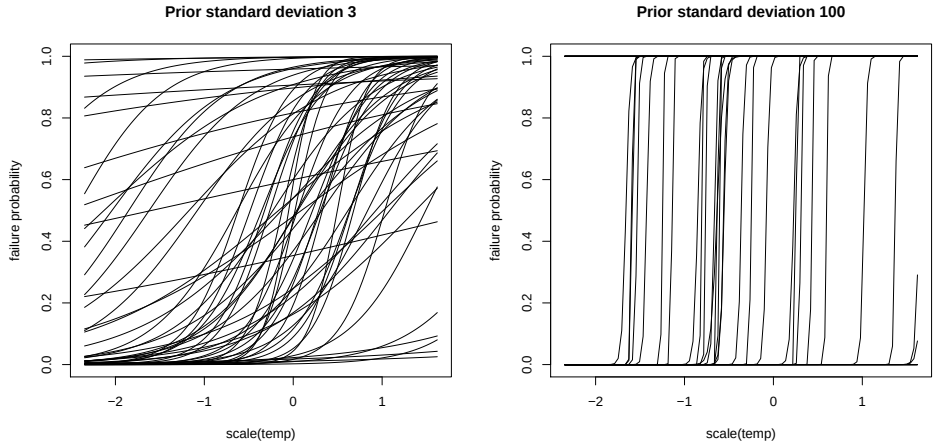


Figure 9: Samples of $\mu(\eta)$, the probability for O-ring failure, drawn form the prior: (Left) prior standard deviation of β_1, β_2 equal 3; (Right) prior standard deviation 100. In both graphs the subset of increasing $\mu(\eta)$ only is plotted for clarity.

The estimators above are computed from samples generated by MCMC in the online example code for this lecture. The resulting estimates of Bayes Factors: 2.7 (naive - pretty good!) 1.9 (harmonic) 2.8 (bridge) 2.7 (R-built-in Laplace-approximation estimator). We report $\hat{B}_{1,2} = 2.75$, weak evidence "barely worth mentioning" for Logit over Probit.

2.4.3 Using Simulation to check a prior

This is an off-topic remark on elicitation which naturally goes here as it is associated with the model in the previous section.

Why the $\beta_1, \beta_2 \sim N(0, 3^2)$ prior? You will commonly see $N(0, v)$ with v large, often presented as a "non-informative" choice. The linear predictor is $\eta_i = \beta_1 + \beta_2 x_i$ with scaled temperature

$$x_i = \texttt{scale(temp)[i]} = (\texttt{temp[i]} - \texttt{mean(temp)})/\texttt{sd(temp)}.$$

Take the logistic link in this example.

Before seeing the data we have some prior knowledge of $\mu(\eta)$, the probability for O-ring failure. We know they can but dont always fail. The probability for failure shouldn't vary too sharply with temperature but must depart substantially from zero and one.

This is a situation where we know something about a function of the parameters, and not so much about the parameters themselves. We can check the prior we have represents the information we have using simulation from the prior. We simulate $\beta_1, \beta_2 \sim N(0, v^2)$ from the prior for a few values of v and look at $\mu(\beta_1 + \beta_2 x)$ as a function of x. In Figure 9, large values of v put high probability on a very steep $\mu(\eta)$-functions. This is implausible on physical grounds. $N(0, 3^2)$ *allows* fairly sharp dependence but favors a weaker effect.

3 Likelihood-free methods: Approximate Bayesian Computation

3.1 Motivation and Definitions

3.1.1 Doubly intractable distributions

We often use approximation methods to approximate the integrals and sums we need to evaluate in Bayesian inference. However some approximation methods go further and approximate the posterior distribution itself. After all, why work hard to fit exactly a model that is wrong anyway?

Definition 3.1. Denote by $\mathcal{Y}$ the space of realisable data vectors $y = (y_1, ..., y_n)$. We emphasise that y is the entire data set, not just one sample, so for multivariate observations we may have $y_i \in \mathbb{R}^d$ etc. In this section we use $y_{obs} \in \mathcal{Y}$ to represent the real observed data. We use symbols like $y, y' \in \mathcal{Y}$ for simulated data.

Definition 3.2. (*doubly intractable problems*) If for $y_{obs} \in \mathcal{Y}$ and $\theta, \theta' \in \Omega$ either of the ratios $p(y_{obs}|\theta)/p(y_{obs}|\theta')$ or $\pi(\theta)/\pi(\theta')$ cannot be evaluated then the posterior is said to be *doubly intractable*.

Remark 3.3. This commonly arises in two main settings.

(1) Un-normalised likelihood: the observation model is given in the form

$$p(y_{obs}|\theta) = p(y_{obs}, \theta)/c(\theta)$$

with

$$c(\theta) = \int_{\mathcal{Y}} p(y, \theta) dy$$

and $c(\theta)$ may be intractable. However we sometimes also encounter intractable priors.

(2) Likelihood free: Sometimes the observation model is not defined by a density at all, just a complex simulator - essentially a piece of code that takes as input a parameter θ and some random variables Z with the same dimension as Y and returns as output a simulated realisation of Y. Although $p(y_{obs}|\theta)$ is formally defined by the simulator, we have no idea what the function $p(y|\theta)$ actually is and how it depends on θ. This appears alot in big physics based models including some climate models.

Remark 3.4. The concept of intractability is not well-defined here. When we say "X cannot be evaluated", we mean "X cannot be evaluated in any reasonable time". This may depend on available computing resources etc.

Example 3.5. (*Ising model*) Denote by $\mathcal{Y} = \{0, 1\}^{n^2}$ the set of all binary $n \times n$ "images" $Y = (Y_1, Y_2, ..., Y_{n^2})$, $Y_i \in \{0, 1\}$, where $i = 1, 2, ..., n^2$ is the cell or "pixel" index in the square lattice of image cells. We say two cells i and j are *neighbors* and write $i \sim j$ if the cells share an edge in

1	2	3
4	5	6
7	8	9

0	0	1
0	1	1
0	1	1

Table 1: (Left) cell index labels in a 3×3 square lattice and (Right) a possible realisation of the Ising model.

the lattice. For example $5 \sim 6$ in Table 1 and the neighbors of cell 9 are $\{8, 6\}$. Let

$$\#y = \frac{1}{2} \sum_{i=1}^{n^2} \sum_{j \sim i} \mathbb{I}_{Y_j \neq Y_i},$$

where $\sum_{j\sim i}$ sums over all j such that j is a neighbor of i. Here $\#y$ counts the number of "disagreeing neighbors" in the binary image $Y = y$. In the example in Table 1 we have $\#y = 4$.

The *Ising model* with a free boundary is the following distribution over $\mathcal{Y}$:

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

It has a free boundary because cells on the edge have no neighbors beyond the edge. Here $\theta \geq 0$ is a positive *smoothing parameter* and

$$c(\theta) = \sum_{y \in \mathcal{Y}} \exp(-\theta \#y)$$

is a normalizing constant which we cant compute for n large. There are 2^{n^2} terms in the sum, and no one can solve it (it is an important model in physics, so many have looked at this). [7]

Suppose we have image data $Y = y_{obs}$ with $y_{obs} \in \{0,1\}^{n^2}$ and we want to estimate θ. Consider doing MCMC targeting $\pi(\theta|y_{obs})$ with some prior $\pi(\theta)$. Choose a simple proposal for the scalar parameter θ, say $\theta' \sim U(\theta - a, \theta + a),\ a > 0$. The acceptance probability is

$$\begin{aligned}\alpha(\theta'|\theta) &= \min\left\{1,\ \frac{p(y|\theta')\pi(\theta')}{p(y|\theta)\pi(\theta)}\right\} \\ &= \min\left\{1, \frac{c(\theta)}{c(\theta')}\ \times\ \exp((\theta - \theta')\#y)\ \times\ \frac{\pi(\theta')}{\pi(\theta)}\right\}\end{aligned}$$

and although $p(y)$ cancels, $c(\theta)/c(\theta')$ does not, and so we are left with an acceptance probability we cannot evaluate. [End of Example]

3.1.2 The ABC posterior

Exercise 3.6. (repeat Exercise 1.2) Suppose the data Y and Θ are discrete random variables. Show that the algorithm

1. Set $n = 0$
2. Set $n \leftarrow n + 1$. Simulate $\theta_n \sim \pi(\cdot)$ and $y_n \sim p(\cdot|\theta_n)$
3. If $y_n = y_{obs}$ stop and return $\Theta' = \theta_n$ and $N = n$ and otherwise goto Step 2,

returns Θ' distributed like $\Pr(\Theta' = \theta) = \pi(\theta|y_{obs})$ so according to the exact posterior.

ANS: this is a rejection-sampling algorithm. We stop when $y_n = y_{obs}$ so we stop with probability $p(y_{obs}) = \sum_\theta \pi(\theta)p(y_{obs}|\theta)$. The probability the output is $\Theta' = \theta$ is

$$\Pr(\Theta' = \theta) = \sum_{n=1}^{\infty} \Pr(\Theta' = \theta, N = n)$$

event $\{\Theta' = \theta, N = n\}$ is event $\{\theta_n = \theta, y_1 \neq y_{obs}, ..., y_{n-1} \neq y_{obs}, y_n = y_{obs}\}$

$$= \sum_{n=1}^{\infty} \Pr(\theta_n = \theta, y_1 \neq y_{obs}, ..., y_{n-1} \neq y_{obs}, y_n = y_{obs})$$

[7] There is famously a formula for $c(\theta)$ for the special case of periodic boundary conditions, but that is not an image model we would generally use.

events in each loop are independent so we have a factor $1-p(y_{obs})$ for each rejection, and we draw $\theta_n = \theta,\ y_n = y_{obs}$ at step n with probability $\pi(\theta)p(y_{obs}|\theta)$,

$$= \sum_{n=1}^{\infty} \pi(\theta)p(y_{obs}|\theta)(1-p(y_{obs}))^{n-1}$$

This sum is geometric and equal to $\pi(\theta|y_{obs}),\ \theta \in \Omega$.

The idea in ABC is to relax the requirement that $y_n = y_{obs}$ and stop if $|y_n - y_{obs}|$ is small. We like θ-values with a large likelihood value $p(y_{obs}|\theta)$ so might like θ values with the property that simulated data $y \sim p(\cdot|\theta)$ are close to y_{obs}. ABC was originally an approximation defined by an algorithm, and only later was the approximation really characterised. We use it when the likelihood or prior are intractable but simulation is easy.

Assumption 3.7. (*simulation of* (y,θ) *from the generative model is tractable*) Suppose the posterior $\pi(\theta|y)$ is doubly intractable. We assume it is nevertheless possible to *simulate the generative model* $y \sim p(\cdot|\theta)$ and $\theta \sim \pi(\theta)$.

ABC assumes the exact posterior $\pi(\theta|y_{obs})$ doesn't change much if we make a small change in the data but keep θ fixed. To measure this we start by defining a distance measure between data vectors.

Definition 3.8. (*summary statistics and distance in* $\mathcal{Y}$) Let $S : \mathcal{Y} \to \mathbb{R}^p$ be a vector of $p \geq 1$ summary statistics on the data. For $y, y' \in \mathcal{Y}$, suppose $s = S(y)$ and $s' = S(y')$. We specify a distance measure $D : \mathbb{R}^p \times \mathbb{R}^p \to [0,\infty)$ on pairs s, s'. The distance between y and y' is then $D(s,s')$.

Remark 3.9. If we have sufficient statistics then we use them to specify $S(y)$. This gives us a low-dimensional representation of the data (as $p \ll n$ typically). We have $p(y|\theta) = g(S(y),\theta)f(y)$ with $f(y)$ a function not depending on θ, and so $\pi(\theta|y_{obs}) \propto g(S(y_{obs}),\theta)\pi(\theta)$. This will still be doubly intractable but at least the dimension of data space will be lower.

Remark 3.10. It is easy to simulate (S,θ)-pairs: simulate $\theta \sim \pi(\cdot)$, $Y \sim p(\cdot|\theta)$ and evaluate $S(Y)$.

Definition 3.11. Let $\Delta_\delta(y_{obs})$ be a "ball" of radius δ centred on y_{obs},

$$\Delta_\delta(y_{obs}) = \{y' \in \mathcal{Y} : D(S(y_{obs}), S(y')) \leq \delta\}.$$

The data y_{obs} is a realisation of a random variable $Y \in \mathcal{Y},\ Y \sim p(\cdot|\theta)$ with $\theta \sim \pi(\cdot)$ so we write

$$p(\Delta_\delta(y_{obs})|\theta) = \int_{\Delta_\delta(y_{obs})} p(y|\theta)\, dy.$$

For $y \in \mathcal{Y}$ let $p(y) = \int_\Omega \pi(\theta)p(y'|\theta)d\theta$ denote the prior predictive distribution for Y (ie the marginal likelihood at generic "data" y) so we may write

$$p(\Delta_\delta(y_{obs})) = \int_{\Delta_\delta(y_{obs})} p(y)\, dy.$$

Definition 3.12. (*ABC posterior*) We define the ABC posterior approximation to $\pi(\theta|y_{obs})$ to be

$$\pi_{ABC}(\theta|y_{obs}) = \frac{p(\Delta_\delta(y_{obs})|\theta)\pi(\theta)}{p(\Delta_\delta(y_{obs}))},$$

which we may alternatively write $\pi_{ABC}(\theta|y_{obs}) = \pi(\theta|Y \in \Delta_\delta(y_{obs}))$ by Bayes rule.

Remark 3.13. The ABC-posterior is the posterior we would get if our data was not the statement "$Y = y_{obs}$", but instead the statement "$Y \in \Delta_\delta(y_{obs})$". This is like seeing the data at lower resolution. We cant see the exact data we just know it is in this set $\Delta_\delta(y_{obs})$.

Proposition 3.14. *The ABC posterior can be written the following form,*

$$\pi_{ABC}(\theta|y_{obs}) = \int_{\Delta_\delta(y_{obs})} \pi(\theta|y)p(y|Y \in \Delta_\delta(y_{obs}))dy, \tag{3.1}$$

where

$$p(y|Y \in \Delta_\delta(y_{obs})) = \frac{p(y)\mathbb{I}_{y\in\Delta_\delta(y_{obs})}}{p(\Delta_\delta(y_{obs}))}. \tag{3.2}$$

Proof: Substituting Equation 3.2 in the RHS of Equation 3.1,

$$\begin{aligned}
\int_{\Delta_\delta(y_{obs})} \pi(\theta|y)p(y|Y \in \Delta_\delta(y_{obs}))dy &= \int_{\Delta_\delta(y_{obs})} \frac{\pi(\theta|y)p(y)}{p(\Delta_\delta(y_{obs}))}dy \\
&= \frac{\int_{\Delta_\delta(y_{obs})} p(y|\theta)\pi(\theta)dy}{p(\Delta_\delta(y_{obs}))} \\
&= \frac{p(\Delta_\delta(y_{obs})|\theta)\pi(\theta)}{p(\Delta_\delta(y_{obs}))} \\
&= \pi_{ABC}(\theta|y_{obs}).
\end{aligned}$$

[End of Proof]

3.2 Computational methods for ABC

3.2.1 Simulating the ABC posterior via rejection

We wouldnt make an approximation unless it helped us somehow. The point is that, if δ is not too small, $\pi(\theta|Y \in \Delta_\delta(y_{obs}))$ is often very easy to sample, and we can do it "perfectly" using rejection, even in cases where the observation model is very complex. Rejection-samples are iid and distributed according to the target, so there are no issues with burn-in or mixing as in MCMC.

Proposition 3.15. *ABC Rejection Algorithm*

1. *Set* $n = 0$
2. *Set* $n \leftarrow n + 1$. *Simulate* $\theta_n \sim \pi(\cdot)$ *and* $y_n \sim p(\cdot|\theta_n)$
3. *If* $y_n \in \Delta_\delta(y_{obs})$ *stop and return* $\Theta_{ABC} = \theta_n$, $Y_{ABC} = y_n$ *and* $N = n$ *else goto step 2.*

The ABC-Rejection Algorithm returns samples $(\Theta_{ABC}, Y_{ABC}, N)$ *with* $\Theta_{ABC} \sim \pi(\cdot|Y \in \Delta_\delta(y_{obs}))$.

Remark 3.16. We introduced two new random variables Θ_{ABC} and Y_{ABC}. The point is to distinguish between $\Theta \sim \pi(\cdot|y_{obs})$ and $\Theta_{ABC} \sim \pi_{ABC}(\cdot|y_{obs})$.

Proof: (partitioning on N) For any suitable set $A \subseteq \Omega$, $\Pr(\Theta_{ABC} \in A)$ is the long-run proportion of times we run the Algorithm and get output $\Theta_{ABC} \in A$. We would like to show that $\Pr(\Theta_{ABC} \in A) = \pi_{ABC}(A|y_{obs})$.

If $(\theta, y) \sim \pi(\theta)p(y|\theta)$ then each loop is an independent trial which succeeds with probability $p(\Delta_\delta(y_{obs})) = \Pr(Y \in \Delta_\delta(y_{obs}))$ so $N \sim Geom(p(\Delta_\delta(y)))$.

We get output $\Theta_{ABC} \in A$ and $N = n$ if and only if $y_m \notin \Delta_\delta(y_{obs})$ for $m < n$ (so we don't stop before $N = n$) and $(\theta_n, y_n) \in A \times \Delta_\delta(y_{obs})$ (so we stop at $N = n$ with $\Theta_{ABC} \in A$) so

$$\begin{aligned}
\Pr(\Theta_{ABC} \in A) &= \sum_{n=1}^{\infty} \Pr(\Theta_{ABC} \in A, N = n) \\
&= \sum_{n=1}^{\infty} \Pr((\theta_n, y_n) \in A \times \Delta_\delta(y_{obs}), y_1, ..., y_{n-1} \notin \Delta_\delta(y_{obs})) \\
&= \sum_{n=1}^{\infty} \Pr((\theta, y) \in A \times \Delta_\delta(y_{obs}))(1 - p(\Delta_\delta(y_{obs})))^{n-1}
\end{aligned}$$

where the last line follows because events in every loop are independent of events in every other loop, and we replace (θ_n, y_n) with a generic pair $(\theta, y) \sim \pi(\theta)p(y|\theta)$.

$$\begin{aligned}
\Pr(\Theta_{ABC} \in A) &= \frac{\int_A \int_{\Delta_\delta(y_{obs})} \pi(\theta)p(y|\theta)dyd\theta}{p(\Delta_\delta(y_{obs}))} \\
&= \frac{\int_A p(\Delta_\delta(y_{obs})|\theta)\pi(\theta)d\theta}{p(\Delta_\delta(y_{obs}))} \\
&= \int_A \pi_{ABC}(\theta|y)d\theta \\
&= \pi_{ABC}(A|y) \qquad \text{[EOP]}
\end{aligned}$$

Example 3.17. Here is a very simple example in which the data are five samples from a Poisson with mean λ and we have a Gamma-prior for λ.

Data model: $y_{obs_i} \sim \text{Poisson}(\Lambda)$, iid for $i = 1, 2, ..., n$ with $n = 5$ and truth $\Lambda = 2$.

Data space: $\mathcal{Y} = \{0, 1, 2, ..., \}^n$.

Prior: $\lambda \sim \Gamma(\alpha = 1, \beta = 1)$.

Summary statistic: if if $y = (y_1, ..., y_n)$ for $y \in \mathcal{Y}$ then we take $S(y) = \bar{y}$, the average of $y_1, ..., y_n$.

Distance measure: $D(S(y) - S(y_{obs})) = |S(y) - S(y_{obs})|$ and we will consider tolerance $\delta = 0.5, 1$.

ABC algorithm: here is the algorithm of Proposition 3.15 for this case.

1. Simulate $\lambda \sim \Gamma(\alpha, \beta)$ and $y_i \sim \text{Poisson}(\Lambda), i = 1, 2, ..., n$.
2. If $|\bar{y} - \overline{y_{obs}}| < \delta$ return λ and stop, otherwise goto (1).

Run this algorithm T times returning $\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(T)}$. These are samples from $\pi_{ABC}(\lambda|y_{obs})$.

We can estimate expectations, plot histograms and calculate a HPD set using these samples. Figure 10 shows density plots of samples. When we make δ smaller we include only y-values closer to y_{obs} and so the approximation $\pi_{ABC}(\theta|y_{obs})$ tends to improve (so the blue line at $\delta = 0.5$ looks more like the exact posterior in black than the green line at $\delta = 1$). We are doing Bayesian inference without calculating $p(y_{obs}|\theta)$ or $\pi(\theta)$, so this is sometimes called "likelihood free" inference. We just specify how to simulate parameters and data.

3.2.2 Regression adjustment of samples

Consider the pairs $(\theta, y) \sim \pi(\theta)p(y|\theta)$ generated by ABC. Conditional on y we have $\theta \sim \pi(\theta|y)$. We will adjust this distribution by making a transformation $\theta^{(adj)} = f(\theta)$ so that $\theta^{(adj)} \sim \pi(\cdot|y_{obs})$ (to a better approximation). This is not achievable in general. We will set out some assumptions

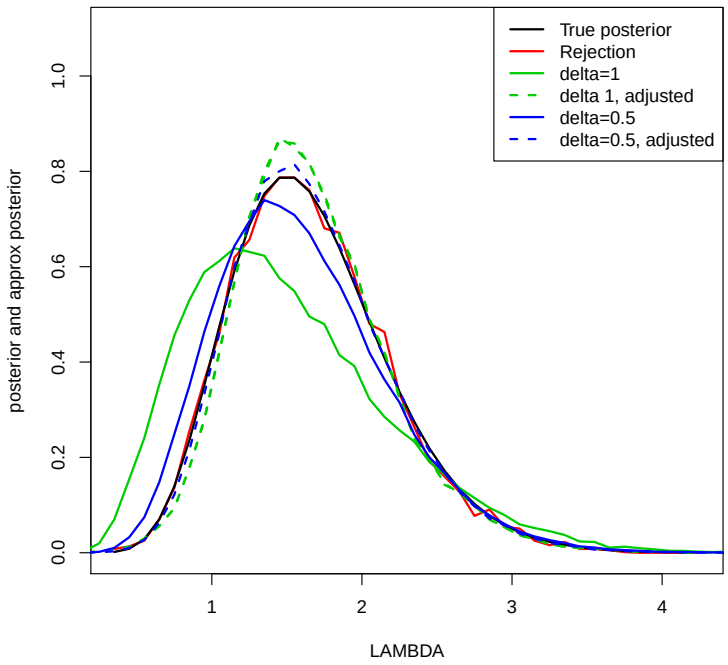


Figure 10: ABC-poisson example: (black curve) exact posterior; (red curve) rejection algorithm in Exercise 3.6; (green curve) samples from π_{ABC} with $\delta = 1$; (blue curve) samples from π_{ABC} with $\delta = 0.5$; (dashed curves) Regression adjusted samples. See code online for details.

under which this operation is exact and straightforward, and use this to motivate the method when the assumptions do not hold but are satisfied to a good approximation. We consider scalar $\theta \in \mathbb{R}$ here.

Assumption 3.18. $s_{obs} = S(y_{obs})$ is a sufficient statistic.

Remark 3.19. Under this assumption we may treat the statistics s_{obs} as the data. Suppose $(\theta, y) \sim \pi(\theta)p(y|\theta)$. Under Assumption 3.18, the posterior mean $\mu(s) = E(\theta|Y = y)$ is a function of $s = S(y)$ alone.

Let ϵ be a random variable with the same distribution as $\theta - \mu(s)$, so its density given y is

$$\epsilon|y \sim \pi(\mu(s) + \epsilon|y).$$

Assumption 3.20. Suppose $s = S(y)$. Changing the "data" from y to y_{obs} shifts the posterior mean from $\mu(s)$ to $\mu(s_{obs})$ but has no other effect on the distribution of θ, so

$$\pi(\mu(s) + \epsilon|y) = \pi(\mu(s_{obs}) + \epsilon|y_{obs})$$

or equivalently $\epsilon|y \sim \epsilon|y_{obs}$.

The distribution of the offset ϵ about the posterior mean doesn't change when we change the data.

Assumption 3.21. The mean is a linear function of s, that is

$$\mu(s) = \mu(s_{obs}) + (s - s_{obs})\beta.$$

Remark 3.22. This all looks like dreamland. However for s such that $D(s, s_{obs}) \leq \delta$ and $\mu(s)$ a smooth function, we can think of this as a local linear approximation good at small δ.

Proposition 3.23. *Under Assumptions 3.18, 3.20 and 3.21, if* $(\theta, y) \sim \pi(\theta)p(y|\theta)$ *with* $s = S(y)$ *and we set*

$$\theta_{adj} = \theta - (s - s_{obs})\beta$$

then $\theta_{adj} \sim \pi(\cdot|y_{obs})$.

Proof: substituting for $(s - s_{obs})\beta$ using Assumption 3.21 we get $\theta_{adj} = \theta + \mu(s_{obs}) - \mu(s)$. In terms of the random variable ϵ, $\theta_{adj} = \mu(s_{obs}) + \epsilon|y$. But the density of $\epsilon|y$ can be taken to be $\pi(\mu(s_{obs}) + \epsilon|y_{obs})$ so making a change of variable from ϵ to θ_{adj}, the density of θ_{adj} is $\pi(\theta_{adj}|y_{obs})$. [EOP]

Remark 3.24. How do we use this in practice? We have lots of pairs $(\theta^{(t)}, s^{(t)})$, $t = 1, ..., T$ with $s^{(t)} = S(y^{(t)})$ output from our ABC-rejection runs. We can regress $\theta^{(t)}$ on $s^{(t)} - s_{obs}$ to estimate the intercept $\widehat{\mu(s_{obs})}$ and slope $\hat{\beta}$ and set

$$\theta^{(t)}_{adj} = \theta^{(t)} - (s^{(t)} - s_{obs})\hat{\beta},\ t = 1, ..., T$$

an approximation to the exact adjustment in Proposition 3.23. The regression correction adjusts the distribution of θ at y to move it onto the distribution of θ at y_{obs}.

Example 3.25. We did this for the Poisson-Gamma posterior in Example 3.17. The code for this is available on the website. The improvement after regression adjustment is significant. Referring to Figure 10 we see the green dashed line ($\delta = 1$, regression-adjusted) is closer to the posterior than the solid green line (unadjusted) and similarly for the blue lines at $\delta = 0.5$.

Remark 3.26. One of the main reasons for doing regression adjustment is that it allows us to take δ quite large and fix the poor approximation using the regression adjustment. Some experimentation is usually necessary (to choose δ small enough so the approximation doesnt change significantly when we make it smaller). We like to take δ large as more samples are retained (in Proposition 3.15 in the ABC-rejection algorithm we throw out samples if $D(S(y), S(y_{obs})) > \delta$), or alternatively, shorter runtimes at a fixed sample size.

3.3 ABC example: the Ising Model

The Ising distribution $Y \sim p(\cdot|\theta)$ is easy to sample for moderate n-values using MCMC. Figure 11 shows three samples $Y \sim p(y|\theta)$ simulated at three different values of θ. These samples are not exactly distributed according to $p(y|\theta)$ as the MCMC wont have converged exactly to the target distribution. However we can make this error negligible by taking long MCMC runs and checking for convergence.

Remark 3.27. Here is an MCMC algorithm simulating $Y^{(1)}, ..., Y^{(K)}$ and targeting $p(y|\theta)$. We will need to run this algorithm *at each value of* N in the ABC algorithm in Proposition 3.15. When $N = n$ our sample y_n will be the final state $Y^{(K)}$ of an MCMC run targeting $p(y|\theta_n)$, since $Y^{(K)} \sim p(\cdot|\theta_n)$ approximately. Notice that the intractable normalising constant (as a function of θ) isn't a problem as we are sampling y at this point not θ.

First we give a proposal distribution, Here is something simple. Suppose the current state is $y \in \{0, 1\}^{n^2}$. 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.

Here is the MCMC algorithm itself. Let $Y^{(t)} = y$. $Y^{(t+1)}$ is determined in the following way.

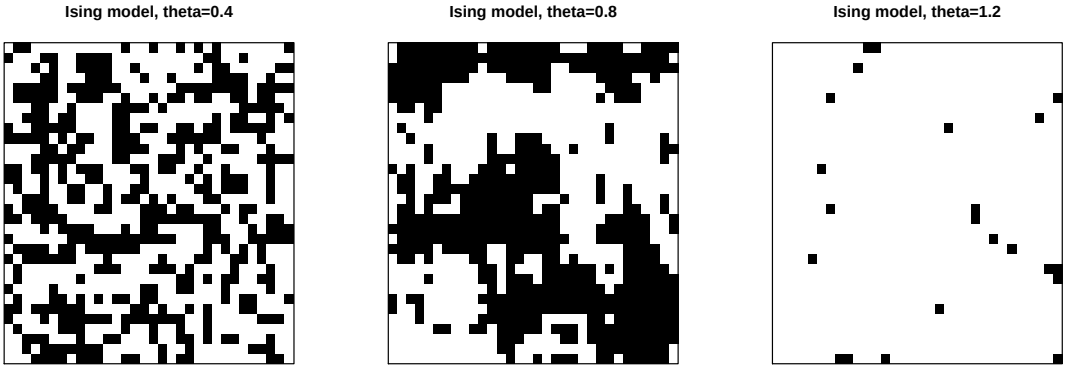


Figure 11: Three samples from the Ising model at different θ-values (see image titles).

1. Simulate $y' \sim q(y'|y)$ as above, and $u \sim U(0,1)$.
2. If $u \leq \alpha(y'|y)$ with
$$\alpha(y'|y) = \min\left\{1,\ \exp(-\theta(\#y' - \#y))\right\}$$
set $Y^{(t+1)} = y'$ and otherwise set $Y^{(t+1)} = y$.

The Markov chain is irreducible (q is irreducible and α is never zero) and aperiodic (rejection is possible), so it is ergodic for $p(y|\theta)$.

An implementation of this algorithm in R is available in the code for this section. Some samples produced using this code are shown in Figure 11.

Remark 3.28. We now define the ABC-rejection sampler for this problem. Our MCMC sampler is embedded inside this algorithm. We scale the ABC-distance by M, the maximum value $\#y$ can take (in a chessboard coloring) so δ is on a scale of $O(1)$.

Data model: $y_{obs} \sim \text{Ising}(\Theta)$ (with $n = 8$ so an 8×8 lattice and truth $\Theta = 0.8$).

Data space: $\mathcal{Y} = \{0,1\}^{n^2}$.

Prior: $\theta \sim \text{Exp}(1)$.

Summary statistic: if $y = (y_1, ..., y_n)$ for $y \in \mathcal{Y}$ then we take $S(y) = \#y$, which is sufficient.

Distance measure: $D(S(y), S(y_{obs})) = |S(y) - S(y_{obs})|/M$ and we take $\delta = 0.05, 0.1$.

ABC algorithm: here is the algorithm of Proposition 3.15 for this case.

1. Simulate $\theta \sim \text{Exp}(1)$ and $y \sim \text{Ising}(\theta)$ (using MCMC in Remark 3.27).
2. If $|\#y - \#y_{obs}| < M\delta$ return θ and stop, otherwise goto (1).

Run this algorithm T times returning $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(T)}$. These are samples from $\pi_{ABC}(\theta|y_{obs})$. We implemented this (see R code). Results are shown in Figure 12. 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 $\delta = 0.05$. It is time-consuming to gather ABC-samples at very small δ-values due to the high rejection rate, so the regression adjustment, which is computed much more rapidly, is helpful. We report the best estimate we have, the regression adjusted posterior at $\delta = 0.05$.

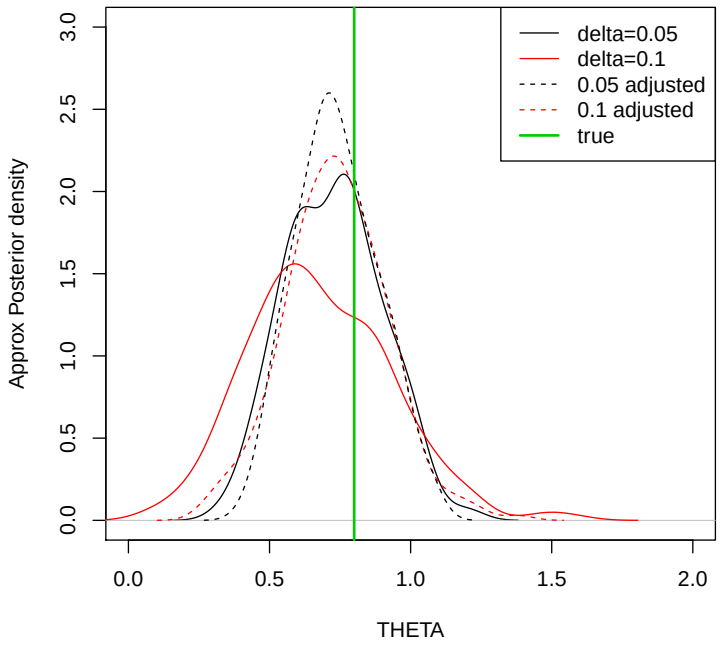


Figure 12: Density plots for samples from π_{ABC} varying δ and applying regression adjustment. The vertical line is the true θ-value in this synthetic-data example.

4 The Savage Axioms

The main point of this chapter is that if we have a collection of prior preferences expressed as "A is more likely than B", and those preferences satisfy the Savage Axioms, then there is a prior probability distribution π satisfying the axioms of probability and expressing those preferences, so $\pi(A) > \pi(B)$ etc.

In preparing this chapter I found *Morris H. DeGroot (2004), "Optimal Statistical Decisions"* particularly useful, and also *Christian Robert (2007), "The Bayesian Choice"*.

4.1 Utility theory

4.1.1 Rewards and Utility

Let $r(\theta, \delta)$ denote the reward if our action is $\delta \in \mathcal{D}$ and the truth is θ. We assume rewards, $r \in\in \mathbb{R}$ are bounded so $\mathbb{R} = [r_{\min}, r_{\max}]$. Rewards often take values in a finite set which we can take to be $\mathbb{R} = \{r_{\min}, r_{\min} + 1, ..., r_{\max}\}$. Let $U(r)$ denote the utility of reward r. In the literature we are referencing (principally Maurice DeGroot, "Optimal Statistical Decisions" (1970), but see also Christian Robert's book "The Bayesian Choice") the utility is also bounded. We wont make explicit use of this assumption, as we will state the main results without proof, but it must hold below.

Utility $U(r),\ r \in \mathbb{R}$ is the opposite of loss. In terms of our notation in Section 1.3.2,

$$L(\theta, \delta) = c - U(r(\theta, \delta)), \tag{4.1}$$

with c the largest attainable value taken by U. We are replacing one function L by two, U and r, which must be elicited.

Under the generative model for the data, $\theta \sim \pi(\cdot)$ and $y \sim p(y|\theta)$. Conditioned on the value of the data, the reward

$$R = r(\theta, \delta(y)), \quad \theta|y \sim \pi(\cdot|y) \tag{4.2}$$

is random as it is a function of the random variable $\theta|y \sim \pi(\cdot|y)$. It has a distribution, $R|y \sim P_{\delta,y}(r)$ say, where $P_{\delta,y}$ is a probability density or probability mass function (PMF). In the continuous case

$$\int_{\mathbb{R}} P_{\delta,y}(r) = 1.$$

If there is no data then the action doesn't depend on data (we would predict θ using the prior), so the reward distribution is determined from the prior. In this case,

$$R = r(\theta, \delta), \quad \theta \sim \pi(\cdot) \tag{4.3}$$

with P_δ again a normalised density or PMF.

Changing the action, or action function, $\delta(y)$ changes the reward distribution, so if we have two different action functions $\delta(y)$ and $\delta'(y)$ then we have different reward distributions P_δ and $P_{\delta'}$ over the same space of rewards. When we are discussing two reward distributions we may drop the the δ-subscript also and write P, P' etc.

The *expected utility* of the action δ at fixed y (for continuous rewards) is

$$E_{P_{\delta,y}}(U(R)) = \int_{\mathbb{R}} U(r) P_{\delta,y}(r) dr \tag{4.4}$$

The expected utility has the opposite sign to the expected posterior loss defined in Equation 1.1 in Section 1.3.2:

$$\begin{aligned} E_{P_{\delta,y}}(U(R))) &= E_{\theta|y}(U(r(\theta,\delta))) \\ &= c - \int_{\Omega} L(\theta,\delta)\pi(\theta|y)d\theta \end{aligned}$$

where first line follows from the definition of the random variable R in Eqn. 4.2 and the second using Eqn. 4.1. Choosing the action δ that maximises the expected utility is the same as choosing the action that minimises the expected posterior loss in Equation 1.1. So we are just doing decision theory.

When there is no data we have R given by Eqn. 4.3 and the expected utility is

$$E_{P_\delta}(U(R))) = E_\theta(U(r(\theta,\delta))).$$

Example 4.1. Consider an urn with 100 balls, colored either red or black. A ball with color θ is drawn uniformly at random and we have to predict the color. Our action predicts the color $\delta \in \{\text{red}, \text{black}\}$. Suppose the reward is

$$r(\theta,\delta) = £\mathbb{I}_{\theta=\delta}.$$

Let $\phi = \Pr(\theta = \text{black})$ be the proportion of black balls in the urn. If we have a prior $\phi \sim \pi_\phi(\cdot)$ for ϕ then our prior for θ is

$$\pi(\text{black}) = E_\phi(E(\mathbb{I}_{\theta=\text{black}}|\phi)) = E(\phi)$$

and $\pi(\text{red}) = 1 - E(\phi)$.

Suppose further the utilities are $U(0) = 0$ and $U(1) = u$ with $u > 0$. How to choose a color?

Possible rewards are $r = 0$ and $r = 1$. If $\delta = \text{black}$, we have

$$P_\delta(r = 1) = \Pr(\theta = \text{black}) = E(\phi)$$

The expected utility of choosing black is

$$\begin{aligned} E_{P_\delta}(U(r(\theta,\text{black}))) &= P_\delta(0)U(0) + P_\delta(1)U(1) \\ &= uE(\phi) \end{aligned}$$

and if $\delta' = \text{red}$ is a different action then

$$E_{P'_\delta}(U(r(\theta,\text{red}))) = uE(1-\phi).$$

The action/choice

$$\delta = \begin{cases} \text{black} & \text{if } E(\phi) > 1/2 \\ \text{red} & \text{if } E(\phi) \leq 1/2 \end{cases}$$

maximises the expected utility.

Example 4.2. In a choice between an average reward $r_0 = E_P(R))$ and a random reward $R \sim P(\cdot)$, for a strictly concave utility we take the average reward, since $E_P(U(R)) < U(E_P(R))$ by Jensen's inequality[8] so $E(U(r)) < U(r_0)$.

[8] I include an Appendix on Jensen's inequality. See Section 4.4.1 in the Appendix at the end of this chapter.

4.2 Definitions of coherence

4.2.1 Coherent belief and coherent inference

Coherent inference chooses the action $\hat{\delta}$ that maximises the expected utility, so we choose

$$\hat{\delta} = \arg\max_{\delta \in \mathcal{D}} E_{P_\delta}(U(r)).$$

Coherent inference is possible if the utility function and reward distributions exist. Suppose the observation model $p(y|\theta)$ exits. The reward distribution $P_{\delta,y}$ determined by Equation 4.2 exists if a prior distribution $\pi(\theta)$ representing our knowledge about θ exists and $p(y)$ is finite (so the posterior is proper). If this prior exists and is unique we have coherent belief.

We separate the question of existence of the posterior when the prior exists - this depends on the existence of a proper posterior - and focus on existence of the prior and the reward distribution defined in Equation 4.3. We are asking if coherent inference is possible without data.

When we elicit a prior for $\theta \in \Omega$, we assume our prior belief is coherent, so a prior representing our preferences exists. We express preference through a preference relation over sets of events in an idealised prediction problem in which the utility exists. Suppose we are interested in assigning probabilities to sets $A \in \mathcal{B}_\Omega,\ A \subseteq \Omega$ so that $\mathcal{B}_\Omega$ is a σ-field containing all the sets of interest, and $(\Omega, \mathcal{B}_\Omega, \pi)$ is a probability space with π the prior distribution. If we can do that then the prior clearly exists.

Consider two sets, $A, B \in \mathcal{B}_\sigma$ so $A, B \subseteq \Omega$, action space $\delta \in \{A, B\}$, reward $r(\theta, \delta) = \mathbb{I}_{\theta \in \delta}$ and utility $U(0) = 0,\ U(1) = u$ for $u > 0$. The reward distribution, $P_\delta = (P_\delta(0), P_\delta(1))$, for action $\delta = B$ under our prior is

$$P_\delta = (1 - \pi(B), \pi(B))$$

The expected utility $E_\theta(U(r(\theta, \delta))),\ \theta \sim \pi(\cdot)$ of the action $\delta = B$ is

$$\begin{aligned} E_\theta(U(r(\theta, \delta))) &= P_\delta(0)U(0) + P_\delta(1)U(1) \\ &\propto u\pi(B), \end{aligned}$$

so we maximise expected utility by choosing B if $\pi(B) > \pi(A)$ etc.

Write $A > B$ if $\pi(A) > \pi(B)$ and $A \sim B$ if $\pi(A) = \pi(B)$. If we *start* with a preference order $<, \sim, >$ over sets, then we seek a prior probability distribution satisfying

$$A > B \quad \Rightarrow \quad \pi(A) > \pi(B)$$

and similarly for the preference order $<, \sim$ or $>$ for every pair of sets $A, B \in \mathcal{B}_\Omega$. We say that the probability distribution $\pi(A),\ A \in \mathcal{B}_\Omega$ *expresses the preference order* in this case. If our preferences $<, \sim, >$ on sets of outcomes satisfy the Savage axioms for subjective probability (given below) then a prior expressing them exists and is unique. If the prior exists then reward distributions exist.

The expected utility determines an ordering on reward distributions $P_\delta(r)$ and $P_{\delta'}(r)$ of two actions δ, δ'. Define $>$, $<$, and $\sim$ so that

$$P_\delta > P_{\delta'} \quad \Leftrightarrow \quad E_{P_\delta}(U(r)) > E_{P_{\delta'}}(U(r)),$$

and

$$P_\delta \sim P_{\delta'} \quad \Leftrightarrow \quad E_{P_\delta}(U(r)) = E_{P_{\delta'}}(U(r)).$$

If reward distributions exist and we *start* with a preference order over reward distributions, then the inference will be coherent if there exists a utility function expressing these preferences in terms of expected utility. We need to find a utility function U that ensures

$$P > P' \quad \Rightarrow \quad E_P(U(R)) > E_{P'}(U(R)),$$

and similarly for all pairs of reward distributions P, P' generated by actions in the action space. If our preferences $<, \sim, >$ on reward distributions satisfy the Savage axioms for utility (given below) then a utility function exists and is unique. If the prior π and utility function U exist then coherent inference is possible. Coherent Bayesian inference requires coherent prior belief.

In our setting the *expected utility hypothesis* says that if we start with given prior preferences over sets $A \in \mathcal{B}_\Omega$ and preferences over reward distributions $P_\delta, \delta \in \mathcal{D}$ then a prior π and utility function U exist so that our preference over actions $\delta \in \mathcal{D}$ corresponds to order by expected utility.

4.2.2 The Ellsberg paradox

This paradox shows that peoples' preferences are in some cases inconsistent with any prior.

Suppose we have two urns, urn "A" has 50 red and 50 black balls, and urn "B" has 100 balls which are all red or black, but the proportions are not known. Bet £1000 on the color-outcome of a ball drawn from an urn so we loose £1000 if we get it wrong (and otherwise gain). The rewards are $r \in \{-1000, 1000\}$. Suppose the rewards have utilities $U(-1000) = 0$ and $U(1000) = u$, with $u > 0$ so the existence of the utility is not in question.

A number of different bets are available to us. In each bet a set of (urn, color) pairs are presented to us. We choose one of the pairs, and win if the chosen color is drawn from the chosen urn.

	Option 1 $\mathsf{color}_{\mathsf{urn}}$	Option 2 $\mathsf{color}_{\mathsf{urn}}$
Bet 1	r_A	b_A
Bet 2	r_B	b_B
Bet 3	r_A	b_B
Bet 4	r_B	b_A

For example, in Bet 4 we have the option of betting on a red ball from urn B or a black ball from urn A.

We are indifferent on Bets 1 and 2 as "red" and "black" are exchangeable labels in each urn. In Bet 3 we have a choice between red from urn A or black from urn B. We know there are 50% red balls in urn A, but we dont know how many black balls there are in urn B. They might all be black, or none of them. In a bet like Bet 3, many people opt for the fixed odds available from urn A, preferring this bet to the subjective uncertainty offered by urn B, so they take r_A and b_A in Bets 3 and 4 respectively. Call this the "standard preference".

Our preferences for each bet tell us about our unstated prior $\pi(\phi)$ for the probability $\phi = \Pr(b_B)$ to draw a black ball from urn B, assuming we are trying to make decisions according to the expected utility. 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 Bet 3, choosing r_A over b_B gives $E_\pi(\phi) < 1/2$. Choosing b_A over r_B in Bet 4 gives $E_\pi(\phi) > 1/2$. The standard preferences over bets lead to contradictory $E_\pi(\phi)$-values so a single common prior cannot exist.

4.2.3 The Allais paradox

Our preferences may be inconsistent with any utility function. In this example the reward distributions are given, so the "prior" exists.

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

$$(r_1, r_2, r_3) = (£0, £500,000, £750,000).$$

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 $p^{(D)} = (0.9, 0, 0.1)$

so for example in the first round there are two lotteries. If you choose lottery (A) you get $£500K$ guaranteed, while if you opt for (B) there is a small chance you go away with nothing, but also a 10% chance of making the big money, $£700K$.

Choices (B) and (D) maximise the expected return but that is not the same as expected utility, unless we had the identity function as a utility, which is unrealistic. Which lotteries would you choose? People commonly choose (A) for a sure thing, and (D) as there is a 1% higher chance of getting zero, but a 10% chance of getting $£700K$ instead of an 11% chance of getting $£500K$. What utility function are they using?

Set the utilities to be $U(r_1) = 0$, $U(r_2) = u$ and $U(r_3) = 1$. In terms of the row vectors (p_1, p_2, p_3) and $(0, u, 1)$, the expected utilities are $E(U) = (p_1, p_2, p_3)(0, u, 1)^T$, so

$$\begin{aligned} E(U|A) &= u & \qquad E(U|B) &= 0.1 + 0.89u \\ E(U|C) &= 0.11u & \qquad E(U|D) &= 0.1 \end{aligned}$$

Preferring (A) to (B) means

$$u > 0.1 + 0.89u \quad \Rightarrow \quad u > 10/11.$$

On the other hand preferring (D) to (C) means

$$0.1 > 0.11u \quad \Rightarrow \quad u < 10/11,$$

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.

4.3 The Savage Axioms

In De Groot, "Optimal Statistical Decisions" (1970) the Axioms are broken down into two sets. If our preferences over $A, B \in \mathcal{S}$ satisfy axioms 1-5 then the prior exists. If our preferences over reward distributions P, P' satisfy axioms 6-10 then the utility functions exists. The Savage axioms are prescriptive - if our preferences do not satisfy the axioms, then they should be altered.

4.3.1 Probability space

In the next section we give the Savage axioms. We will give them in terms of a generic probability space $(S, \mathcal{S}, \pi)$, asking essentially if this space exists. This section is notation and reminder.

Let S be a sample space and let $\mathcal{S}$ be a σ-field of sets in S satisfying $S \in \mathcal{S}$, $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S}$. If

$$A_1, A_2, ... \in \mathcal{S} \text{ then } \bigcup_{i=1}^{\infty} A_i \in \mathcal{S}.$$

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 $\pi : \mathcal{S} \to [0, 1]$ satisfying the axioms of probability: (i) $p(A) \geq 0,\ A \in \mathcal{S}$; (ii) $\pi(S) = 1$; and (iii) if $A_1, A_2, ... \in \mathcal{S}$ are disjoint then

$$\pi(\cup_i A_i) = \sum_i \pi(A_i) \qquad \text{"countably additive"}$$

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

4.3.2 Axioms of qualitative probability

We begin with the Savage axioms for qualitative probability.

Theorem 4.3. *Let S be a sample space and let $\mathcal{S}$ be a σ-field of sets in S. Let a system $>, <, \sim$ of preferences over $A, B \in \mathcal{S}$ be given. A probability distribution $\pi(A), A \in \mathcal{S}$ expressing these preferences exists and is unique if and only if the preference relations satisfy Savage Axioms 1-5 given below and in Appendix 4.4.2.*

Proof (not examinable): see De Groot, (1970).

Our main interest will be on Axioms 1-3 and existence of the prior following from Axioms 1-5.

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

Remark: if every pair of sets $A, B \in \mathcal{S}$ is ordered by preference or equal in preference, how can we have a $\geq$ symbol in the next Axiom? There is no contradiction. For example every pair of reals $x, y \in \Re$ satisfy exactly one of $x > y$, $x < y$ or $= y$ and $x \geq y$ simply means one of $>$ or $=$ hold.

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 $A_1 \cup A_2 > B_1 \cup B_2$.

Remark: If A can happen two exclusive ways, and each is more likely than corresponding events leading to B then A is more likely then B.

Axiom 3. If $A \in \mathcal{S}$ then $\emptyset \leq A$. Furthermore $\emptyset < S$.

Remark: $\emptyset < S$ is needed to avoid the trivial case where $A \sim \emptyset$ for all $A \in \mathcal{S}$. See CR-TBC p158-159 and de Groot section 6.2 especially page 72.

Remark: It follows (PS3) from Axioms 1-3 that the order is transitive (if $A \leq B$ and $B \leq C$ then $A \leq C$) and $A \leq B \Rightarrow A^c \geq B^c$. Transitivity feels important - you might suspect that transitivity is key, and hope that we could replace Axiom 2 with transitivity. This is not the case. It may be shown by counter-example that there exist transitive preferences which do not satisfy Axioms 1-3 so these axioms are already stronger than transitivity.

Example 4.4. Here is an example of the axioms in action. We saw in the Ellsberg paradox that no prior existed for standard preferences. These standard preferences cant satisfy the axioms, as otherwise a prior would exist.

If $S = \{(r,r),(r,b),(b,r),(b,b))\}$ is the set of outcomes when a ball is drawn from each urn, then

$$r_A = \{(r,r),(r,b)\}, \quad r_B = \{(r,r),(b,r)\},$$
$$b_A = \{(b,r),(b,b)\}, \quad b_B = \{(r,b),(b,b)\}.$$

Mapping onto the objects in Axiom 2, let

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

The elicited preferences preferred r_A to b_B (so $A_1 > B_1$) and b_A to r_B (so $A_2 > B_2$). By Axiom 2 we have

$$A_1 \cup A_2 > B_1 \cup B_2,$$

but $A_2 = A_1^c$ and $B_2 = B_1^c$ so $S > S$, a contradiction.

Axioms 4 and 5 and the existence proof for the prior are outside the scope of the course. I include some comments on these as an Appendix in Section 4.4.2.

4.3.3 The Savage Axioms 1-5 and the Axioms of probability

The following may be useful to clarify the relation between the first five Savage axioms and the Axioms of probability. If prior preferences satisfy the Savage axioms then $(S, \mathcal{S}, \pi)$ is a probability space for some unique π expressing the preferences and satisfying the Axioms of probability Section 4.3.1. In this setting we can work out preference relations using either set of axioms. In this section we give an example illustrating this point.

Suppose in the following the given preference relations over sets satisfy the Savage axioms.

Exercise 4.5. (PS3) Let A_1, B_1, D be sets in $\mathcal{S}$. Show that if $A_1 \cap D = B_1 \cap D = \emptyset$ then $A_1 \cup D > B_1 \cup D$ iff $A_1 > B_1$.

From the exercise we can add the same set to both sides of an inequality, if it doesn't intersect the other sets, and we can remove the same set from both sides.

Example 4.6. Suppose for two sets $A, B \in \mathcal{S}$ we have $A \subseteq B$ (as sets). It follows directly, - *ie from the Savage axioms alone* - that $A \leq B$ (in preference order). To see this suppose $A > B$. By

the exercise we can remove A from both sides to get $\emptyset > B \setminus A$ with $B \setminus A \in \mathcal{S}$. This contradicts Axiom 3.

Notice that we could alternatively prove this from the Axioms of probability. Since the Savage axioms are satisfied, $\pi(\cdot)$ expressing the preferences exists. But then $B = A \cup (B \setminus A)$ so $\pi(B) = \pi(A) + \pi(B \setminus A)$ with $\pi(B \setminus A) \geq 0$ by the Axioms of probability and hence $\pi(A) \leq \pi(B)$. However π expresses this system of preferences so $A \leq B$. [End of example]

4.3.4 Axioms of utility

Let $\mathcal{P}$ denote a set of distributions P over rewards $r \in \mathbb{R}$ where $\mathbb{R} = r_{min}, r_{min} + 1, ..., r_{max}$ or $\mathbb{R} = [r_{min}, r_{max}]$. For example, P corresponds to a choice of lottery distribution in the Allais paradox. By our assumption on the support for the reward distribution,

$$P([r_{\min}, r_{\max}]) = \Pr(r_{\min} \leq R \leq r_{\max}) = 1, \quad \text{for } R \sim P.$$

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

$$P \geq P' \quad \Leftrightarrow \quad E_P(U(r)) \geq E_{P'}(U(r)) \tag{4.5}$$

based on expected utility. If we *start* with a set of preference relations $>$, $\sim$, $<$ over $P \in \mathcal{P}$ (so fix the LHS of Equation 4.5), as we did for the Allais paradox, where we chose our preferred lottery, we can ask if there exists a utility function that satisfies the relations imposed on the RHS? We say in this case that the utility function expresses the given preferences over reward distributions.

Theorem 4.7. *There exists a utility function U which expresses our preference relations over $P \in \mathcal{P}$ if and only if our preferences satisfy Savage Axioms 6-10 given below and in Appendix 4.4.3.*

Proof (outside the syllabus): the presentation here follows Christian Robert (2007) "The Bayesian Choice". I find this easier to follow than de Groot (1970) which is itself generally very clear!

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

Axiom 6: For any two reward distributions $P, P' \in \mathcal{P}$ exactly one of the following relations must hold: $P > P'$ or $P < P'$ or $P \sim P'$.

Axiom 7: 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 8: $P' \geq P''$ if and only if

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

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

Example 4.8. We saw that for the preferences we took in the Allais paradox, no utility exists, so they must violate the axioms. The choices we made there yield preferences which violate Axiom 8. If we take

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

then the reward distributions for lotteries A-D can be written

$$\begin{array}{lcll} p^{(A)} & = & (0,1,0) & = & 0.11P' + 0.89P \\ p^{(B)} & = & (0.01,0.89,0.1) & = & 0.11P'' + 0.89P \\ p^{(C)} & = & (0.89,0.11,0) & = & 0.11P' + 0.89\tilde{P} \\ p^{(D)} & = & (0.9,0,0.1) & = & 0.11P'' + 0.89\tilde{P} \end{array}$$

Since P is the same for options (A) and (B) and we prefer (A) we have $P' > P''$ by Axiom 8. Similarly, $\tilde{P}$ is the same for (C) and (D) so a preference for (D) leads to $P' < P''$, a contradiction.

It seems Allais posed this example to Savage himself, who preferred (A) to (B) and (D) to (C)! [End of example]

Axioms 9 and 10 and the existence proof for the utility are outside the scope of the course. I include some comments on these as an Appendix in Section 4.4.3.

4.3.5 Conclusions

What should be make of the Savage axioms? Should we take the Savage axioms as prescriptive? Well, if they dont hold then we dont have coherent belief, or we cant make coherent inference, or both. By the relation between loss and utility, coherent inference is inference based on Decision Theory and that seems desirable. Do the paradoxes lead us to dismiss these notions of coherence? Allais in particular has motivated research on how to generalise the idea of utility.

However, more fundamentally the SA are a careful answer to a question we dont often have to ask. In Bayesian inference, existence of the prior is not usually an issue and there is some natural loss function for the task in hand. We do use preference relations over sets to help us elicit a prior (I used the example "is $p > 0.99$ or is $p \leq 0.99$" in the check-list) but we don't consciously elicit a prior by expressing preferences for every pair of sets of events $\{\theta \in A\},\ \{\theta \in B\}$. When we derive a prior from a simple physical model for the process generating θ (as in Section 1.4) we largely avoid this sort of consideration.

We very often have a candidate prior, perhaps with hyper-parameters, and we choose the hyper-parameters to make the prior representative of prior knowledge. We dont have enough certainty about our prior information to be able to express preferences over every pair of sets in the σ-field. For a more interesting question in practice try "is this prior representative of available information?".

4.4 Appendices

4.4.1 Appendix for Section 4.2.1 - Jensen's inequality

Proposition 4.9. *If $f(x)$ is a concave function and $E(X), E(f(X)) < \infty$ then*

$$E(f(X)) \leq f(E(X)).$$

Proof (scalar $X \in \mathbb{R}$): if $f : \Re \to \Re$ is concave if for all $x, x_0 \in \Re$,

$$f(x) \leq f(x_0) + (x - x_0)f'(x_0)$$

ie, the function lies below its tangent taken at any point $(x_0, f(x_0))$ on the curve. Since this holds for all $x, x_0 \in \Re$, take $x = X$ and $x_0 = E(X)$ and expectations on both sides to get the inequality.

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 simply a constant random variable (so not $X = c$ with probability one) then

$$E(f(X)) < f(E(X)).$$

4.4.2 Appendix for Section 4.3.2 - Axioms of qualitative probability

The following material, included for completeness, is outside the scope of the course.

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

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

The last axiom assigns a probability to any set of outcomes. I state this informally here to give the idea. de Groot is explicit about how the space is extended to accommodate a uniform random variable.

Axiom 5. For each $p \in [0, 1]$ there is $B_p \in \mathcal{S}$ satisfying $\pi(B_p) = p$.

Axiom 5 refers to π, which may not exist! The trick here is to add a set of events B_p with known probabilities to the set of possible outcomes. For example if the original sample space was $\{rain,\ sunshine\}$ then the extended outcomes are events like $\{rain\} \cap \{X \leq p\}$, with $X \sim U(0, 1)$. We know $\pi(\{X \leq p\}) = p$ so there is an event in the extended space for each probability $p \in [0, 1]$.

The existence and uniqueness proof uses these extra events. It replaces 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 $\mathcal{S}$. However A1-A4 must now hold for sets in the new σ-algebra. We can then fix $\pi(A), A \in \mathcal{S}$ by finding a matching event with known probability. Given A, we find B_p satisfying $A \sim B_p$ and set $\pi(A) = p$. Axioms 1-4 then ensure the probability space $(S, \mathcal{S}, \pi)$ defined in this way will satisfy the axioms of probability.

4.4.3 Appendix for Section 4.3.4 - Axioms of utility

The following material, included for completeness, is outside the scope of the course.

I omit Axioms 9 and 10. See CR-TBC/deGroot for detail. In Axiom 9 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 A6-9 apply, this is relatively easy. A candidate utility function is now given, and the expectation $E(U|P)$ is now well-defined for general P.

Axiom 10 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 the same expected utility but having just two possible outcomes, so $\tilde{P}(r_1) + \tilde{P}(r_2) = 1$. Using A6-9 it can be shown that the equivalent lottery is

$$\tilde{P}(r) \sim \frac{E(U|P) - U(r_1)}{U(r_2) - U(r_1)} \mathbb{I}_{r=r_2} + \frac{U(r_2) - E(U|P)}{U(r_2) - U(r_1)} \mathbb{I}_{r=r_1}$$

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

5 Exchangeability

The main point of this chapter is made in the final section: exchangeability plays a similar role in subjective Bayesian inference to the role that repeated trials play in defining probability in Frequentist inference. If the data are part of an infinite exchangeable sequence then there *is* a generative model $p(y|\theta)p(\theta)$ and in particular a prior $\pi(\theta)$ exists.

Exchangeability and the Savage axioms make a pair. They both give sufficient conditions for a prior to exist. Exchangeability just says roughly speaking "if the data are exchangeable then a prior exists". The Savage axioms say "if your preferences are coherent then a prior exists".

In preparing this section I found the following references particularly useful. For the big picture see *Steffen Lauritzen (2007) "Exchangeability and de Finetti's Theorem", Oxford Graduate Lecture Series, web-link.* For a concise statement of the proof see *David A. Stephens (2006) "The de Finetti 0-1 representation", Statistical Theory II lectures, Imperial College, web-link* and for the original paper with this version of the proof see *David Heath & William Sudderth (1976) "De Finetti's Theorem on Exchangeable Variables", The American Statistician, 30:4, 188-189, web-link.*

5.1 Exchangeability and Infinite Exchangeable Sequences

5.1.1 Exchangeability in finite sequences

Definition 5.1. Consider a finite sequence $X_i \in \mathcal{X}, i = 1, ..., n$ of $n \geq 1$ random variables. The random variables are *exchangeable* if their joint distribution is unchanged by permutation of the indices, so

$$(X_1, X_2, ..., X_n) \sim (X_{\sigma_1}, X_{\sigma_2}, ..., X_{\sigma_n})$$

for every permutation $\sigma \in \mathcal{P}_{[n]},\ \sigma = (\sigma_1, \sigma_2, ..., \sigma_n)$ of the numbers $[n] = \{1, 2, ..., n\}$.

Any realisation of $(X_1, X_2, ..., X_n)$ is equally likely to be seen in any order. For example, suppose $X_1, ..., X_n$ are discrete and have joint probability mass function (PMF) $p_{1:n}(x_1, ..., x_n)$ on $\mathcal{X}^n$ so

$$P(X_1 = x_1, ..., X_n = x_n) = p_{1:n}(x_1, ..., x_m).$$

The subscript $1:n = (1, 2, ..., n)$ emphasises the variables present and the order in which they are taken so

$$P(X_{\sigma_1} = x_1, ..., X_{\sigma_n} = x_n) = p_\sigma(x_1, ..., x_n).$$

Random variables $X_1, ..., X_n$ are then *exchangeable if and only if* for every $(x_1, ..., x_n) \in \mathcal{X}^n$,

$$p_{1:n}(x_1, ..., x_n) = p_{1:n}(x_{\sigma_1}, ..., x_{\sigma_n}), \quad \text{for each } \sigma \in \mathcal{P}_{[n]}.$$

The same condition holds if instead $X_1, ..., X_n$ were continuous exchangeable random variables with joint density $p_{1:n}(x_1, ..., x_n)$ on $\mathcal{X}^n$.

Example 5.2. If $\{X_i\}_{i=1}^n$ are iid they are exchangeable, but exchangeable variables need not be independent.

For $i = 1, ..., n$ let $X_i \in \{green = 0, red = 1\}$ be the sequence of colors of n balls sampled without replacement from an urn containing N balls, exactly K of which are red. Let $x = (x_1, ..., x_n)$ for $x \in \{0,\ 1\}^n$ and $k(x) = \sum_i x_i$. I redid this bit to avoid assuming exchangeability in deriving this expression (and also fix a typo). For $k(x) = k$ let $i_1, ..., i_k$ give the indices of the red-ball selections

so $x_{i_a} = 1,\ a = 1, ..., k$. Similarly $x_{j_b} = 0,\ b = 1, ..., n-k$ for green. When the a'th red ball was selected there were $K-(a-1)$ reds among $N-(i_a-1)$ balls for a factor $(K-a+1)/(N-i_a+1)$ in the probability for the sequence. Gathering up the factors,

$$p_{1:n}(x_1, ..., x_n) = \prod_{a=1}^{k} \frac{K-(a-1)}{N-(i_a-1)} \times \prod_{b=1}^{n-k} \frac{N-K-(b-1)}{N-(j_b-1)} \tag{5.1}$$

$$\frac{K!}{(K-k)!} \times \frac{(N-K)!}{(N-K-(n-k))!} \times \frac{(N-n)!}{N!} \tag{5.2}$$

$$= \binom{n}{k}^{-1} \times \binom{K}{k}\binom{N-K}{n-k} / \binom{N}{n}, \tag{5.3}$$

$$= \Pr(X_1 = x_1, ..., X_n = x_n | k(X) = k) \Pr(k(X) = k), \tag{5.4}$$

when $k(x) = k$, and we identified the distribution of the number of red balls $k(X)$ as hypergeometric in the last step. Since $k(x)$ is the same for $x = (x_1, ..., x_n)$ taken in any order, $p_{1:n}(x_1, ..., x_n) = p_{1:n}(x_{\sigma_1}, ..., x_{\sigma_n})$ so $X_1, ..., X_n$ are exchangeable. The variables are not independent since for example,

$$P(X_2 = 1 | X_1 = 1) = \frac{K-1}{N-1}$$

but

$$P(X_2 = 1 | X_1 = 0) = \frac{K}{N-1}.$$

Exercise 5.3. Suppose $K \geq 3$ so there are at least three red balls in the urn. Show that the probability the last three balls are red, so $X_{n-2} = 1, X_{n-1} = 1$ and $X_n = 1$, is

$$p_{n-2:n}(1,1,1) = \frac{K(K-1)(K-2)}{N(N-1)(N-2)}.$$

ANS: this is simply the probability the first three are red.

All marginals of exchangeable random variables are the same. The proof is left as an exercise.

5.1.2 Infinite Exchangeable sequences

Definition 5.4. An *infinite exchangeable sequence* of random variables $\{X_i\}_{i=1}^{\infty}$ is an infinite sequence of random variables in which $X_1, X_2, ..., X_n$ are exchangeable for every $n \geq 1$.

Every subset of the infinite sequence of random variables is then exchangeable, and any realisation is equally likely to be seen in any order.

Example 5.5. An iid sequence of binary random variables $X_i \sim \mathsf{Bern}(\theta),\ i = 1, 2, 3, ...$ is clearly an infinite exchangeable sequence since for any $n \geq 1$ we have

$$p_{1:n}(x_1, ..., x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

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

5.1.3 Exchangeability in a Hierarchical model

This is an example where the "sequence" structure is not obvious. For $n = 1, 2, ...$ let $0_n = (0, ..., 0)$ be a vector of n zeros. Let $\Sigma^{(n)}$ be an $n \times n$ covariance matrix with unit variance $\Sigma^{(n)}_{i,i} = 1,\ i = 1, ..., n$ and for $0 \leq \rho < 1$ and $i, j = 1, ..., n,\ i \neq j$, equal positive covariances $\Sigma^{(n)}_{i,j} = \rho$. If $x = (x_1, ..., x_n)$, write $p_{1:n}(x_1, ..., x_n) = p_{1:n}(x)$ and suppose

$$p_{1:n}(x) = N(x; 0_n, \Sigma^{(n)}) \tag{5.5}$$

with $N(x; 0_n, \Sigma^{(n)})$ the multivariate normal density at $x \in \mathbb{R}^n$. We now show that there is an infinite exchangeable sequence $X_1, X_2, X_3, ...$ with marginals given by Equation 5.5 for every n.

We first check that $p_{1:n}(x) = p_{1:n}(x_\sigma)$ for $\sigma \in \mathcal{P}_n$ with $x_\sigma = (x_{\sigma_1}, ..., x_{\sigma_n})$. We have

$$\begin{aligned} p_{1:n}(x_{\sigma_1}, ..., x_{\sigma_n}) &= N(x_\sigma; 0_n, \Sigma^{(n)}) \\ &= N(x; 0_n, \Sigma^{(n)}) \\ &= p_{1:n}(x_1, ..., x_n) \end{aligned}$$

The second line follows because all the variances and covariances are equal so the order in which $(x_1, ..., x_n)$ are presented does not matter. You can check this if you like.

Exercise 5.6. (optional) Let P_σ be the permutation matrix satisfying $x_\sigma = P_\sigma x$. The transpose of a permutation matrix is its inverse, $P_\sigma^T = P_\sigma^{-1}$. Show that $x_\sigma^T \Sigma^{-1} x_\sigma = x^T (P_\sigma^{-1} \Sigma P_\sigma)^{-1} x$ and explain why $P_\sigma^{-1} \Sigma P_\sigma = \Sigma$ for $\Sigma = \Sigma^{(n)}$ as above. Hence show that $x_\sigma^T \Sigma^{-1} x_\sigma = x^T \Sigma^{-1} x$ and $N(x_\sigma; 0_n, \Sigma^{(n)}) = N(x; 0_n, \Sigma^{(n)})$.

Are we done? We are not. We need to show that there actually exists an infinite sequence of random variables $X_1, X_2, X_3, ...$ with marginals given by Equation 5.5 so that

$$(X_1, ..., X_n) \sim N(0_n, \Sigma^{(n)}) \tag{5.6}$$

for every $n \geq 1$. We can prove this by construction. Simulate $\theta \sim N(0, \rho)$ and set

$$X_i = \theta + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, 1 - \rho), \quad \text{iid for } i = 1, 2, 3....$$

These X's are distributed as in Equation 5.6. We only need to check the means and covariances as they are linear combinations of normal random variables so jointly normal. Checking the moments, we have $E(X_i) = 0$, $\mathsf{var}(X_i) = 1$ and $\mathsf{cov}(X_i, X_j) = \rho$ so indeed $(X_1, ..., X_n) \sim N(0_n, \Sigma^{(n)})$ for every n.

This construction makes it obvious that $p_{1:n}(x) = p_{1:n}(x_\sigma)$ for all $x \in \mathbb{R}^n,\ \sigma \in \mathcal{P}_n$: the X's are independent conditional on θ so any realisation is equally likely to appear in any order. Explicitly, integrating over θ, we have

$$p_{1:n}(x_1, ..., x_n) = \int_{-\infty}^{\infty} \prod_{i=1}^{n} N(x_i; \theta, 1 - \rho) N(\theta; 0, \rho)\, d\theta$$

for the joint density of $X_1, ..., X_n$, and this alternative form must equal $N(x; 0_n, \Sigma^{(n)})$ for all $x \in \mathbb{R}^n$. Now it is clear that the density is invariant under $x \to x_\sigma$ as we just shuffle terms in the product when we permute the indices.

Exercise 5.7. The covariance matrix $\Sigma^{(n)}$ is positive definite for every $n \geq 1$. Show that the parameter ρ must be non-negative. To show this verify that

$$v_1 = (1, 1, 1, ..., 1),\ v_2 = (1, -1, 0, 0, ..., 0),\ v_3 = (0, 1, -1, 0, ..., 0),\ ...\ , v_n = (0, ..., 0, 1, -1)$$

are eigenvectors in $\Sigma^{(n)} v_i = \lambda_i v_i,\ i = 1, ..., n$ with eigenvalues respectively

$$\lambda_1 = 1 + (n-1)\rho$$

and $\lambda_i = 1 - \rho,\ i = 2, ..., n$. Hence show that $-1/(n-1) < \rho < 1$ is necessary and sufficient for positive definite $\Sigma^{(n)}$ at any fixed n, so $0 \leq \rho < 1$ is necessary and sufficient for this to hold for all $n \geq 1$.

5.1.4 Marginal consistency

A family of probability distributions is marginally consistent if every marginal of every distribution in the set is also in the family.

Definition 5.8. (binary case) Let $\mathcal{O}_{[n]}$ be the set of all nonempty subsets of $[n]$. The sets $o \in \mathcal{O}_{[n]}$ are ordered, so if $o = (i_1, ..., i_m)$ for some $1 \leq m \leq n$ then $i_1 < i_2 < ... < i_m$. Suppose $x_i \in \{0, 1\},\ i = 1, ..., n$. Let $\mathcal{F}_{[n]} = \cup_{o \in \mathcal{O}_{[n]}} \{p_o(\cdot)\}$ be a given family of probability mass functions. Let $\bar{o} = [n] \setminus o$ be the complement of o and

$$q_o(x_{o_1}, ..., x_{o_m}) = \sum_{x_{\bar{o}_1} \in \{0,1\}} ... \sum_{x_{\bar{o}_{n-m}} \in \{0,1\}} p_{1:n}(x_1, ..., x_n)$$

be the marginal PMF for the random variables $X_o^{([n])}$ if $(X_1^{([n])}, ..., X_n^{([n])}) \sim p_{1:n}$. The family $\mathcal{F}_{[n]}$ is marginally consistent if and only if $q_o(x_o) = p_o(x_o)$ for every $x_o \in \{0, 1\}^m$ and all $o \in \mathcal{O}_{[n]}$.

Remark 5.9. Marginally consistent families of distributions are sometimes called projective families in the literature.

Example 5.10. Suppose $\mathcal{F}_{[n]}$ is marginally consistent for every n. Then we must have

$$p_{1:n}(x_1, ..., x_n) = p_{1:(n+1)}(x_1, ..., x_n, 0) + p_{1:(n+1)}(x_1, ..., x_n, 1).$$

[EOE]

Notice the superscript $[n]$ in $(X_1^{([n])}, ..., X_n^{([n])}) \sim p_{1:n}$. The point here is that the random variables $X^{(o)} = (X_{o_1}^{(o)}, ..., X_{o_m}^{(o)}),\ X^{(o)} \sim p_o$ need not be the same as the corresponding components $X_o^{([n])}$ of $X_{1:n}^{([n])} \sim p_{1:n}$ since $X_o^{([n])} \sim q_o$. However, if $q_o(x_o) = p_o(x_o)$ for all x_o and every o then $X^{(o)} \sim X_o^{([n])}$ and we can just take $X = X^{([n])}$ and drop the superscript.

If we *start* with a collection of binary random variables $X_1, X_2, X_3, ...$ and we *define*

$$p_{1:n}(x_1, ..., x_n) = \Pr(X_1 = x_1, ..., X_n = x_n)$$

for each $n \geq 1$ then the probability mass functions $p_{1:n},\ n \geq 1$ are always marginally consistent - this follows from the axioms of probability, and holds for any sequence $X_1, X_2, X_3, ...$ of random variables, whether or not they are exchangeable.

However, if we *start* with a collection of arbitrary probability mass functions there may not exist a sequence of random variables with the stated marginals. If the probability mass functions are not

marginally consistent there cant be such a sequence. When we wrote down $p_{1:n}$ in the Hierarchical Normal example, Section 5.1.3 we showed that these were the marginals of an infinite sequence of random variables so we directly addressed this question. This comes up quite often in prior elicitation when the prior we choose depends on the number of variables.

Example 5.11. Here is an example (an Ising model on a complete graph) where the probability mass functions are invariant under permutation of arguments at every $n \geq 1$ but no infinite sequence with the stated marginals can possibly exist as they are not marginally consistent.

Suppose that $x = (x_1, ..., x_n)$ with $x \in \{0,1\}^n$. Let $k(x) = \sum_{i=1}^n x_i$ and

$$p_{1:n}(x) = c_n 2^{-k(x)(n-k(x))},$$

with c_n a normalising constant with $c_3 = 2/7$ and $c_4 = 8/27$.

Consider the sequence of probability mass functions $p_{1:n}$, $n \geq 1$. Clearly $p_{1:n}(x) = p_{1:n}(x_\sigma)$, $\sigma \in \mathcal{P}_n$ for every $n \geq 1$ as $k(x) = k(x_\sigma)$, so if a sequence $X_1, X_2, X_3, ...$ exists with PMF's $\Pr\{X_1 = x_1, ..., X_n = x_n\} = p_{1:n}(x_1, ..., x_n)$ then it would be an IES as $X_1, ..., X_n$ would be exchangeable for each $n \geq 1$. However, these PMF's would have to be marginally consistent and they are not. For example, we must have

$$p_{1:3}(0,0,0) = p_{1:4}(0,0,0,0) + p_{1:4}(0,0,0,1).$$

However, $p_{1:3}(0,0,0) = 2/7$ but

$$p_{1:4}(0,0,0,0) + p_{1:4}(0,0,0,1) = 8/27(1 + 1/8) = 1/3.$$

These are not equal, so the given PMF's are not marginally consistent, and hence there cant exist random variables with distributions $(X_1, ..., X_4) \sim p_{1:4}$ with $(X_1, ..., X_3) \sim p_{1:3}$.

Exercise 5.12. For $x = (x_1, ..., x_n)$ with $x \in \{0,1\}^n$, show

$$p_{1:n}(x) = \frac{k(x)!(n-k(x))!}{(n+1)!}, \quad n \geq 1$$

is marginally consistent and determines an IES.
ANS: We recognise the (inverse of the) normalising constant of a Beta(α, β) distribution with $\alpha = k + 1$ and $\beta = n - k + 1$ so

$$\begin{aligned} p_{1:n}(x) &= \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta \\ &= \int_0^1 \theta^{k(x)}(1-\theta)^{n-k(x)} d\theta \\ &= \int_0^1 \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} d\theta \end{aligned}$$

and this gives a generative model $\theta \sim U(0,1)$, $X_i \sim \text{Bernoulli}(\theta)$, $i = 1, 2, 3, ...$, for the X's. This generates an infinite sequence of exchangeable random variables with the stated marginals (the "if" direction of de Finetti below). It follows by the axioms of probability that $p_{1:n}$ are marginally consistent and it is also evident from the last integral expression (see PS3). This is a Polya urn (Section 5.2.1) with $b = 1$ black balls, $w = 1$ white and $A = 1$ added at each step.

5.2 de Finetti's Theorem

This is a theorem characterising infinite exchangeable sequences (IES's).

Example 5.13. We saw in the hierarchical normal in Section 5.1.3 that if

$$(X_1, ..., X_n) \sim N(0_n, \Sigma^{(n)}),$$

for $\Sigma^{(n)}$ a covariance matrix with constant diagonal equal 1 and off diagonal equal $\rho \in [0, 1)$, and every $n \geq 1$ then $X_1, X_2, X_3, ...$ is an IES (Infinite Exchangeable Sequence). Permutation invariance of the density was obvious once we saw that

$$N(x; 0_n, \Sigma^{(n)}) = \int_{-\infty}^{\infty} \prod_{i=1}^{n} N(x_i; \theta, 1-\rho) N(\theta; 0, \rho)\, d\theta,$$

so this infinite exchangeable sequence was actually distributed as a continuous mixture of iid random variables.

de Finetti's Theorem says this representation exists for all IES's. It is given here for the case of IES's of binary random variables, but holds much more generally, as the example above illustrates.

Theorem 5.14. *Let $X_1, X_2, ...X_n, ...$ be an infinite sequence of binary random variables with probability mass functions*

$$p_{1:n}(x_1, ..., x_n) = \Pr(X_1 = x_1, ..., X_n = x_n),\ n \geq 1.$$

The sequence is exchangeable if and only if there exists a CDF $F(\theta)$ on $[0, 1]$ such that

$$p_{1:n}(x_1, ..., x_n) = \int_0^1 \left[\prod_{i=1}^{n} p(x_i|\theta) \right] dF(\theta) \tag{5.7}$$

with

$$F(\theta) = \Pr(\Theta \leq \theta) \quad \textit{where} \quad \Theta = \lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} X_i.$$

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

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

In short, "an infinite exchangeable sequence is distributed as a mixture of iid random variables". If $X_1 = x_1, ..., X_n = x_n$ are a realisation of the first n random variables in an infinite exchangeable sequence then there exists a generative model for the X's in which

$$\Theta \sim F$$
$$X_i|\Theta = \theta \sim p(\cdot|\theta), \quad \text{iid for } i = 1, ..., n.$$

The theorem says F and $p(\cdot|\theta)$ must exist to make this hold. It extends to cover infinite exchangeable sequences of random vectors (for example, if X_i are continuous multivariate random variables) with a multivariate parameter θ.

The expression $dF(\theta)$ may be off-putting but it is useful to express the generality of the result. For example, if $F(\theta)$ is the CDF of a probability density $\pi(\cdot)$ then $dF(\theta) = \pi(\theta)d\theta$, and the de Finetti representation is just

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

In general F is just some unknown distribution which puts probability mass on measurable subsets of parameter space Ω.

5.2.1 The Polya urn

This is a process generating an infinite exchangeable sequence and illustrates de Finetti in action.

Consider an urn containing b black and w white balls. We sample a ball and return it to the urn with an additional A balls of same color. For $i = 1, 2, 3, ...$ let $X_i = 1$ if the ith ball is black and $X_i = 0$ if white.

The random variables variables $X_1, X_2, X_3, ...$ are not independent but they are exchangeable. Consider the probability for sequences $0, 0, 1, 1$ and $1, 1, 0, 0$,

$$p_{1:4}(0, 0, 1, 1) = \frac{w}{(w+b)} \frac{w+A}{(w+b+A)} \frac{b}{(w+b+2A)} \frac{b+A}{(w+b+3A)}$$

$$p_{1:4}(1, 1, 0, 0) = \frac{b}{(w+b)} \frac{b+A}{(w+b+A)} \frac{w}{(w+b+2A)} \frac{w+A}{(w+b+3A)}$$

Permuting arrivals just shuffles the numerator factors, they all still appear exactly once. This clearly generalises to sequences of arbitrary length.

Exercise 5.15. (i) Let $k(x) = \sum_{i=1}^n x_i$. Show for the Polya urn that

$$p_{1:n}(x_1, ..., x_n) = \frac{\prod_{i=0}^{k(x)-1}(b + iA) \prod_{j=0}^{n-k(x)-1}(w + jA)}{\prod_{i=0}^{n-1} b + w + iA}.$$

and write this in terms of Γ-functions using $x\Gamma(x) = \Gamma(x + 1)$.

(ii) Use this result to show that $X_1, X_2, X_3, ...$ is an infinite exchangeable sequence.

(iii) Find the de Finetti representation for $p_{1:n}$. Show that

$$p_{1:n}(x_1, ..., x_n) = \int \prod_i \theta^{x_i}(1 - \theta)^{1-x_i} \mathsf{Beta}(\theta; b/A, w/A)d\theta.$$

Hint: Do the integral in (iii) above and check it equals the expression in (i).

5.2.2 Proof of de Finetti's Theorem

Equation 5.7 is clearly sufficient for $X_1, X_2, X_3, ...$ to be an IES from the permutation symmetry on the RHS. Permuting the indices just reorders the factors in the product.

We need to show the more interesting fact that if $X_1, X_2, X_3, ...$ is an IES then Equation 5.7 holds. Our approach will be to take n of the first N outcomes in the sequence, write down the joint pmf of $X_1, ..., X_n$ and show that this converges to the RHS of Equation 5.7 as $N \to \infty$.

Proposition 5.16. *Let*

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

and let r, s *be two integers* $0 \leq r \leq s \leq N$. *If* $x_1 = 1, x_2 = 1, ..., x_r = 1, x_{r+1} = 0, x_{r+2} = 0, ..., x_n = 0$ *so we have* r *1's and* $n - r$ *0's, then*

$$\Pr(S_n = r) = \binom{n}{r} p_{1:n}(x_1, ..., x_n) \tag{5.8}$$

and

$$\Pr(S_n = r | S_N = s) = \binom{s}{r}\binom{N-s}{n-r} / \binom{N}{n}. \tag{5.9}$$

Proof: There are n choose r distinct permutations of r 1's and $n - r$ 0's giving $S_n = r$ and they all have the same probability as we can assume exchangeability in this direction. This gives Equation 5.8. For Equation 5.9, again, by the exchangeability of $X_1, ..., X_N$, if we have $S_N = s$ 1's in the first N of the X's then they are equally likely to appear in any order, so the conditional distribution of $S_n = r$ given $S_N = s$ 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. This is hypergeometric as above. [End of proof of Proposition]

When $S_n = r$ the sequence $X_1, ..., X_N$ contains at least $n - r$ 0's (in $X_1, ..., X_n$) so

$$\begin{aligned}\Pr(S_n = r) &= \sum_{s=r}^{N-(n-r)} \Pr(S_n = r | S_N = s) \Pr(S_N = s) \\ &= \sum_{s=r}^{N-(n-r)} \Pr(S_n = r | S_N/N = \theta(s)) \Pr(S_N/N = \theta(s)),\end{aligned} \tag{5.10}$$

where $\theta(s) \equiv s/N$. This is true for every $N \geq n$. If we take the limit $N \to \infty$ on the RHS in Equation 5.10 with n and r fixed, the sum will converge to an integral and give us an integral like Equation 5.7. That is the approach we now take.

Define a random variable $\Theta_N \sim S_N/N$ with the same distribution as S_N/N. This takes values $\Theta_N \in \{0, 1/N, 2/N, ..., 1\}$ and has a CDF F_N with

$$\begin{aligned}F_N(\theta) &= \Pr(\Theta_N \leq \theta), \\ &= \sum_{s=0}^{N} \Pr(S_N = N\theta(s)) \mathbb{I}_{\theta(s) \leq \theta},\end{aligned} \tag{5.11}$$

for $\theta \in [0, 1]$. We now give a "density" f_N for the distribution F_N and express $\Pr(S_n = r)$ in terms of f_N.

Definition 5.17. For $0 < \tau < 1$ denote by $\delta_\tau(\theta)$ the Dirac delta-function giving the density of a distribution δ_τ putting a unit point mass at $\theta = \tau$. This distribution is defined by its action in integrals: if $g(\theta)$ is a function continuous at $\theta = \tau$, then

$$\int_0^1 g(\theta)\delta_\tau(\theta)d\theta = g(\tau).$$

As we take θ increasing past $\theta(s)$ for some s in $0, 1, ..., N$ in Equation 5.11, the CDF $F_N(\theta)$ jumps by $\Pr(S_N = N\theta(s))$ as one more term is added to the sum. The CDF F_N is not differentiable at

discontinuities, but the "density"

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

assigns correct probability to sets $[0, \tau]$ so that

$$F_N(\tau) = \int_0^\tau f_N(\theta)d\theta.$$

The discontinuities at $\theta(s)$ in F_N are associated with point masses $\Pr(S_N = N\theta)\delta_{\theta(s)}$ in f_N.

Exercise 5.18. Show that for $0 \le a < b \le 1$ we have $\int_a^b f_N(\theta)d\theta = F_N(b) - F_N(a)$.

Proposition 5.19. *Let $dF_N(\theta) \equiv f_N(\theta)d\theta$. For $0 \le r \le N$ it holds that*

$$\Pr(S_n = r) = \int_{r/N}^{1-(n-r)/N} \Pr(S_n = r|S_N = N\theta)dF_N(\theta). \tag{5.12}$$

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

$$\begin{aligned}
\int_{r/N}^{1-(n-r)/N} \Pr(S_n = r|S_N = N\theta)dF_N(\theta) &= \int_0^1 \mathbb{I}_{r \le N\theta \le N-(n-r)} \Pr(S_n = r|S_N = N\theta)f_N(\theta)d\theta \\
&= \int_0^1 g_N(\theta)f_N(\theta)d\theta \\
&= \sum_{s=0}^{N} \int_0^1 g_N(\theta) \Pr(S_N = N\theta)\delta_{\theta(s)}(\theta)d\theta \\
&= \sum_{s=0}^{N} g_N(\theta(s)) \Pr(S_N = N\theta(s)) \\
&= \sum_{s=r}^{N-(n-r)} \Pr(S_n = r|S_N = N\theta(s)) \Pr(S_N = N\theta(s)) \\
&= \Pr(S_n = r)
\end{aligned}$$

[End of proof of proposition]

Proposition 5.20. *There exists a distribution $F(\theta)$ such that*

$$\Pr(S_n = r) = \binom{n}{r} \int_0^1 \theta^r(1-\theta)^{n-r}dF(\theta), \tag{5.13}$$

for every r, n satisfying $0 \le r \le n$.

Proof: The expression we have for $\Pr(S_n = r)$ in Proposition 5.19 holds for every $N \ge n$. Since $\Pr(S_n = r)$ does not depend on N, the RHS of Equation 5.12 does not depend on N and so trivially its limit as $N \to \infty$ exists and is equal $\Pr(S_n = r)$. If the limits of F_N and $\Pr(S_n = r|S_N = N\theta)$ exist then we can substitute them in the integral. At fixed θ, the hypergeometric probability $\Pr(S_n = r|S_N = N\theta)$ converges uniformly as a function of θ to the binomial probability,

$$\lim_{N\to\infty} \Pr(S_n = r|S_N = N\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}.$$

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 n from a large population of size N.

Helly's Theorem (Feller (1966) *Probability Theory and Applications* II) gives us "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 on some diverging subsequence $N_1 \leq N_2 \leq N_3 \leq ...$ of N-values. On the subsequence,

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

where $\Theta \sim F$ and $\Theta = \lim_{N\to\infty} N^{-1}S_N$. [End of proof of proposition]

Equating the RHS of Equations 5.8 and 5.13 for $\Pr(S_n = r)$ and cancelling n choose r we obtain

$$p_{1:n}(x_1, ..., x_n) = \int_0^1 \theta^r (1-\theta)^{n-r} dF(\theta)$$

where $r = \sum_{i=1}^n x_i$, the result of de Finetti given in Equation 5.7.

Helly's Theorem does not identify a unique limit $F_N \to F$. However a distribution on a bounded interval is uniquely determined by its moments, and since Equation 5.7 fixes all the moments, $E(\theta^n), n = 1, 2, ...$, of F when we take $r = n$ and $n = 1, 2, ...$ it follows that F is unique.

5.3 Bayesian inference

We come to the point of this Chapter.

If the *data* $X_1 = x_1, ..., X_n = x_n$ are a realisation of n samples in an infinite exchangeable sequence then there exists a generative model

$$\begin{aligned} \Theta &\sim F \\ X_i|\Theta = \theta &\sim p(\cdot|\theta), \quad \text{iid for } i = 1, ..., n \end{aligned}$$

for the data. These distributions all exist by de Finetti. If $x = (x_1, ..., x_n)$, the likelihood is

$$p(x|\theta) = \prod_i p(x_i|\theta).$$

Here F is the distribution giving the true generative model for θ, and we should use this as our prior. It could be called "nature's prior". Our own prior π may not coincide with the true generative model F. However, de Finetti gives the form for the prior predictive distribution of the data

$$p_{1:n}(x_1, ..., x_n) = \int p(x_1, ..., x_n|\theta) dF(\theta).$$

This is of course the marginal likelihood. Suppose we have some priors $\pi(\theta|m),\ m \in \mathcal{M}$ and we carry out model selection using the marginal likelihood

$$p(x|M = m) = \int p(x|\theta)\pi(\theta|M = m)d\theta.$$

When we choose a model $M = m$, we are forming an estimate $\pi(\theta|M = m)$ for F, the unknown true generative model for the parameter, so we are estimating the prior! We will come back to this when we look at non-parametric Bayesian inference in Chapter 8.

We also get an expression for the posterior in terms of F. Suppose we have seen $x_{1:m} = (x_1, ..., x_m)$ and we wish to predict $x_{m+1:n} = (x_{m+1}, ..., x_n)$. The posterior predictive distribution is

$$\begin{aligned} p(x_{m+1:n}|x_{1:m}) &= p(x_{1:n})/p(x_{1:m}) \\ &= \int p(x_{m+1:n}|\theta)\frac{p(x_{1:m}|\theta)dF(\theta)}{p(x_{1:m})} \\ &= \int p(x_{m+1:n}|\theta)d\tilde{F}(\theta). \end{aligned}$$

After we get the data $X_1 = x_1, ...X_m = x_m$ there exists a generative model

$$\begin{aligned} \Theta|X_{1:m} &\sim \tilde{F}(\theta) \\ X_i|\Theta = \theta &\sim p(\cdot|\theta), \quad \text{iid for } i = m+1, ..., n. \end{aligned}$$

Here

$$d\tilde{F}(\theta) \propto p(x_1, ..., x_m|\theta)dF(\theta)$$

is the updated true generative model for the parameter $\Theta|X_{1:m}$, or in other words, the posterior. de Finetti tells us that Bayesian inference is possible in this exchangeable setting.

One objection is that nature's prior $dF(\theta)$ may exist but our prior $\pi(d\theta)$ doesnt match it. However, the same could be said of the likelihood. We do our best in our statistical modelling to match nature's generative model for the parameter and data.

6 Model averaging

We can allow for uncertainty in which model is the right model when we estimate a function $h(\theta)$ of the parameters θ. This full quantification of uncertainty is one of the strengths of Bayesian inference, though it comes at the cost of some heavy computation. We can also get a posterior distribution over models so we don't have to do model selection, we can instead give an HPD confidence-set for models!

6.1 Model averaging distributions and decisions

6.1.1 Distributions over models and parameters

Recall the setup for model selection in Section 1.3.5 where we introduced model selection. Suppose $\theta \in \Omega_m$ when the model is $m \in \mathcal{M}$. So far we have worked with $\pi(\theta|m)$, the prior for θ when the unknown true model M is model m, $p(y|\theta, m)$, the observation model for y given θ,

$$p(y|m) = \int_{\Omega_m} p(y|\theta, m)p(\theta|m)d\theta$$

the marginal likelihood under model $M = m$,

$$\pi(\theta|y, m) = \frac{p(y|\theta, m)\pi(\theta|m)}{p(y|m)}$$

the posterior for θ given $M = m$ and

$$\pi(m|y) = \frac{p(y|m)\pi_M(m)}{p(y)} \tag{6.1}$$

the posterior probability the true model is m, where $\pi_M(m) = \Pr(M = m)$ is the prior probability for M to equal m and

$$p(y) = \sum_{m\in\mathcal{M}} p(y|m)\pi_M(m)$$

is the model-averaged marginal likelihood. When we speak of the "true model" we mean nature's true generative model for the data. For example, in Chapter 5, in Section 5.3, we saw that a true generative model exists and is unique when the data came from an infinite exchangeable sequence. It is unlikely that the true model will be among the set of models $\mathcal{M}$ we are considering. When we do model selection we prefer models which make the prior predictive probability $p(y|m)$ for the data large. When we do model averaging we (automatically) put more weight on models giving a larger prior predictive probability for the data.

Suppose we are interested in the expectation $E_{\theta|y}(h(\theta))$. Instead of selecting a model and estimating $E_{\theta|y,m}E(h(\theta))$ in that model, we integrate over the model uncertainty.

Definition 6.1. The extended parameter space including the model index is

$$\begin{aligned}\Omega^* &= \bigcup_{m\in\mathcal{M}} \bigcup_{\theta\in\Omega_m} \{(\theta, m)\} \\ &= \bigcup_{m\in\mathcal{M}} \{m \times \Omega_m\}\end{aligned}$$

The joint posterior distribution for the model and parameter is

$$\pi(\theta, m|y) = \pi(\theta|y, m)\pi(m|y),\ (\theta, m) \in \Omega^*. \tag{6.2}$$

Definition 6.2. The model-averaged posterior[9] for the parameter is

$$\pi(\theta|y) = \sum_{m\in\mathcal{M}} \pi(\theta, m|y),\ \theta \in \Omega \tag{6.3}$$

where now θ could come from any one of the model spaces $\Omega_m,\ m \in \mathcal{M}$ so

$$\Omega = \bigcup_{m\in\mathcal{M}} \Omega_m.$$

Remark 6.3. When we expand the model averaged posterior in terms of the marginal likelihoods, by substituting Equation 6.2 into Equation 6.3 and using Equation 6.1 to expand $\pi(m|y)$,

$$\pi(\theta|y) \propto \sum_{m\in\mathcal{M}} \pi(\theta|y, m)p(y|m)\pi_M(m), \tag{6.4}$$

we see that terms in the sum with larger $p(y|m)$-values get a higher weight.

Remark 6.4. For computation we typically start from the joint distribution in the form

$$\pi(\theta, m|y) \propto p(y|\theta, m)\pi(\theta|m)\pi(m), \qquad (\theta, m) \in \Omega^*, \tag{6.5}$$

since everything here is usually tractable. The normalising constant is $p(y)$ above.

Example 6.5. *(averaging over link functions)* Recall our two models for the Challenger O-ring data,

$$y_i \sim \text{Bernoulli}(\mu_m(\beta_1 + \beta_2 x_i))\ i = 1, ..., n,$$

with x_i scaled temperature and a choice $\mu_m(\beta_1 + \beta_2 x_i)\ m = 1, 2$ of link functions with μ_1 the logit link and μ_2 the probit link. We will estimate the model averaged posterior $\pi(\beta|y)$ (and plot it for β_1). We are in effect averaging over the choice of link function. Take $\pi(m = 1) = \pi(m = 2) = 1/2$ model priors.

The two model parameter spaces are in this case equal, $\Omega_1, \Omega_2 = \mathbb{R}^2$ since the linear predictor $\beta_1 + \beta_2 x_i$ doesnt change when we change the link function. The model averaged posterior is given by summing over m in Equation 6.3

$$\pi(\beta|y) = \pi(\beta|m = 1, y)\,\pi(m = 1|y) + \pi(\beta|m = 2, y)\,\pi(m = 2|y).$$

We sum over m in Equation 6.3 rather than Equation 6.5 as we actually calculated the Bayes Factor $B_{1,2}$ for the model comparison, and obtained $B_{1,2} = 0.2.75$. Also, $p(y|m = 1)/p(y|m = 2) = \pi(m = 1|y)/\pi(m = 2|y)$ as the model priors cancel, and $\pi(m = 2|y) = 1 - \pi(m = 1|y)$, so we have

$$\pi(y|m = 1) = \frac{B_{1,2}}{1 + B_{1,2}}$$

and

$$\pi(\beta|y) = \pi(\beta|m = 1, y)\frac{B_{1,2}}{1 + B_{1,2}} \ +\ \pi(\beta|m = 1, y)\frac{1}{1 + B_{1,2}}.$$

Posterior densities for $\beta_1|y$ estimated for each model using MCMC are illustrated in Figure 13. The model-averaged posterior is slightly weighted towards Model 1 (Logistic).

[9] here I take $\pi(\theta, m|y) = 0$ if $\theta \notin \Omega_m$

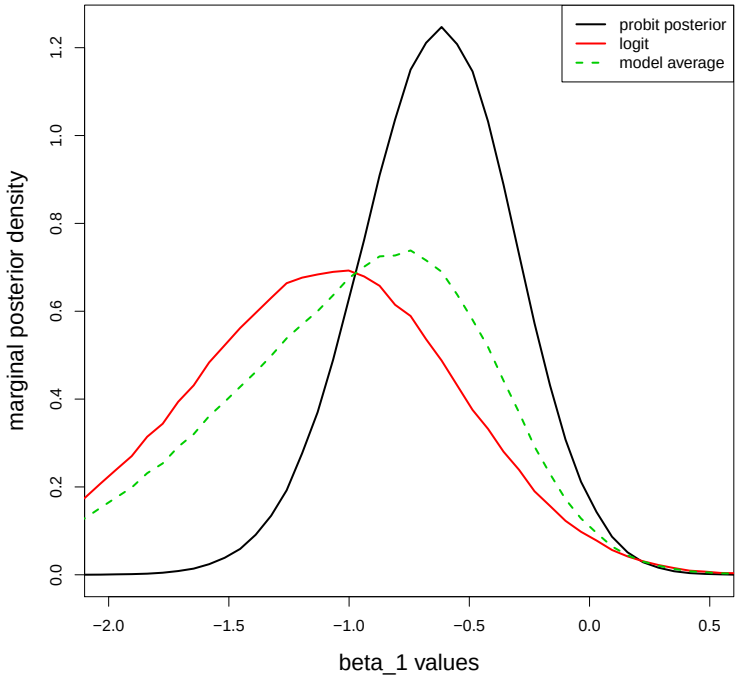


Figure 13: Posterior under logit (red) and probit (back) link functions and model-averaged posterior (green dashed) for Bernoulli-GLM regression of the binary O-ring data.

6.1.2 Model averaging is preferred to inference after model selection

Under squared error loss we should report the model averaged expectation. Suppose we want to estimate the value of some function $h(\theta)$. Here $h : \Omega \to \mathbb{R}$ so h has to be defined on each of the spaces $\Omega_m,\ m \in \mathcal{M}$. We will see some natural examples below. The posterior mean would be

$$E_{\theta,m|y}(h(\theta)) = \sum_{m\in\mathcal{M}} \int_{\Omega_m} h(\theta)\pi(\theta, m|y)\, d\theta.$$

Compare this with the single-model posterior mean at some selected model m^* say,

$$E_{\theta|y,m^*}(h(\theta)) = \int_{\Omega_{m^*}} h(\theta)\pi(\theta|y, m^*)\, d\theta.$$

If we do inference after model selection and report the posterior mean for the model $M = m^*$ we get a higher Bayes Risk.

Proposition 6.6. *If the loss for estimating δ when the truth is h is $(h-\delta)^2$ then the Bayes risk $\rho(\pi, \delta(y))$ allowing for model and parameter uncertainty is minimised by $E_{\theta,m|y}(h(\theta))$ and $\rho(\pi, E_{\theta|y,m}(h)) \geq \rho(\pi, E_{\theta,m|y}(h))$ for every $m \in \mathcal{M}$.*

Proof: Recall that the Bayes risk is minimised by the estimator minimising the expected posterior loss $\rho(\pi, \delta|y)$ at every $y \in \mathcal{Y}$. This is

$$\rho(\pi, \delta|y) = \sum_{m\in\mathcal{M}} \int_{\Omega_m} (\delta - h(\theta))^2 \pi(\theta, m|y)\, d\theta.$$

Now the calculation is the same as it was for the original case without model averaging, except that we now have an extra discrete parameter m, and we differentiate through the sum and integral, not just the intergral,

$$\frac{\partial \rho}{\partial \delta} = \sum_{m \in \mathcal{M}} \int_{\Omega_m} (2\delta - 2h(\theta))\pi(\theta, m|y)\, d\theta$$

and this is zero and when $\delta(y) = E_{\theta,m|y}(h(\theta))$. This is the unique minimum value for δ at y and so any other estimator cant be better. We could have an estimator $\delta'(y)$ which differs from $\delta(y)$ on null sets in $p(y)$ and get the same Bayes risk, but apart from those, the model averaged posterior mean minimises the Bayes risk. EOP

6.2 Model averaging with spike-and-slab priors.

You might find the text by Peter Hoff "A First Course in Bayesian Statistical Methods", Springer (2009), and in particular Section 9.3.1, useful for this bit. It is useful in simple cases where removing a parameter is the same as setting it equal to zero, and the components of the parameter have independent priors.

6.2.1 Spike and slab priors for regression

Consider model averaging in a regression problem with the setup,

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

with $z = (z_1, ..., z_p)$ a vector of binary indicator variables, and X an $n \times p$ matrix of continuous covariates. Here $z_i \in \{0, 1\}$ switches on and off the effect due to covariate X_i. For example, $z = (1, 1, 0, 0, ..., 0)$ gives standard linear regression, with $E(Y_i) = \theta_1 + \theta_2 X_{i,2}$.

There are 2^p models and our model index m is now effectively $z \in \mathcal{Z}$, which takes values in

$$\mathcal{Z} = \{0, 1\}^p.$$

The joint posterior in the form given by Equation 6.5 is

$$\pi(\theta, \sigma, z|y) \propto p(y|\theta, \sigma, z)\pi(\theta)\pi_s(\sigma)\pi_Z(z), \tag{6.6}$$

with $p(y|\theta, \sigma, z) = N(y; X\theta_z, \sigma^2 I_n)$ and I will suppose for simplicity that $\pi(\theta)\pi_s(\sigma)$ and $\pi_Z(z)$ are all jointly independent. The parameter space for $(\theta, \sigma) \in \Omega_z$ in model z is always $\Omega_z = \mathbb{R}^p \times \mathbb{R}^+$ so the parameter θ has the same the same dimension in every model. The joint parameter/model distribution $\pi(\theta, \sigma, z|y)$ is defined on the space

$$\Omega^* = \mathbb{R}^p \times \mathbb{R}^+ \times \mathcal{Z}.$$

Remark 6.7. Normally when we consider different models in regression, with different subsets of effects, the dimension of the parameter space Ω_z varies across models $z \in \mathcal{Z}$. The idea of introducing the auxiliary variables $z \in \mathcal{Z}$ is that *the models all have the same parameter space*, $(\theta, \sigma) \in \Omega_z$ with $\Omega_z = \Re^p \times \Re^+$ and this makes it easy (in principle) to analyse using our standard MCMC tools. If $z_i = 0$ then the posterior distribution for θ_i is just given by its prior, as the likelihood doesn't depend on θ_i when $z_i = 0$.

Let's see how it looks if we don't include the unused θ-components. Let $\theta_{z=1} = (\theta_i)_{i\in\{j\in[p]:z_j=1\}}$ and similarly for $\theta_{z=0}$. We can write

$$\pi(\theta_{z=0}, \theta_{z=1}, \sigma, z|y) \propto p(y|\theta_{z=1}, \sigma, z)\pi_s(\sigma)\pi_Z(z) \prod_{i:z_i=1} \pi(\theta_i) \prod_{j:z_j=0} \pi(\theta_j)$$

and integrate out $\theta_{z=0}$ and get

$$\pi(\theta_{z=1}, \sigma, z|y) \propto p(y|\theta_{z=1}, \sigma, z)\pi_s(\sigma)\pi_Z(z) \prod_{i:z_i=1} \pi(\theta_i).$$

Now let $m = (i \in [p] : z_i = 1)$ be the ordered list of indices for the covariates we include, say $m = (i_1, ..., i_k)$ when $\sum_i z_i = k$. In this notation the space of models $m \in \mathcal{M}$ where $\mathcal{M}$ is the set of all subsets of $[p] = \{1, ..., p\}$. Let $\phi = \theta_{z=1}$, $\phi \in R^k$ when there are $|m| = k$ components in ϕ. This leaves

$$\pi(\phi, \sigma, m|y) \propto p(y|\phi, \sigma, m)\pi(\phi)\pi_s(\sigma)\pi_M(m) \tag{6.7}$$

where

$$\pi_M(m) = \prod_{i\in m} \pi_Z(z_i = 1) \prod_{j\in[p]\setminus m} \pi_Z(z_j = 0).$$

and $p(y|\phi, \sigma, m) = N(y; X_{:,m}\phi, \sigma^2 I_n)$ where $X_{:,m}$ is the matrix we get by taking the columns X_i of X corresponding to $z_i = 1$, so $X_{:,m} = [X_i]_{i\in m}$ and $X_{:,m}\phi = X\theta_z$. Written this way the parameter space for the joint distribution $\pi(\phi, \sigma, m|y)$ is $(\phi, \sigma, m) \in \tilde{\Omega}^*$ where

$$\tilde{\Omega}^* = \bigcup_{m\in\mathcal{M}} \{m \times R^{|m|} \times R^+\}.$$

Now let's reparameterise back to the θ_z variables. We get back to that (simpler) representation using a familiar trick: we introduce auxiliary variables $\psi \in R^{p-|m|}$ into the model and take

$$\pi(\phi, \psi, \sigma, m|y) = \pi(\phi, \sigma, m|y)\pi(\psi).$$

Now $(\phi, \psi) \in R^p$ so the dimension is fixed. The mapping

$$\theta_i = \begin{cases} \phi_i & \text{if } i \in m \\ \psi_i & \text{otherwise} \end{cases},$$

and $z_i = \mathbb{I}_{i\in m}$ for $i \in [p]$ "takes us back" to where we began, with $X_{\theta_z} = X_{:,m}\phi$.

Remark 6.8. This θ-prior is called a "spike and slab" prior. Suppose

$$\pi(\theta, \sigma, z) = \pi_\sigma(\sigma) \prod_i \pi(\theta_i)\pi_Z(z_i)$$

(all independent a priori). The parameters appearing in the regression are θ_z. Consider the prior distribution of one of the actual regression effects $\tilde{\theta}_i = \theta_i z_i$. We get its prior CDF by summing over the possible values of $z_i = 0, 1$. Let $w = p_i(0)$ give the prior probability $z_i = 0$. We have

$$\Pr(\tilde{\Theta}_i \le c) = w\mathbb{I}_{c\ge 0} + (1 - w) \int_{-\infty}^{c} \pi_i(\tilde{\theta}_i)d\tilde{\theta}_i$$

We get the prior density by differentiating with respect to c,

$$\pi_{\tilde{\Theta}_i}(\tilde{\theta}_i) = w\delta_0(\tilde{\theta}_i) + (1 - w)\pi_i(\tilde{\theta}_i),$$

in terms of the delta-function notation introduced in Chapter 5, Definition 5.17, so the prior density for the regression parameters $\tilde{\theta} = \theta_z$ is "spike" plus "slab".

6.2.2 Polynomial regression of strike-insurance claim data

We illustrate the method for model averaging polynomial regression on data $x_i, Y_i,\ i = 1, ..., n$ for average claims paid per policy for Strikes in the USA in the years 1951-1980. The response Y_i is average claims paid in year $x_i,\ i = 1, ..., n$. In this example we do polynomial regression, so $X = [X_{i,j}]_{i=1,...,n}^{j=1,...,p}$ with $X_{i,j} = x_i^{j-1}$ (the highest power is $p-1$). The model is

$$Y_i = \sum_{j=1}^{p} z_j \theta_j x_{i,j}^{j-1} \ + \ \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$ and otherwise as above. In this model $z_i \in \{0, 1\}$ switches on and off the effect for the power $i-1$. For example, $z = (1, 1, 1, 0, ..., 0)$ gives regression with a quadratic, as $E(Y_i) = \theta_1 + \theta_2 x_i + \theta_3 x_i^2$. I scaled both x and y to mean zero and standard deviation one and allow up to quintic terms, $X = (1, x, x^2, x^3, x^4, x^5)$ so $p = 6$.

Priors: For the sake of example take $\theta_i \sim N(0, 9)$ and $\sigma \sim 1/\sigma$ in the parameter priors. We take the model prior

$$\pi(z) = \prod_{i=1}^{p} \xi^{z_i} (1-\xi)^{1-z_i}$$

with $\xi = c/p$ (with c small, I use $c = 3$ for $p = 5$ below). Our prior for z gives an expected number c of covariates playing a role in the fit. They are all equally likely to show an effect (in fact exchangeable).

Posterior: The joint distribution of the model index z and parameters θ, σ is, from Equation 6.6,

$$\pi(\theta, z, \sigma | y) \propto N(y; X\theta_z, \sigma^2) \times N(\theta; 0, 9I_p) \times \sigma^{-1} \times \xi^{|z|}(1-\xi)^{p-|z|}.$$

MCMC targeting $\pi(\theta, z, \sigma|y)$: random-walk MH-MCMC. Suppose $X_t = (\theta, z, \sigma)$ is the Markov chain state at step t. Cycle through a θ-update, a z-update and a σ-update sequentially.

θ-update: fix $a > 0$. At each step choose $i \sim U\{1, 2, ..., p\}$ and simulate a proposal

$$\theta_i' \sim U(\theta_i - a, \theta_i + a).$$

The new regression parameters are

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

The acceptance probability is

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

Everything else cancels as the other parameters have not changed.

z-update: choose $i \sim U\{1, 2, ..., p\}$ and set $z_i' = 1 - z_i$ (all 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 acceptance probability

$$\alpha_{z_i}(\theta_{z'} | \theta_z) = \min\left\{1, \frac{N(y; X\theta_{z'}, \sigma^2)\xi^{z_i'}(1-\xi)^{1-z_i'}}{N(y; X\theta_z, \sigma^2)\xi^{z_i}(1-\xi)^{1-z_i}}\right\}.$$

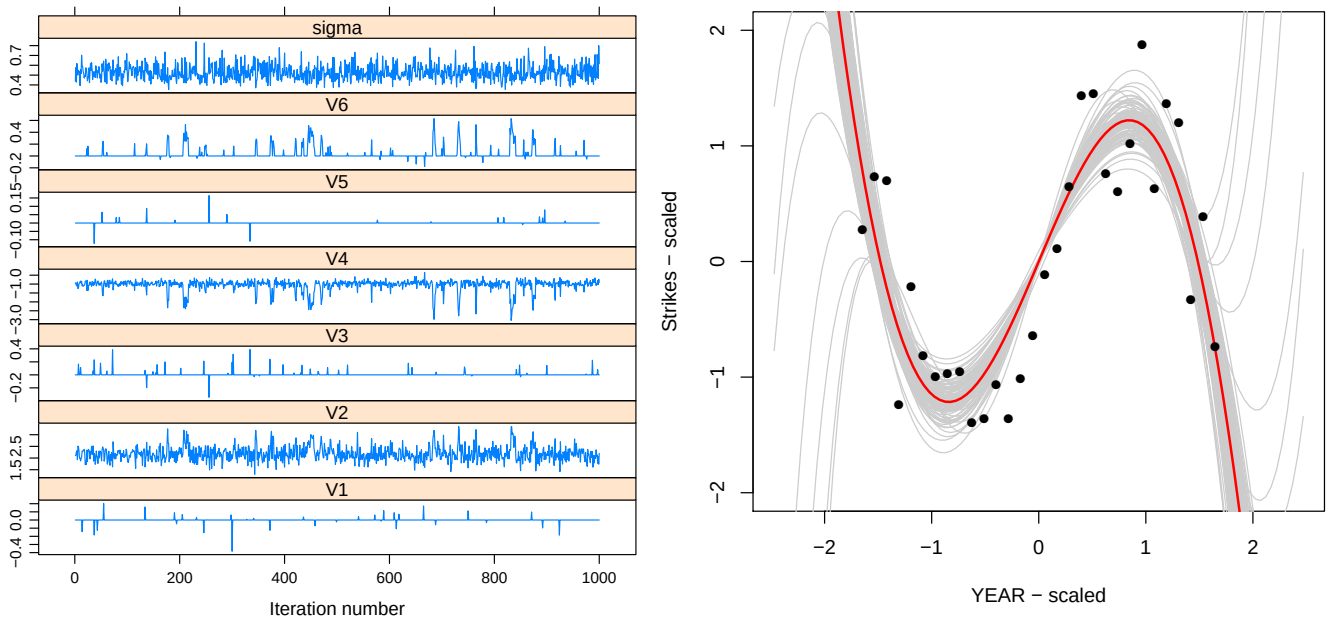


Figure 14: Polynomial regression of insurance claims data: (Left) MCMC traces; (Right) fit to data (black dots) showing sampled polynomial fits for the MCMC output (grey curves) and model-averaged posterior mean polynomial fit (red curve).

σ-update: In the code online I use a "random walk on a log scale" (comes up in a later lecture). We could use random-walk similar to the θ-update.

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

$$(\theta^{(t)}, z^{(t)}, \sigma^{(t)}) \sim \pi(\theta, z, \sigma|y) \quad t = 1, 2, ..., T.$$

The MCMC traces for $\theta_z^{(t)}$ are shown in Figure 14 The traces show the model variation. Notice that when $z_i = 0$, $\theta_{z,i} = 0$. There is little evidence in the data for a quartic term, so $z_5 = 0$ frequently and we see $\theta_{z,5}$ sits on zero. The same for the intercept (bottom graph, left panel) though in this case I guess it is the covariate scaling.

Suppose we want to estimate the fit at a new location $x \in \mathbb{R}$. Let $v(x) = (1, x, x^2, x^3, x^4, x^5)^\top$. Given θ and z, the regression model for a new observation Y_x at x is $E(Y_x|\theta, \sigma, z) = v(x)^\top \theta_z$ where

$$v(x)^\top \theta_z = z_1\theta_1 + z_2\theta_2 x + z_3\theta_3 x^2 + z_4\theta_4 x^3 + z_5\theta_5 x^4 + z_6\theta_6 x^6,$$

so if $\mu(x) = E(Y_x|y)$ is the expectation of the posterior predictive number of (scaled) strikes per year in (scaled) year x then

$$\mu(x) = E_{\theta,\sigma,z|y}(E(Y_x|\theta, \sigma, z))$$

so $\mu(x) = E_{\theta,\sigma,z|y}(v(x)^\top \theta_z)$ or

$$\mu(x) = v(x)^\top E_{\theta,\sigma,z|y}(\theta_z).$$

We estimate this in the obvious way from our MCMC samples, with

$$\hat{\mu}(x) = v(x)^\top \frac{1}{T} \sum_{t=1}^{T} \theta_z^{(t)}.$$

We plot $\hat{\mu}(x)$ against x to show the estimated posterior mean at each x as a function of x (red curve, right panel of Figure 14). In order to visualise the variation we plot $\mu(x;\theta_z^{(t)}) = v(x)\theta_z^{(t)}$ against x for several selected t-values sampled from the MCMC run (grey curves). This shows the polynomial-state for each of those samples.

The posterior probabilities $\pi(z|y)$ for the models can be estimated from the $z^{(t)}$ values by as proportion of times z appears in the MCMC output,

$$\hat{\pi}_z = T^{-1} \sum_t \mathbb{I}_{z^{(t)}=z}$$

which converges (a.s.) to $\pi(z|y)$. The results (as percentages) sorted by magnitude are

```
010100 010101 011100 110100 010001 010110 011101
  74.4   14.6    3.2    2.3    1.9    1.3    0.6
110101 010111 111100 011110 110110 111101 111110
   0.6    0.2    0.2    0.1    0.1    0.1    0.1
010011 011001 011111 110001 110111 111011 111111
   0.0    0.0    0.0    0.0    0.0    0.0    0.0
```

The top ranked (ie MAP) model is unsurprisingly cubic.

The analysis might be improved by restricting the model space to nested models in which lower order terms are included automatically, so $z \in \{000000, 100000, 110000, ..., 111111\}$. Also the MCMC mixed rather slowly on account of the parameterisation and scaling and more thought could go into this.

7 Reversible-Jump MCMC

See Givens and Hoeting, *Computational Statistics*, Wiley, (2013) for a clear introduction and Robert and Casella, *Monte Carlo Statistical Methods*, Springer (2004) for something more like the following. There is also a detailed presentation of RJ-MCMC in Piere Jacob's old lecture notes for Advanced Simulation where the topic was once taught. These can be found on the Advanced Simulation canvas page.

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,

$$\pi(\theta, m|y) \propto p(y|\theta, m)\pi(\theta|m)\pi(m), \quad \theta \in \Omega_m, m \in \mathcal{M},$$

the dimension of the parameter θ may vary depending on the model. We saw this in Section 6.2.1 where

model index	model	parameter	parameter space
$m = 1$	$Y = \theta_1 + \epsilon$	$\theta = (\theta_1)$	$\Omega_1 = \Re$
$m = 2$	$Y = \theta_1 + \theta_2 x + \epsilon$	$\theta = (\theta_1, \theta_2)$	$\Omega_2 = \Re^2$

We say that "The number of things we dont know is one of things we dont know" because we dont know if we have to estimate one parameter or two. This presents some computational issues. For example if we use MCMC to sample $(\theta, m) \sim \pi(\cdot|y)$ then the MCMC algorithm must jump between spaces of different dimension to allow the dimension of θ to vary. The trick of using latent indicator variables and spike and slab priors described in the last section is too restricted. We may wish to use other priors. In this example $\mathcal{M} = \{1, 2\}$ and

$$(\theta, m) \in \Omega^*,$$

where

$$\Omega^* = \bigcup_{m \in \mathcal{M}} \bigcup_{\theta \in \Omega_m} \{(\theta, m)\}.$$

We can alternatively write

$$\Omega = (\Omega_1 \times \{1\}) \cup (\Omega_2 \times \{2\})$$

in terms of two simple product spaces.

7.2 MCMC with a Jacobian

We begin by revisiting fixed-dimension MCMC and generalising it a bit so we can see some of the ideas in a familiar setting. In this section we replace the proposal distribution $q(\theta'|\theta)$ we had in MCMC with something equivalent but more easily generalised.

7.2.1 Proposals from transformations

I rearranged this slightly. The point is that everything follows from $\theta' = \psi_1(\theta, u)$ and the requirement that, for any pair (θ, θ') connected in this way, there exists a unique u' such that $\theta = \psi_1(\theta', u')$.

I simply assume $\theta'(\theta,u)$ and $u'(\theta,u)$ are differentiable, and leave this for the algorithm designer to check (like irreducibility).

We target a density $\pi(d\theta)=\pi(\theta)d\theta$ on $\Omega=\mathbb{R}^p$ (will be $\pi(\theta|y)$ but drop the y for now).

Definition 7.1. (*proposal variable and function*) Let $\mathcal{U}=\mathbb{R}^p$ or a given region of $\mathbb{R}^p$. Let $g(u)$, $u\in\mathcal{U}$ be a density on $\mathcal{U}$. For $\theta\in\Omega$ and $u\in\mathcal{U}$ let $\psi_1(\theta,u)$ be an invertible differentiable function of its arguments mapping Ω to itself given u so that $\psi_1:\Omega\times\mathcal{U}\to\Omega$. Given θ, the proposal simulates $u\sim g(u)$ and sets $\theta'=\psi_1(\theta,u)$. Call u the proposal variable and ψ_1 the proposal function.

Remark 7.2. Here $\psi_1(\theta,u)$ being invertible means the mapping $(\theta,u)\to\theta'$ at fixed θ is one-to-one and invertible, so for each $\theta'\in\{\psi_1(\theta,u):\ u\in\mathcal{U}\}$ there is a unique u such that $\theta'=\psi_1(\theta,u)$.

Example 7.3. For $a>0$ let $u\sim U(0,1)$ and set $\theta'=\theta+a(2u-1)$ to get our standard "random-walk" proposal $\theta'\sim U(\theta-a,\theta+a)$. Here

$$g(u)=\mathbb{I}_{0<u<1},\quad \psi_1(\theta,u)=\theta+a(2u-1).$$

This is invertible at fixed θ since $u=(a+\theta'-\theta)/2a$.

Proposition 7.4. *The conditional distribution of θ' given θ under the proposal in Definition 7.1 is given at $\theta'=\psi_1(\theta,u)$ by*

$$q(d\theta'|\theta)=g(u)du. \tag{7.1}$$

The density for θ' determined by the pair g,ψ_1 is

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

where $u=u(\theta')$ on the RHS solves $\theta'=\psi_1(\theta,u)$ and

$$\frac{\partial\theta'}{\partial u}=\frac{\partial\psi_1(\theta,u)}{\partial u}.$$

Proof: This is just a change of variables from u to θ' in which θ plays the role of a parameter in the change of variables. We apply the rule for changing variables in a density. This gives a Jacobian factor $|\partial\theta'/\partial u|^{-1}=|\partial u/\partial\theta'|$. [EOP]

Remark 7.5. We include the $d\theta'$ and du to make the point that the distribution $q(d\theta'|\theta)$ is given in terms of the distribution of u, not just the density but the element of volume du for integration. All these expressions properly live under an integral sign.

Remark 7.6. This is what was happening on the computer anyway - almost all standard distributions are simulated by simulating $U(0,1)$ variables and applying a transformation.

Example 7.7. Continuing Example 7.3, where $\theta'=\theta+a(2u-1)$ and $g(u)=\mathbb{I}_{0<u<1}$,

$$\begin{aligned}q(\theta'|\theta)&=\mathbb{I}_{0<(a+\theta'-\theta)/2a<1}\left|\frac{\partial\theta'}{\partial u}\right|^{-1}\\&=\frac{\mathbb{I}_{\theta-a<\theta'<\theta+a}}{2a}\end{aligned}$$

since $|\partial\theta'/\partial u|=2a$. That is $\theta'\sim U(\theta-a,\theta+a)$.

Definition 7.8. Suppose $\theta'=\psi_1(\theta,u)$ for some $u\in\mathcal{U}$ so $\theta\to\theta'$ is possible. We *require* that there exists a unique $u'\in\mathcal{U}$ giving the u-value reversing the move so that $\theta=\psi_1(\theta',u')$. Equivalently, since $\theta'=\psi_1(\theta,u)$, we let u' solve

$$\theta=\psi_1(\psi_1(\theta,u),u'). \tag{7.2}$$

The solution for u' will depend on θ and u and we write $u'=\psi_2(\theta,u)$ with $\psi_2(\theta,u)$ a function mapping (θ,u) into $\mathcal{U}$ so that $\psi_2:\Omega\times\mathcal{U}\to\mathcal{U}$. We call u' the proposal variable for the reverse update.

Remark 7.9. It is up to us to set things up so that a unique solution u' to Equation 7.2 exists, and $\psi_2(\theta, u)$ is differentiable. Actually we have been doing this already when we did MCMC for a continuous parameter as the next example shows.

Example 7.10. Continuing Example 7.3, here $\psi_2(\theta, u) = 1 - u$, since $u' = 1 - u$ will "take us back": $\psi_1(\theta', u') = \theta' + a(2u' - 1)$ so starting from the RHS of Equation 7.2 we have

$$\begin{aligned}\psi_1(\psi_1(\theta, u), u') &= \psi_1(\theta + a(2u - 1), u') \\ &= \theta + a(2u - 1) + a(2u' - 1)\end{aligned}$$

so $\psi_1(\psi_1(\theta, u), u') = \theta$ gives the relation $a(2u - 1) + a(2u' - 1) = 0$ which gives $u' = 1 - u$. It follows that Equation 7.2 has the unique solution $u' = \psi_2(\theta, u)$ with $\psi_2(\theta, u) = 1 - u$.

We have defined a mapping between pairs $(\theta', u') = \psi(\theta, u)$ with

$$\psi(\theta, u) = (\psi_1(\theta, u), \psi_2(\theta, u)).$$

We will need this mapping to be invertible (so the Jacobian $\partial\psi(\theta, u)/\partial(\theta, u)$ exists). We actually need the stronger requirement that $(\theta, u) = \psi(\theta', u')$ (so the inverse exists and is equal to ψ itself). Now $(\theta, u) = (\psi_1(\theta', u'), \psi_2(\theta', u'))$, and we already have $\theta = \psi_1(\theta', u')$ from Definition 7.8, so we need $u = \psi_2(\theta', u')$.

Proposition 7.11. *If $\theta' = \psi_1(\theta, u)$ and $u' = \psi_2(\theta, u)$ then $\psi_2(\theta', u') = u$.*

Proof: by Definition 7.8 there is $u' \in \mathcal{U}$ such that $\theta = \psi_2(\theta', u')$, so $\theta' \to \theta$ is possible. Then by the condition in the same definition, there exists a unique $x \in \mathcal{U}$ giving the u-value reversing the move so that x solves $\theta' = \psi_1(\theta, x)$. We know this has a unique solution, namely $x = u$, and since u then solves $\theta' = \psi_1(\psi_1(\theta', u'), u)$ we have $u = \psi_2(\theta', u')$. [EOP]

Proposition 7.12. *The function $\psi = (\psi_1, \psi_2)$ mapping $\psi : \Omega \times \mathcal{U} \to \Omega \times \mathcal{U}$ is an involution,*

$$(\theta, u) = \psi(\psi(\theta, u)),$$

that is, it is a function which is its own inverse.

Proof: By $\psi(\psi(\theta, u))$ we mean $\psi(\psi_1(\theta, u), \psi_2(\theta, u))$ which has two components. The first is

$$\psi_1(\psi_1(\theta, u), \psi_2(\theta, u)) \;=\; \psi_1(\theta', u') \;=\; \theta$$

by Equation 7.2. The second component is

$$\psi_2(\psi_1(\theta, u), \psi_2(\theta, u)) = \psi_2(\theta', u') = u$$

by Proposition 7.11.[EOP]

Example 7.13. If $\psi(\theta, u) = (\theta + a(2u - 1),\ 1 - u)$ as in Example 7.3 then

$$\begin{aligned}\psi(\psi(\theta, u)) &= \psi(\theta + a(2u - 1), (1 - u)) \\ &= (\theta + a(2u - 1) + a(2(1 - u) - 1), 1 - (1 - u)) \\ &= (\theta, u)\end{aligned}$$

7.2.2 MCMC using transformations

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 ψ_1 and let q be whatever it is. This is often easier to do well. If you havnt read Section 2.8, now would be a good time.

We write down the Metropolis Hastings Algorithm in a slightly different form from Section 2, using our proposal functions and proposal variables, and show the algorithm satisfies detailed balance.

Theorem 7.14. *Let $(\theta', u') = \psi(\theta, u)$ be an invertible, differentiable involution for $\theta, \theta' \in \Omega$ and $u, u' \in \mathcal{U}$. The MCMC update with proposal $u \sim g(u)$, $(\theta', u') = \psi(\theta, u)$ and acceptance probability*

$$\alpha(\theta'|\theta) = \min\left\{1,\ r(\theta', u'|\theta, u)\right\} \tag{7.3}$$

with $r(\theta', u'|\theta, u)$ given by

$$r(\theta', u'|\theta, u) = \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)} J_\psi(\theta, u), \tag{7.4}$$

and J_ψ the Jacobian for the transformation $(\theta', u') = \psi(\theta, u)$,

$$J_\psi(\theta, u) = \left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right|, \tag{7.5}$$

satisfies detailed balance in Equation 2.4 with respect to $\pi(\theta)$.

Proof: we need to verify detailed balance, Eqn 2.4. Change variables on the RHS in Eqn 2.4 from θ' to u at fixed θ. Using Eqn 7.1,

$$\pi(d\theta)q(d\theta'|\theta)\alpha(\theta'|\theta) = \pi(\theta)g(u)\alpha(\theta'(\theta, u)|\theta)dud\theta.$$

Suppose WLOG that $r(\theta', u'|\theta, u) \leq 1$, and use the abbreviated notation

$$(\theta'(\theta, u), u'(\theta, u)) = (\psi_1(\theta, u), \psi_2(\theta, u)).$$

Then the RHS of (2.4) is

$$\begin{aligned}
\pi(\theta)g(u)\alpha(\theta'(\theta, u)|\theta)dud\theta &= \pi(\theta)g(u)\frac{\pi(\theta'(\theta, u))g(u'(\theta, u))}{\pi(\theta)g(u)} J_\psi(\theta, u)dud\theta \\
&= \pi(\theta'(\theta, u))g(u'(\theta, u))\left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right| dud\theta \\
&= \pi(\theta')g(u')du'd\theta'
\end{aligned}$$

since the Jacobian we have is correct for the change of variables.

We assumed $r(\theta', u'|\theta, u) \leq 1$ (WLOG). Suppose it holds that

$$r(\theta', u'|\theta, u) = 1/r(\theta, u|\theta', u'), \tag{7.6}$$

so that $r(\theta, u|\theta', u') \geq 1$ and $\alpha(\theta, u|\theta', u') = 1$. The LHS of Eqn 2.4 is then

$$\pi(d\theta')q(d\theta|\theta')\alpha(\theta|\theta') = \pi(\theta')g(u')du'd\theta',$$

using Eqn 7.1 with $(\theta', u') \leftrightarrow (\theta, u)$ swapped, so LHS=RHS and detailed balance is satisfied.

We assumed Eqn 7.6 holds. We now verify this. We have

$$1/r(\theta,u|\theta',u') = \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)} J_\psi(\theta',u')^{-1},$$

which is a function of θ, u via $(\theta',u') = \psi(\theta,u)$, and

$$r(\theta',u'|\theta,u) = \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)} J_\psi(\theta,u).$$

These are equal if the Jacobian factors are equal. We now quote a standard result for the Jacobian of an invertable differentiable tranformation, usually expressed as "the Jacobian of the inverse transformation is the inverse of the Jacobian for the transformation". In our setting,

$$\begin{aligned} J_\psi(\theta,u) &= J_{\psi^{-1}}(\theta',u')^{-1} \\ &= J_\psi(\theta',u')^{-1} \end{aligned}$$

where the second line follows because ψ is an involution so ψ is ψ^{-1}. Lets do that again a bit more explicitly: take

$$(\theta,u) = \psi(\psi(\theta,u))$$

and differentiate the column vector (θ,u) WRT the row vector $\partial/\partial(\theta,u)$ both sides. Using the chain rule,

$$\begin{aligned} \frac{\partial(\theta,u)}{\partial(\theta,u)} &= \frac{\partial}{\partial(\theta,u)}\psi(\psi(\theta,u)) \\ I_{\dim(\theta,u)} &= \frac{\partial\psi(\theta,u)}{\partial(\theta,u)}\frac{\partial\psi(\psi(\theta,u))}{\partial(\psi(\theta,u)))} \\ &= \frac{\partial(\theta',u')}{\partial(\theta,u)}\left.\frac{\partial(\theta,u)}{\partial(\theta',u')}\right|_{(\theta',u')=\psi(\theta,u)}, \end{aligned}$$

and so inverting the matrix on the right (as we know we can),

$$\left[\frac{\partial(\theta,u)}{\partial(\theta',u')}\right]^{-1} = \frac{\partial(\theta',u')}{\partial(\theta,u)},$$

which is $J_\psi(\theta,u) = J_\psi(\theta',u')^{-1}$.

Finally when we make a change of variables we should transform the integration domain. Since we have shown that the integrands in Eqn 2.4 are equal, we have detailed balance if the integration domains are equal. This must hold as all the mappings are invertible, and is easily checked, so we have the desired result. [EOP]

Remark 7.15. The Jacobian must be non-singular, so that the change of variables is well defined. We get that because we insist the transformations are invertible (and differentiable). One very basic requirement is $\dim(\theta',u') = \dim(\theta,u)$ so that the Jacobian matrix in Equation 7.5 is a square matrix. This is called "dimension matching".

Example 7.16. (*Random walk on a log scale*) Suppose we are targeting $\pi(\theta) = e^{-\theta}$, $\theta > 0$ so $\theta \sim \exp(1)$ 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. To work out $\psi_2(\theta,u) = 1/u$ we simply observe that $\theta' = u\theta$ so $\theta = \theta'/u$ so $u' = 1/u$ is the value of the proposal variable for the reverse move. The Jacobian equals $1/u$ since

$$\left|\frac{\partial(\theta',u')}{\partial(\theta,u)}\right| = \begin{vmatrix} u & \theta \\ 0 & -1/u^2 \end{vmatrix} = 1/u,$$

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

$$\alpha(\theta'|\theta) = \min\left\{1, \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)}\left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right|\right\} = \min\left\{1, e^{-\theta'+\theta}u^{-1}\right\},$$

set $X_{t+1} = \theta'$ and otherwise $X_{t+1} = \theta$.

Factors of $g(u)/g(u')$ cancel in α. This proposal is useful if simulating a density which is peaked or diverges at a boundary and I used a variant of this in the radiocarbon dating example MCMC code. This whole framework is a useful rephrasing of MCMC.

Exercise 7.17. show that the Jacobian for the simple random walk proposal $u \sim U(-a, a)$ with transformation $(\theta', u') = (\theta + u, -u)$ is equal one.

Exercise 7.18. In the polynomial regression example in Section 6.2.2, the update for σ is

$$u \sim U(\delta, 1/\delta), \qquad \sigma' = u\sigma$$

with $0 < \delta < 1$ a constant we can choose (and adjust for efficient MCMC). Calculate the acceptance probability (answer in the code for that material, but this is pretty much the same as Example 7.16).

7.2.3 Matched proposals

We saw in Section 2.1.8 that we can mix proposals, breaking down the overall kernel

$$K(\theta, d\theta') = \sum_{i=1}^{N} \xi_i K_i(\theta, d\theta')$$

into a mixture of all the different transitions $K_i(\theta, d\theta'),\ i = 1, 2, ...N$. For DB we need

$$\sum_i \pi(d\theta)\xi_i K_i(\theta, d\theta') \ = \ \sum_i \pi(d\theta')\xi_i K_i(\theta', d\theta) \tag{7.7}$$

We can now "pair up" the kernels.

Definition 7.19. (*matched kernels*) Let $\sigma \in \mathcal{P}_N$ be a permutation of $\{1, ..., N\}$ satisfying $\sigma_{\sigma_i} = 1$, so if $\sigma_i = j$ then $\sigma_j = i$ (for example $\sigma = (2, 1, 6, 5, 4, 3)$). Suppose we pair K_i with $K_{\sigma_i},\ i = 1, ..., N$. Let $q_i(d\theta'|\theta) = q_i(\theta'|\theta)d\theta',\ i = 1, ..., N$ be proposal distributions and densities and let

$$\alpha_i(\theta'|\theta) = \min\left\{1, \ \frac{\pi(\theta')\xi_{\sigma_i}q_{\sigma_i}(\theta|\theta')}{\pi(\theta)\xi_i q_i(\theta'|\theta)}\right\}.$$

Let

$$c_i(\theta) = 1 - \int_\Omega \alpha_i(\theta'|\theta)dq_i(\theta'|\theta)$$

so that

$$K_i(\theta, d\theta') = \alpha_i(\theta'|\theta)dq_i(\theta'|\theta) + c_i(\theta)\delta_\theta(d\theta').$$

Remark 7.20. We have set things up so that if $\sigma_i = j$ and we ask what is paired with j then that will be $\sigma_j = i$, which is what we need. These are all Metropolis-Hastings kernels with the right (paired) acceptance probabilities.

Proposition 7.21. *Detailed balance in Equation 7.7 is satisfied if the following distributions match in (i, σ_i) pairs,*

$$\pi(d\theta)\xi_i\alpha_i(\theta'|\theta)q_i(d\theta'|\theta) \ = \ \pi(d\theta')\xi_{\sigma_i}\alpha_{\sigma_i}(\theta|\theta')q_{\sigma_i}(d\theta|\theta'), \quad i = 1, ..., N.$$

Proof: Referring to Equation 7.7, the terms involving c_i will cancel in detailed balance as they did in Proposition 2.10 (*ie*, between $K_i(\theta, d\theta')$ and $K_i(\theta', d\theta)$), so detailed balance is the condition that the integrands,

$$\sum_i \pi(d\theta)\xi_i\alpha_i(\theta'|\theta)dq_i(\theta'|\theta) \;=\; \sum_i \pi(d\theta')\xi_i\alpha_i(\theta|\theta')dq_i(\theta|\theta'),$$

match in sum. This works because each term on the left has a unique matching pair on the right. Suppose WLOG that $\alpha_i(\theta'|\theta) \leq 1$. In terms of the densities,

$$\begin{aligned}\pi(\theta)\xi_i\alpha_i(\theta'|\theta)q_i(\theta'|\theta) &= \pi(\theta)\xi_i\frac{\pi(\theta')\xi_{\sigma_i}q_{\sigma_i}(\theta|\theta')}{\pi(\theta)\xi_i q_i(\theta'|\theta)}q_i(\theta'|\theta)\\ &= \pi(\theta')\xi_{\sigma_i}q_{\sigma_i}(\theta|\theta')\\ &= \pi(\theta')\xi_{\sigma_i}\alpha_{\sigma_i}(\theta|\theta')q_{\sigma_i}(\theta|\theta')\end{aligned}$$

since $\alpha_{\sigma_i}(\theta|\theta') = 1$. We should check this. When we apply the formula for α_i with $i \to \sigma_i$ we get

$$\begin{aligned}\alpha_{\sigma_i}(\theta|\theta') &= \min\left\{1,\; \frac{\pi(\theta)\xi_{\sigma_{\sigma_i}}q_{\sigma_{\sigma_i}}(\theta'|\theta)}{\pi(\theta')\xi_{\sigma_i}q_{\sigma_i}(\theta|\theta')}\right\}\\ &= \min\left\{1,\; \frac{\pi(\theta)\xi_i q_i(\theta'|\theta)}{\pi(\theta')\xi_{\sigma_i}q_{\sigma_i}(\theta|\theta')}\right\}\\ &= 1\end{aligned}$$

since we required σ to be a matching so that $\sigma_{\sigma_i} = i$, and since we assumed the Hastings ratio for $\theta \to \theta'$ was less than or equal one. [EOP]

Remark 7.22. An interesting consequence is that *the individual kernels* $K_i,\ i = 1, ..., N$ *do not need to satisfy detailed balance* individually. In Example 7.24 below we will have two kernels in which if $\theta < \theta' < 2\theta$ then one kernel (if we choose (a)) allows $\theta \to \theta'$ and cannot generate the transition $\theta' \to \theta$, and the other (if we choose (b)) allows $\theta' \to \theta$ and cannot generate the transition $\theta \to \theta'$.

We can extend this to our generate-transform setting. Suppose $N = 2$ above so we have just two kernels and $\xi_1 = \rho$ and $\xi_2 = 1 - \rho$. The matching is just $\sigma = (2, 1)$. We draw $u \sim g_1,\ u \in \mathcal{U}_1$ with probability ρ and otherwise $u \sim g_2,\ u \in \mathcal{U}_2$. We use the same differentiable, invertible transformation $\theta' = \psi_1(\theta, u)$ to get the new state in both cases and again $u' = \psi_2(\theta, u)$. This time $u' \in \mathcal{U}_2$ if $u \in \mathcal{U}_1$ and $u' \in \mathcal{U}_1$ if $u \in \mathcal{U}_2$ so $\psi = (\psi_1, \psi_2)$ maps the space

$$(\Omega \times \mathcal{U}_1) \cup (\Omega \times \mathcal{U}_2)$$

back onto itself. We arrange things so this is an involution again so this side of things hasnt changed. We put a subscript $i = 1, 2$ on $q_i(d\theta'|\theta)$ to indicate which g is used, so the transform in Equation 7.1 is $q_i(d\theta'|\theta) = g_i(u)du$ and we can write

$$q_i(\theta'|\theta) = g_i(u)\left|\frac{\partial\theta'}{\partial u}\right|^{-1}.$$

Detailed balance in Proposition 7.21 becomes

$$\pi(d\theta)\rho q_1(d\theta'|\theta)\alpha_1(\theta'|\theta) = \pi(d\theta')(1-\rho)q_2(d\theta|\theta')\alpha_2(\theta|\theta') \tag{7.8}$$

Proposition 7.23. *Let* $(\theta', u') = \psi(\theta, u)$ *be an invertible, differentiable involution on* $(\Omega \times \mathcal{U}_1) \cup (\Omega \times \mathcal{U}_2)$. *Suppose* $X_t = \theta$. *Then* X_{t+1} *is determined in the following way.*

1. *With probability ρ simulate $u \sim g_1(\cdot)$ and otherwise (ie with probability $1-\rho$) simulate $u \sim g_2(\cdot)$. Set $(\theta', u') = \psi(\theta, u)$.*
2. *(a) If we chose $u \sim g_1(\cdot)$ then accept θ' (and set $X_{t+1} = \theta'$) with probability*

$$\alpha_1(\theta'|\theta) = \min\left\{1,\ \frac{\pi(\theta')(1-\rho)g_2(u')}{\pi(\theta)\rho g_1(u)}\left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right|\right\}.$$

(b) If we chose $u \sim g_2(\cdot)$ then accept θ' (and set $X_{t+1} = \theta'$) with probability

$$\alpha_2(\theta'|\theta) = \min\left\{1,\ \frac{\pi(\theta')\rho g_1(u')}{\pi(\theta)(1-\rho)g_2(u)}\left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right|\right\}.$$

Otherwise set $X_{t+1} = \theta$.

This update satisfies detailed balance in Proposition 7.21 between the pair of transition kernels associated with proposal distributions g_1 and g_2.

Proof: The proof is essentially the same as before. The $\pi(\theta)g_1(u)$-factors cancel and leave us with the $\pi(\theta')g_2(u')$ factors in the numerator, which is what we need. The ψ transformation has been set up so that that side of things goes through as before. [EOP]

Example 7.24. (*transformation with a matched pair of proposals*) 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 $\rho = 1/2$. 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 $u \sim U(1,2)$ else (b) $u \sim U(0.5, 1)$. Set $\theta' = u\theta$.
2. if we chose (a) then

$$g_2(u')/g_1(u) = \frac{U(u'; 0.5, 1)}{U(u; 1, 2)} = 2$$

and we accept θ' wp

$$\alpha(\theta'|\theta) = \min\{1, 2 \times e^{-\theta'+\theta}u^{-1}\},$$

and if we chose (b) then $g_1(u')/g_2(u) = 0.5$ and we accept θ' wp

$$\alpha(\theta'|\theta) = \min\{1, 0.5 \times e^{-\theta'+\theta}u^{-1}\}.$$

7.3 Reversible Jump MCMC

7.3.1 A shortcut to RJ-MCMC

Before we do RJ-MCMC properly, lets see how straightforward it can be. We need a proposal and an acceptance probability for $(\theta, m) \to (\theta', m')$.

Remark 7.25. The probability distribution of $\theta', m'|\theta, m$ is often easy to write down though it often involves discrete and continuous elements and so is a product of densities and probability mass functions. Call this $Q(\theta', m'|\theta, m)$. To form Q we simply write down the product for the probabilities for each of the steps that take us from $(\theta, m) \to (\theta', m')$. If the distributions are interpreted correctly then we can actually write

$$\alpha(\theta', m'|\theta, m) = \min\left\{1, \frac{\pi(\theta', m'|y)\, Q(\theta, m|\theta', m')}{\pi(\theta, m|y)\, Q(\theta', m'|\theta, m)}\right\} \tag{7.9}$$

Example 7.26. To be concrete, consider the variable dimension setup in Remark 6.7. The space of models $m \in \mathcal{M}$ is the set of all subsets of $[p] = \{1, ..., p\}$ and $\theta \in R^k$ when there are $|m| = k$ components in θ. The joint model-parameter posterior is given in Equation 6.7,

$$\pi(\theta, \sigma, m|y) \propto p(y|\theta, \sigma, m)\pi(\theta)\pi_s(\sigma)\pi_M(m)$$

where $\pi_M(m)$ is given (we took $\pi_M(m) = \xi^k(1-\xi)^{p-k}$ in Section 6.2.1), and the likelihood is $p(y|\theta, \sigma, m) = N(y; X_{:,m}\theta, \sigma^2 I_n)$ where $X_{:,m}$ is the matrix we get by taking columns X_i, $i \in m$ of X, so $X_{:,m} = [X_i]_{i\in m}$. The state space for the joint distribution $\pi(\theta, \sigma, m|y)$ is $(\theta, \sigma, m) \in \tilde{\Omega}^*$ where

$$\tilde{\Omega}^* = \bigcup_{m\in\mathcal{M}} \{m \times R^{|m|} \times R^+\}.$$

Suppose we propose $m \to m'$ by tossing a coin; if heads then propose to add a component and otherwise delete one (if $m = \emptyset$ then we cant delete and if $m = [p]$ then we cant add, so we reject in these cases). If we are adding pick $i \in [p] \setminus m$ at random and set $m' = m \cup \{i\}$ (and m' is an ordered set so put i in the right place). Choose a value for the new parameter $\theta'_i \sim \pi(\theta'_i)$ (the prior) and set $\theta' = \theta \cup \theta'_i$ (in the same position as i is placed in m'). If we are deleting then pick $i \in m$ and set $m' = m \setminus i$ and $\theta' = \theta \setminus \theta_i$.

In order to calculate Q we just write down the probabilities for the sequences of events we realised to get from (θ, m) to (θ', m'). We have

$$Q(\theta', m'|\theta, m) = \begin{cases} 1/2 \times 1/(p - |m|) \times \pi(\theta'_i) & \text{if we choose to add,} \\ 1/2 \times 1/|m| & \text{if we choose to delete,} \end{cases}$$

and "going back" from the new state, $|m'| = |m| + 1$ if we added and $|m'| = |m| - 1$ if we deleted. In order to reverse the move we have to pick the component we changed so,

$$\begin{aligned} Q(\theta, m|\theta', m') &= \begin{cases} 1/2 \times 1/(p - |m'|) \times \pi(\theta_i) & \text{if we chose to delete,} \\ 1/2 \times 1/|m'| & \text{if we chose to add.} \end{cases} \\ &= \begin{cases} 1/2 \times 1/(p - |m| + 1) \times \pi(\theta_i), \\ 1/2 \times 1/(|m| + 1). \end{cases} \end{aligned}$$

The acceptance probabilities are

$$\begin{aligned} \alpha(\theta', m'|\theta, m) &= 1 \wedge \frac{\pi(\theta', \sigma, m'|y) \times 1/2 \times 1/(|m| + 1)}{\pi(\theta, \sigma, m|y) \times 1/2 \times 1/(p - |m|) \times \pi(\theta'_i)}, \\ &= 1 \wedge \frac{\pi(\theta')\pi_M(m')p(y|\theta', \sigma, m')(p - |m|)}{\pi(\theta)\pi_M(m)p(y|\theta, \sigma, m)(|m| + 1)\pi(\theta'_i)} \quad \text{if we chose to add,} \end{aligned}$$

and

$$\begin{aligned} \alpha(\theta', m'|\theta, m) &= 1 \wedge \frac{\pi(\theta', \sigma, m'|y) \times 1/2 \times 1/(p - |m| + 1) \times \pi(\theta_i)}{\pi(\theta, \sigma, m|y) \times 1/2 \times 1/|m|} \\ &= 1 \wedge \frac{\pi(\theta')\pi_M(m')p(y|\theta', \sigma, m')|m|\pi(\theta_i)}{\pi(\theta)\pi_M(m)p(y|\theta, \sigma, m)(p - |m| + 1)} \quad \text{if we chose to delete} \end{aligned}$$

I implemented this, using the same θ-updates at fixed m as before and the same σ update. As we discussed this is the same model as we analysed in Section 6.2.1 (the marginal integrating out $\theta_{z=0}$). I got the same results (up to MC error). The distribution over models was

```
010100 010101 011100 110100 010110 011101 010111
  77.9   12.1    3.9    2.7    1.1    0.6    0.4

110101 111100 010001 011110 110110 111101 111110
   0.3    0.3    0.2    0.2    0.1    0.1    0.1
```

You can check MCMC diagnostic plots using the code for this lecture. [End of Example]

7.3.2 Some typical reversible jump proposals

Consider MCMC targeting a general model-averaging posterior

$$\pi(\theta, m|y) \propto p(y|\theta, m)\pi(\theta|m)\pi(m),\ (\theta, m) \in \Omega^*$$

with $m \in \mathcal{M}$ and $\mathcal{M}$ the set of all subsets of $\{1, 2, 3, ...p\}$ where p may be taken to be infinite. Write $m = (i_1, ..., i_{|m|})$. Given m, $\theta \in \Omega_m$ with $\theta = (\theta_{i_1}, ..., \theta_{i_{|m|}})$, so the model m tells us which parameters we have in the parameter vector. Sometimes the parameters labels are exchangeable and we really only need to know the number $|m|$ of elements in the set. However sometimes the parameters are linked to specific covariates (as above) and we need to know which is which. We further suppose that $\theta_i \in R^k,\ i = 1, ..., |m|$, for $k \geq 1$ fixed (in the example below $k = 3$), so each component of θ has k dimensions. In this setup $\Omega_m = \mathbb{R}^{k|m|}$ and we write $\dim(\theta) = k|m|$ meaning that $\pi(d\theta|y, m) = \pi(\theta|y, m)d\theta$ with $d\theta$ the element of volume in $\mathbb{R}^{k|m|}$. The full space of parameter $\times$ model is

$$\Omega^* = \cup_{m\in\mathcal{M}} \cup_{\theta\in\Omega_m} \{(\theta, m)\}.$$

The MCMC state is $X_t = (\theta, m)$ for some state $(\theta, m) \in \Omega^*$. When we move between models the dimension of θ changes. We can use the proposal matching setup of the last section to do this.

Let $\rho_{m,m'}$ give the probability to propose a move to model m' given the current model is m. In the following $\rho_{m,m'} = 0$ unless $||m| - |m'|| = 1$ (add or delete one component of θ), so we change the dimension of θ by k-dimensions. This is enough to allow the chain to move around the space of parameters and models (so, irreducible). We assume below that m and m' are fixed (so m is known and the choice of m' has been made). So for example ψ_1 and ψ_2 are specific to a given pair m and m'.

Adding dimensions to the state

Consider an update proposal *adding* an entry to the vector so $(\theta, m, u) \to (\theta', m', u')$ with $m' = (i_1, ...i_j, i, i_{j+1}, ..., i_{|m|})$, so we add component $i_j < i < i_{j+1}$ to the model. Here $\theta = (\theta_{i_1}, ..., \theta_{i_{|m|}})$ and

$$\theta' = (\theta_{i_1}, ..., \theta_{i_j}, \theta'_i, \theta_{i_{j+1}}, ..., \theta_{i_{|m|}}),$$

with $\theta' \in \mathbb{R}^{|m|k+k}$.

The generating density for the update is $g_{m,m'}(u),\ u \in \mathcal{U}_{m,m'}$, with $g_{m,m'}$ a density with respect to the element of volume du in $\mathcal{U}_{m,m'} = \mathbb{R}^k$. We take $u \sim g_{m,m'}(\cdot)$ and set $\theta' = \psi_1(\theta, u)$ with $\psi_1 : \Omega_m \times \mathcal{U}_{m,m'} \to \Omega_{m'}$ in this update. The function ψ_1 is just

$$\psi_1(\theta, u) = (\theta_{i_1}, ..., \theta_{i_j}, \theta'_i(\theta, u), \theta_{i_{j+1}}, ..., \theta_{i_{|m|}})),$$

so it adds on a component into θ at position $i = m' \setminus m$ which is given as a function of θ and u, and keeps the rest of the θ's the same. We need the proposal transform $\theta' = \psi_1(\theta, u)$ to be invertible

for u at fixed θ so we must have $\dim(\theta, u) = \dim(\theta')$. We are going from $|m|k$ dimensions in θ to $|m|k + k$ dimensions in θ', so we must ensure $\dim(u) = k$, so u makes up the missing dimensions.

Removing dimensions from the state

We choose the reverse move $(\theta', m', u') \to (\theta, m, u)$ with probability $\rho_{m',m}$. Let $\mathcal{U}_{m',m} = \{\emptyset\}$ and $g_{m',m}(\emptyset) = 1$. These choices reflect the fact that if we choose this update then we have nothing to do but delete the entry $i = m' \setminus m$. There are no random variables u' involved after the move is chosen. We have $\theta = \psi_1(\theta', u')$ with $\psi_1 : \Omega_{m'} \times \mathcal{U}_{m',m} \to \Omega_m$ in this update. Given the argument $u' = \emptyset$, ψ_1 just removes θ'_i from θ' so

$$\psi_1(\theta', \emptyset) = (\theta_{i_1}, ..., \theta_{i_j}, \theta_{i_{j+1}}, ..., \theta_{i_{|m|}}).$$

In the addition update we specified the values taken by ψ_1 over $\Omega_m \times \mathcal{U}_{m,m'}$. Now we are specifying the values it takes over $\Omega_{m'} \times \mathcal{U}_{m',m}$.

Now take a general pair $\theta \in \Omega_m$ and $\theta' \in \Omega_{m'}$ that differ by the addition of exactly one element. For ψ_2 we see that u' solves $\theta = \psi_1(\theta', u')$. This is achieved by the choice $u' = \emptyset$. For the other move, u solves $\theta' = \psi_1(\theta, u)$. Since u only affects the i-component, u solves $\theta'_i = \theta'_i(\theta, u)$. We need to know the specific proposal to make any further progress.

Transformations

For $(\theta, u) \in \Omega_m \times \mathcal{U}_{m,m'}$ and $(\theta', u') \in \Omega_{m'} \times \mathcal{U}_{m',m}$ let

$$(\theta', u') = \psi(\theta, u)$$

with

$$\psi(\theta, u) = (\psi_1(\theta, u), \psi_2(\theta, u))$$

the invertible mapping for this update-pair. We require

$$(\theta', \emptyset) = (\psi_1(\theta, u), \psi_2(\theta, u)),\ \theta' \in \Omega_{m'}$$

and

$$(\theta, u) = (\psi_1(\theta', \emptyset), \psi_2(\theta', \emptyset)).$$

We have set up ψ as a differentiable involution mapping

$$\left(\Omega_m \times \mathcal{U}_{m,m'}\right) \cup \left(\Omega_{m'} \times \mathcal{U}_{m',m}\right)$$

back to itself. If we start in $\Omega_m \times \mathcal{U}_{m,m'}$ then ψ maps to a state in $\Omega_{m'} \times \mathcal{U}_{m',m}$ and if we apply the same transformation to this state in $\Omega_{m'} \times \mathcal{U}_{m',m}$ then we arrive back in $\Omega_m \times \mathcal{U}_{m,m'}$ at the state where we started.

7.3.3 The RJ-MCMC algorithm

Proposition 7.27. *Reversible Jump MCMC algorithm*

Suppose $X_t = (\theta, m)$. The state X_{t+1} is determined as follows.

1. *Sample $m' \sim \rho_{m,m'},\ m' \in \mathcal{M}$. Simulate $u \sim g_{m,m'}(\cdot)$.*
2. *Set $(\theta', u') = \psi(\theta, u)$.*
   (a) If $|m'| = |m| + 1$ (increase dimension) set

$$\alpha(\theta', m'|\theta, m) = \min\left\{1, \frac{\pi(\theta', m'|y)\rho_{m',m}}{\pi(\theta, m|y)\rho_{m,m'}g_{m,m'}(u)} J_\psi(\theta, u)\right\}$$

(b) If $|m'| = |m| - 1$ (decrease dimension) set

$$\alpha(\theta', m'|\theta, m) = \min\left\{1, \frac{\pi(\theta', m'|y)\rho_{m',m}g_{m',m}(u')}{\pi(\theta, m|y)\rho_{m,m'}} J_\psi(\theta, \emptyset)\right\}$$

3. With probability $\alpha(\theta', m'|\theta, m)$ set $X_{t+1} = (\theta', m')$ and otherwise set $X_{t+1} = (\theta, m)$.

This update satisfies detailed balance in Equation 7.10 below with respect to $\pi(\theta, m|y)$ between the pair of transition kernels associated with the addition and deletion proposal kernels.

Remark 7.28. By $J_\psi(\theta, \emptyset)$ in the case where we are deleting I mean $|\partial(\theta', u')/\partial\theta|$. This will be non-singular because we deleted, so $|m'| = |m| - 1$, and u' has the extra dimensions ensuring $\dim(\theta', u') = \dim(\theta)$. Similarly $J_\psi(\theta, u) = |\partial\theta'(\theta, u)/\partial(\theta, u)|$.

Proof: Detailed balance is

$$\pi(d\theta', m'|y)\rho_{m',m}q_{m',m}(d\theta|\theta')\alpha(\theta, m|\theta', m') = \pi(d\theta, m|y)\rho_{m,m'}q_{m,m'}(d\theta'|\theta)\alpha(\theta', m'|\theta, m). \quad (7.10)$$

Consider the case where $|m'| = |m| + 1$ so $u' = \emptyset$ as set out in the addition-update above. We have

$$q_{m,m'}(d\theta'|\theta) = \delta_\theta(d\theta'_m)g_{m,m'}(u)du$$

since the m-components of θ' are unchanged, $\theta'_m = (\theta'_j)_{j\in m} = \theta$. For the reverse move

$$q_{m',m}(d\theta|\theta') = \delta_{\theta'_m}(d\theta)$$

since we make θ by setting $\theta_j = \theta'_j$ for $j \in m$ when we delete. Proceeding as in the proof of Theorem 7.14, and assuming $\alpha(\theta', m'|\theta, m) \leq 1$ (WLOG), the RHS of Equation 7.10 is

$$\begin{aligned}\pi(d\theta, m|y)\rho_{m,m'}q_{m,m'}(d\theta'|\theta)\alpha(\theta', m'|\theta, m) &= \pi(\theta, m|y)\rho_{m,m'}g_{m,m'}(u)\delta_{\theta_m}(d\theta'_m)\frac{\pi(\theta', m'|y)\rho_{m',m}}{\pi(\theta, m|y)\rho_{m,m'}g_{m,m'}(u)}\left|\frac{\partial\theta'(\theta, u)}{\partial(\theta, u)}\right| dud\theta \\ &= \pi(\theta', m'|y)\rho_{m',m}\delta_{\theta_m}(d\theta'_m)\left|\frac{\partial\theta'(\theta, u)}{\partial(\theta, u)}\right| dud\theta.\end{aligned}$$

We next show that

$$\left|\frac{\partial\theta'(\theta, u)}{\partial(\theta, u)}\right| = \left|\frac{\partial\theta'_i}{\partial u}\right|.$$

Suppose for simplicity $i = |m| + 1$ so θ'_i is the last entry in θ'. This follows because

$$\left|\frac{\partial\theta'(\theta, u)}{\partial(\theta, u)}\right| = \begin{vmatrix} \frac{\partial\theta'_m}{\partial\theta} & \frac{\partial\theta'_m}{\partial u} \\ \frac{\partial\theta'_i}{\partial\theta} & \frac{\partial\theta'_i}{\partial u} \end{vmatrix}$$

and since $\frac{\partial\theta'_m}{\partial u} = 0_{k|m|\times k}$ and $\frac{\partial\theta'_m}{\partial\theta} = I_{k|m|\times k|m|}$ the determinant (which is $J_\psi(\theta, u)$) must be $J_\psi(\theta, u) = |\partial\theta'_i/\partial u|$. The RHS of detailed balance is then

$$\pi(d\theta, m|y)\rho_{m,m'}q_{m,m'}(d\theta'|\theta)\alpha(\theta', m'|\theta, m) = \pi(\theta', m'|y)\rho_{m',m}\delta_\theta(d\theta'_m)\left|\frac{\partial\theta'_i}{\partial u}\right| dud\theta.$$

(use the Jacobian to change variables from $dud\theta$ to $d\theta'_i d\theta'_m = d\theta'$ with $\delta_{\theta'_m}(d\theta)$ enforcing $\theta = \theta'_m$)

$$\begin{aligned}&= \pi(\theta', m'|y)\rho_{m',m}\delta_{\theta'_m}(d\theta)d\theta', \\ &= \pi(d\theta', m'|y)\rho_{m',m}q_{m',m}(d\theta|\theta').\end{aligned}$$

which is the left hand side, since $\alpha(\theta', m'|\theta, m) \leq 1$ implies $\alpha(\theta, m|\theta', m') = 1$ by the same reasoning as for Theorem 7.14 (ie exploiting the fact that $J_\psi(\theta, u) = J_\psi^{-1}(\theta', \emptyset)$ for a transform ψ which is an involution). Again, we have not tracked the integration domains as we should, but the transformations are all invertible so when we transform $d\theta' \to d\theta du \to d\theta'$ we must end up integrating over the same sets we started with. [EOP]

Exercise 7.29. Check the acceptance probability for the case where $|m'| = |m| - 1$ and we have a deletion update.

7.4 The Galaxy radial velocity data - RJ MCMC for mixture models

The Galaxy radial velocity data are shown in Figure 17. It is natural to model this via a mixture of normals in which mixture components might capture different classes of galaxy. We make this a small case study in the application of RJ-MCMC. We are hoping that different classes might have different typical radial velocities. However we do not know the number of components in the mixture.

7.4.1 Observation model

We model our data $y_i \in \Re, i = 1, 2, ..., n$ as independent samples from a mixture model with m components $N(\mu_j, {\sigma_j}^2)$, and mixture weights $w_j, j = 1, 2, ..., m,\ w_j > 0, \sum_{j=1}^m w_j = 1$. Given $m \in 1, 2, 3, ...$ the mixture parameters are

$$\mu = (\mu_1, ..., \mu_m), \quad \sigma = (\sigma_1, ..., \sigma_m), \quad w = (w_1, ..., w_m).$$

The observation model is

$$p(y|\mu, \sigma, w, m) = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} w_j N(y_i; \mu_j, {\sigma_j}^2) \right].$$

If $\theta_j = (\mu_j, \sigma_j, w_j),\ j = 1, ..., m$ then $\theta = \{\theta_1, ..., \theta_m\}$ and this is an unordered set, as any permutation gives the same set of clusters. Any particular vector of θ's has $m!$ copies and we would like to treat these as equivalent states.

Notice the model index is just a positive integer m now, not an ordered set, as it was in the regression example in Section 7.3.1. This is because we dont need to keep track of the order of the θ's because any permutation of the indices $(\theta_1, ..., \theta_m) \to (\theta_{\sigma_1}, ..., \theta_{\sigma_m})$ gives the same likelihood. That wasnt true in Section 7.3.1 as the link between parameters and covariates (columns of X) meant we had to maintain the order in θ.

7.4.2 Priors

We need a prior π_M over models. We suppose some prior knowledge of the number of different galaxy classes which might play a role. For this example we take $\pi_M(m) = \text{Poisson}(m; \lambda | m > 0)$ with $\lambda = 10$. This is centred at 10, and tails off above about 20 clusters. In this example our focus is the RJ-MCMC.

For the parameter priors $\pi(\mu, \sigma, w|m)$, we take

$$w \sim \text{Dirichlet}(\alpha 1_{\text{m}}) \qquad \text{with } 1_m \text{ a vector of } m \text{ ones}$$

and take $\alpha = 1$ so w are uniform probabilities, summing to one due to the Dirichlet prior. Take

$$\mu \sim N(\mu_0 1_m, v_0 I_m), \quad \text{with } \mu_0 = 20 \text{ and } v_0 = 10^2,$$

covering the data in $[0, 40]$ at 2σ - I assume the scale of the response is known *a priori* - and

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

again informed by the scale: the average cluster has a standard deviation equal 3; the choice of shape parameter equal 1.5 (in particular, greater then one) rules out very dense clusters at small σ; the small rate equal 0.5 gives a relatively heavy tail, and the standard deviation of the prior for σ is about 2.5.

7.4.3 Mixture model posterior

The posterior for the model and parameters (θ, m), with $\theta_j = (\mu_j, \sigma_j, w_j)$, $j = 1, ..., m$, $\theta = \{\theta_1, ..., \theta_m\}$ and $m \in \{1, 2, 3, ...\}$ is

$$\begin{aligned}\pi(\theta, m|y) &\propto p(y|\theta, m)\pi(\theta|m)\pi(m) \\ &\propto p(y|\mu, \sigma, w, m) \times \text{Dirichlet}(w; \alpha 1_m) \times \prod_{j=1}^{m} N(\mu_j; \mu_0, v_0) \times \text{Gamma}(\sigma_j; 1.5, 0.5) \\ &\times \quad \text{Poisson}(m; \lambda) \times m! \end{aligned} \tag{7.11}$$

The extra $m!$ is needed as the RHS is the posterior probability for a particular vector of θ's. If we want to treat all permutations as equally likely we need to upweight the probability by this factor. We have effectively summed the RHS over all equivalent permutations. Note that for $m > 0$,

$$\text{Poisson}(m; \lambda | m > 0) \propto \text{Poisson}(m; \lambda),$$

as a function of m at fixed λ, so the condition that we have at least one cluster can be dropped. We just constrain the space so that $m > 0$.

7.4.4 RJ MCMC algorithm targeting a normal mixture posterior

In the following we use the "shortcut" in Remark 7.25 to calculate the acceptance probabilities. This skips the Jacobian calculations and working through $g(u)$ etc. We will essentially just take the product of the probabilities or densities for the events which lead from θ, m to θ', m' and call this $Q(\theta', m'|\theta, m)$ in Equation 7.9.

Suppose the state is $X_t = (\theta, m)$, $\theta = \{(\mu_i, \sigma_i, w_i)\}_{i=1}^m$. For irreducibility we need 3 fixed dimension moves and 2 variable dimension moves.

Step 1 Choose an update uniformly at random, move $\sim U\{1, 2, ..., 5\}$.

If move $= 1$ *add a component (increase state dimension by three).*

Step 2(up) We generate $\theta'_{m+1} = (\mu'_{m+1}, \sigma'_{m+1}, w'_{m+1})$ and set $\theta' = \theta \cup \theta'_{m+1}$ and $m' = m + 1$.

Step 2(up)a Simulate $\mu'_{m+1}, \sigma'_{m+1} \sim q_{\mu\sigma}(\mu'_{m+1}, \sigma'_{m+1})$ (use Normal-Gamma prior above). Set $\mu' = (\mu, \mu'_{m+1})$ and $\sigma' = (\sigma, \sigma'_{m+1})$.

Step 2(up)b To complete $\theta'_{m+1} = (\mu'_{m+1}, \sigma'_{m+1}, w'_{m+1})$, we have to assign a value to w'_{m+1}, maintaining $\sum_j w'_j = 1$. 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 distribution Q in Equation 7.9 for (μ', σ', w', m') given (μ, σ, w, m) is

$$Q(\mu', \sigma', w', m'|\mu, \sigma, w, m) = \rho_{m,m'} q_{\mu\sigma}(\mu'_{m+1}, \sigma'_{m+1}) \times \frac{1}{m} \times \frac{1}{w_j}.$$

The last factor $1/w_j$ appears because we chose $w'_{m+1} \sim U(0, w_j)$ with normalised density $1/w_j$.

Step 2(up)c In the reverse move (move 2(down) below, decreasing dimension) pick component $i \in \{1, ..., m'\}$ at random, delete it and add its weight w'_i to randomly chosen component $j \in \{1, ..., m'\}$, $j \neq i$. The probability (mass) to propose this exact reverse move back to (μ, σ, w, m) given (μ', σ', w', m') is

$$Q(\mu, \sigma, w, m|\mu', \sigma', w', m') = \rho_{m',m} \frac{1}{m(m+1)},$$

since we choose i from $m' = m+1$ and j from $m' - 1 = m$.

Step 3(up) Accept the proposal (μ', σ', w', m') with probability

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

where

$$\alpha^+ = \min\left\{1, \frac{\pi(\mu', \sigma', w', m'|y)\rho_{m',m}\frac{1}{m(m+1)}}{\pi(\mu, \sigma, w, m|y)\rho_{m,m'}q_{\mu\sigma}(\mu'_{m+1}, \sigma'_{m+1}) \times \frac{1}{m} \times \frac{1}{w_j}}\right\}$$

from Equation 7.9.

If move $= 2$ *delete a component (decrease state dimension by three).*

Step 2(down) Set $m' = m - 1$ (if $m' = 0$, reject the move and set $X_{t+1} = X_t$).

Step 2(down)a Simulate $i \sim U\{1, 2, ..., m\}$. Set $\mu' = \mu_{-i}$, $\sigma' = \sigma_{-i}$.

Step 2(down)b Simulate $j \sim U\{1, 2, ..., i-1, i+1, ..., m\}$, replace $w'_j \leftarrow w_j + w_i$ and set $w' = w_{-i}$.

The probability for the forward proposal is

$$Q(\mu', \sigma', w', m'|\mu, \sigma, w, m) = \rho_{m,m'}/m(m-1),$$

and to reverse

$$Q(\mu, \sigma, w, m|\mu', \sigma', w', m') = \rho_{m',m} q_{\mu\sigma}(\mu_i, \sigma_i) \times \frac{1}{m-1} \times \frac{1}{w_i + w_j}.$$

Step 3(down) Accept the proposal (μ', σ', w', m') with probability

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

where, after cancelation,

$$\alpha^- = \min\left\{1, \frac{\pi(\mu', \sigma', w', m'|y) m q_{\mu\sigma}(\mu_i, \sigma_i)}{\pi(\mu, \sigma, w, m|y)(w_i + w_j)}\right\}$$

We have additionally moves 3-5 which act on μ, σ and w respectively in fixed dimension moves.

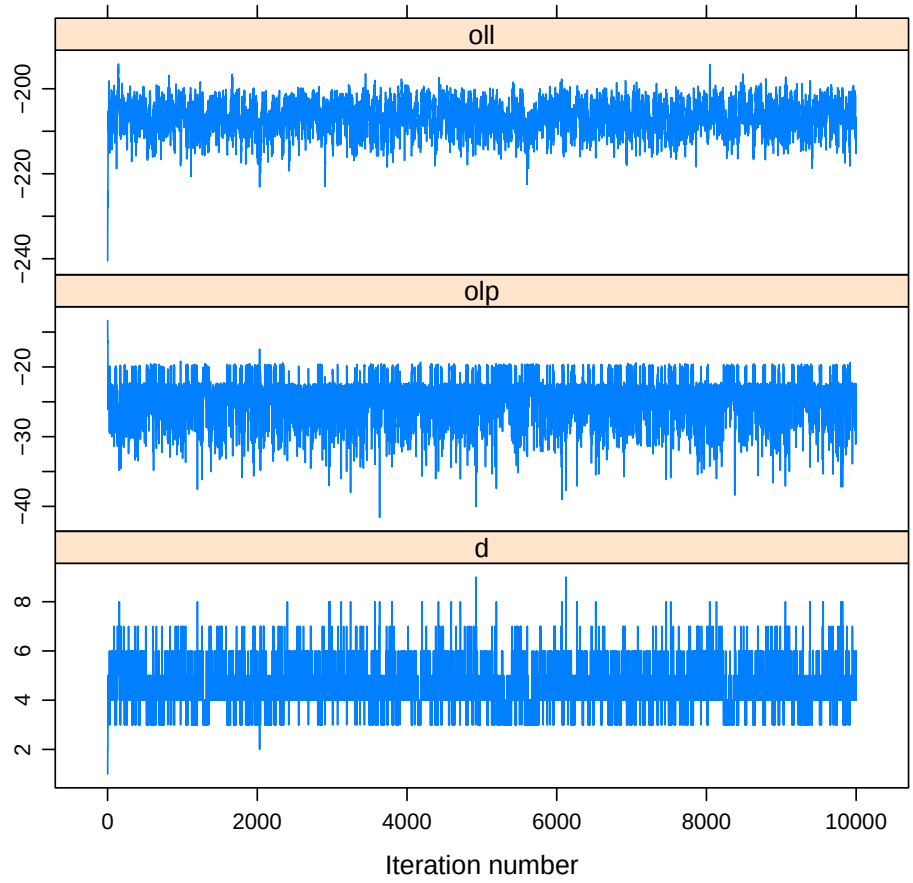


Figure 15: MCMC traces for the log-prior, log-likelihood and number of components, (as the number of parameters vary, the parameters themselves are not easily plotted).

7.4.5 RJ-MCMC fitting a normal mixture model for the Galaxy data

We apply the RJ-MCMC algorithm in Section 7.4.4 to target the posterior in Equation 7.11. R-code is available on the course website. We 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. MCMC output traces are shown in Figure 15. Convergence seems reasonable. The posterior distribution of the number of mixture components is shown in Figure 16. Our analysis averages over the number of clusters and alows us to estimate the likely number of distinct clusters. Figure 17 shows an estimate $\widehat{p(y'|y)}$ of the posterior predictive distribution $p(y'|y)$ (black line) at each point y' on the x-axis. Since

$$p(y'|y) = \sum_m \int p(y'|\theta, m)\pi(\theta, m|y)d\theta$$

we use the natural estimate

$$\widehat{p(y'|y)} = \frac{1}{T}\sum_{t=1}^{T} p(y'|\theta^{(t)}, m^{(t)}).$$

We expect the distribution of the data to match the posterior predictive distribution, and the fit seems reasonable.

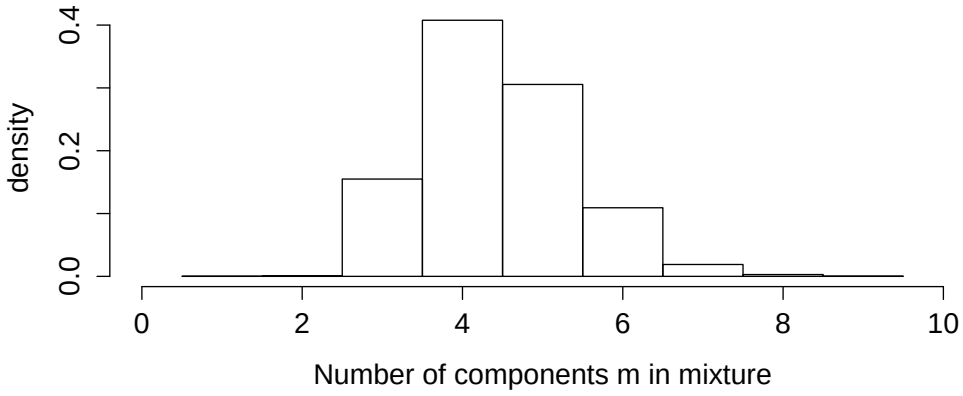


Figure 16: RJ-MCMC for the mixture model for the Galaxy radial velocity data: posterior distribution over the number of components. 3-6 components is the number favored.

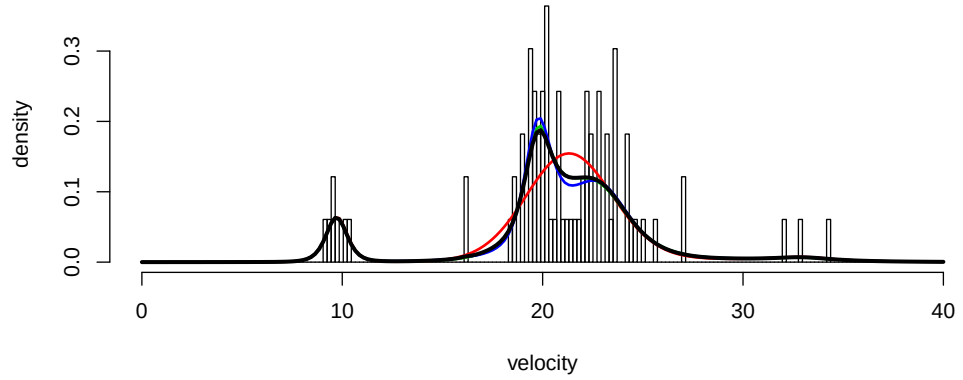


Figure 17: Galaxy radial velocity data - measurements of radial velocities for $n = 82$ galaxies shown as a frequency histogram. Posterior predictive densities overlaid on the data. The underlying histogram in black is a histogram of the data, y. Other lines are $p(y'|y, m)$-estimates (m=3 red, m=4 green, m=5 blue).

8 The Dirichlet Process

The aim of this Chapter to is introduce Bayesian non-parametrics through a particular model, the Dirichlet Process (DP). BNP with the DP is model averaging, with randomly variable dimension. The classical analysis uses an MCMC algorithm that is actually doing a kind of RJ-MCMC. The most useful result in this chapter is probably Theorem 8.44, a simple expression for the DP-posterior which is generally easy to apply. When we actually carry out Bayesian inference using a DP-prior we jump straight to this expression. One doesn't actually need the preceding theory, though it is certainly helpful to understand the properties of a DP and when it may be a suitable model choice.

8.1 Motivation

With enough data you reject any *parametric* model. Small model violations (skew, correlation, heavy tails, number of mixture components) built-in by parametric model assumptions may become glaring as the data set grows large. Non-parametric (NP) models allow fitting with an unbounded number of parameters. 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 wont adapt to every form of model mispecification. The model contains information which may bias the inference. This is good if the model biases towards the truth!

We will look at one example of a NP model and Bayesian Non-Parametric (BNP) fitting: the Dirichlet process mixture for density estimation and clustering. The model changes but the methodological framework we have developed (model averaging, marginal likelihoods...) is the same.

Here is an informal view of where we are going. Recall the discussion in Section 5.3. If data $y_1, y_2, ...$ is an infinite exchangeable sequence then

$$p(y_1, ..., y_n) = \int_\Omega p(y_1, ..., y_n|\theta) dG(\theta)$$

for some distribution G. We called G "nature's prior". These distributions and the posterior distribution

$$d\pi(\theta|y_1, ..., y_n) \propto p(y_1, ..., y_n|\theta) dG(\theta)$$

all exist. However we don't know G (like we didn't know θ). In Bayesian inference, unknowns become random variables so we move the analysis up one level and estimate G.

Let $\mathcal{G}$ be the space of probability distributions we allow as possible generative models for the parameter (so, possible priors) and suppose $G \in \mathcal{G}$ is the unknown true generative distribution for the parameter. Let Π be a probability distribution over $\mathcal{G}$. This is a distribution over distributions. We have actually seen this before, when we did model averaging. If $M \sim \pi_M$ then $\pi(\theta|M)$ is a random prior probability distribution. In that setting the model space $\mathcal{M}$ was at least countable.

We will now allow the model space to be much larger. A given prior $\pi \in \mathcal{G}$ is a realisation $G = \pi$ of this random distribution, $G \sim \Pi$, putting random probability mass $\pi(A)$ on measurable sets $A \subseteq \Omega$. We have a prior $d\Pi(G)$ for our prior!

We might *very informally* treat G as another parameter and write down the joint prior distribution of θ and G,

$$d\Pi(G, \theta) = dG(\theta)d\Pi(G)$$

in which case the posterior would be

$$d\Pi(G, \theta|y) \propto p(y|\theta)dG(\theta)d\Pi(G).$$

It would in general be hopeless to express the infinite dimensional G explicitly on a computer, so we will work with the marginal

$$d\pi(\theta) \propto \int_{\mathcal{G}} dG(\theta)d\Pi(G)$$

and set things up so that $d\pi(\theta) = \pi(\theta)d\theta$ so that there is a density wrt to $d\theta$. We calculate and work on the marginal $d\pi(\theta)$ and posterior $d\pi(\theta|y) \propto p(y|\theta)d\pi(\theta)$. Prior elicitation moves up a level in the hierarchy to the choice of Π, a distribution over distributions.

8.2 The Dirichlet Process and the Chinese Restaurant Process

8.2.1 The Dirichlet *Distribution*

Here are some properties of the "ordinary" Dirichlet Distribution we will use. Let $w = (w_1, ..., w_M)$ with $w \in \{v \in (0,1)^M : \sum_{k=1}^M v_k = 1\}$. Suppose $w \sim \mathsf{Dirichlet}(\alpha_1, ..., \alpha_M)$. The density of w is

$$\pi(w_1, w_2, ..., w_M) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} w_1^{\alpha_1 - 1} w_2^{\alpha_2 - 1} ... w_M^{\alpha_M - 1}.$$

By the *agglomerative property*, if $w_1, ..., w_M \sim \mathsf{Dirichlet}(\alpha_1, ..., \alpha_M)$ then

$$w_1 + w_2, w_3, ..., w_M \sim \mathsf{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, ..., \alpha_M).$$

The Dirichlet distribution is a *conjugate prior for the multinomial distribution*: if $(n_1, ..., n_M) \sim \mathsf{Multinom}(n, w)$ with $n = \sum_k n_k$ and $w \sim \mathsf{Dirichlet}(\alpha_1, ..., \alpha_M)$ then

$$\pi(w|n_1, ..., n_M) \propto w_1^{\alpha_1 + n_1 - 1} w_2^{\alpha_2 + n_2 - 1} ... w_M^{\alpha_M + n_M - 1},$$

so $w|n_1, ..., n_M \sim \mathsf{Dirichlet}(\alpha_1 + n_1, ..., \alpha_M + n_M)$.

8.2.2 The Multinomial Dirichlet process

We want to define a random probability distribution G, that is, we want to distribute probability randomly in parameter space Ω (which is always $\mathbb{R}^p$ or an open subset here). We will build this up using a simple base distribution H. Let θ_k^*, $k = 1, 2, 3, ...$ be continuous random variables each having probability space $(\Omega, \mathcal{B}, H)$. Here $H(d\theta_k^*) = h(\theta_k^*)d\theta_k^*$ with H some simple parametric base distribution on Ω with density h. We need H to be in some sense "simple", but otherwise generalisation is straightforward.

Definition 8.1. The Multinomial Dirichlet Process $G_M \sim \Pi_M(\alpha, H)$ is the following process: Let $M \geq 1$ and $\alpha > 0$ be given.

1. For $k = 1, ..., M$, sample $\theta_k^* \sim H$.
2. Sample $w_1, ..., w_M \sim \text{Dirichlet}(\alpha/M)$.
3. Set $dG_M(\theta) = \sum_{k=1}^M w_k \delta_{\theta_k^*}(d\theta)$. We alternatively write $G_M = \sum_{k=1}^M w_k \delta_{\theta_k^*}$.

This simply drops M random probability masses $w_k,\ k = 1, ..., M$ at a random locations θ_k^* in Ω.

Remark 8.2. Here G_M gives the distribution of θ so if we take $A \in \mathcal{B}$ a set of parameters values then $\Pr(\theta \in A|G_M) = G_M(A)$ will be random. It is equal to the sum of the weights $w_k : \theta_k^* \in A$:

$$\begin{aligned} G_M(A) &= \int_\Omega \mathbb{I}_{\theta \in A} dG_M(\theta) \\ &= \int_A \sum_{k=1}^M w_k \delta_{\theta_k^*}(d\theta) \\ &= \sum_{k=1}^M w_k \mathbb{I}_{\theta_k^* \in A}. \end{aligned} \tag{8.1}$$

So indeed the process assigns a random probability mass $G_M(A)$ to measurable sets $A \subseteq \Omega$.

By putting a prior on w and θ^* we determine a prior *on a probability distribution*. The choice $w \sim \text{Dirichlet}(\alpha/M, ..., \alpha/M)$, $\theta_k^* \sim H,\ k = 1, ..., M$ determines a Multinomial Dirichlet Process (MDP) G_M. Let us look at some of its properties.

Proposition 8.3. *The random distribution G_M is "centred" on the base distribution H in the sense that $E(G_M(A)) = H(A)$ where*

$$H(A) = \int_\Omega \mathbb{I}_{\theta \in A} h(\theta) d\theta.$$

Proof: From Eqn 8.1, since w and θ^* are independent, and looking up $E(w_k) = (\alpha_k/M)/\sum_j(\alpha_j/M)$,

$$\begin{aligned} E(G_M(A)) &= \sum_{k=1}^M E(w_k \mathbb{I}_{\theta_k^* \in A}) \\ &= \sum_{k=1}^M E(w_k) E(\mathbb{I}_{\theta_k^* \in A}) \\ &= \sum_k \frac{\alpha_k}{\sum_j \alpha_j} H(A) \\ &= H(A) \end{aligned} \tag{EOP}$$

Exercise 8.4. Use the Dirichlet-variance formula (look this up) to calculate $\text{var}(G_M(A))$.

Remark 8.5. Another important property of the MDP and the DP below is that it is "atomic". It puts atoms of probability mass at points in Ω. One consequence of this is that, if $G_M \sim \Pi_M(\alpha, H)$ and $\theta_1, \theta_2 \sim G_M$, then marginally $\Pr(\theta_1 = \theta_2) > 0$.

Exercise 8.6. Show that $\Pr(\theta_1 = \theta_2|\theta^*, w) = \sum_k w_k^2$ and hence $\Pr(\theta_1 = \theta_2) > 0$ for $\alpha > 0$.

8.2.3 The Dirichlet *Process*

Definition 8.7. (*Dirichlet process*): $G \sim \Pi(\alpha, H)$ is a Dirichlet Process (DP) iff for all partitions $A_1, ..., A_r$ of Ω (with $A_k \in \mathcal{B},\ k = 1, ..., r$), we have

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

Perhaps surprisingly, there exists a unique process satisfying these conditions.

Remark 8.8. The definition says that $G(A_1), ..., G(A_r)$ (which are random, as G is random) have a joint distribution given by a Dirichlet distribution. This is at least possible, because $0 \leq G(A_k) \leq 1,\ k = 1, ..., r$ and $\sum_k G(A_k) = 1$ since $A_k,\ k = 1, ..., r$ is a partition of Ω.

Remark 8.9. If the process exists (it does) and $G \sim \Pi$ then $\theta_1, ..., \theta_n, ... \sim G$ (iid) is an infinite exchangeable sequence (IES) by construction. Conversely, we give an algorithm below realising an IES $\theta_1, ..., \theta_n$ directly given α, H. By de Finetti (in a statement of the theorem more general than the one we wrote down) a random variable G and distribution $d\Pi(G)$ must exist. From the properties of the sequence we construct we can show the random variable G satisfies the definition, so the Dirichlet process $\Pi(\alpha, H)$ exists.

Remark 8.10. The DP is also obtained in a concrete construction as the limit $M \to \infty$ of the Multinomial Dirichlet process G_M. The definition above "starts afresh". We wont prove this as part of the course - an outline can be found in the Appendix. The proof shows that

$$(G_M(A_1), ..., G_M(A_r)) \xrightarrow{D} \text{Dirichlet}(\alpha H(A_1), ..., \alpha H(A_r))$$

as $M \to \infty$. I include this statement because it gives use some idea of what a realisation of a DP "looks like". It is an infinite set $\{\theta_k^*\}_{k \in \{1,2,3,...\}}$ of atoms of probability in $\Omega = \mathbb{R}^p$ with θ_k^* having weight $w_k \geq 0$, and $\sum_{k=1}^{\infty} w_k = 1$.

8.2.4 Some properties of the Dirichlet Process

A number of properties follow immediately from Definition 8.7.

Proposition 8.11. *If $G \sim \Pi(\alpha, H)$ then for any set $A \in \mathcal{B}$ we have $E(G(A)) = H(A)$*

Proof: Since $H(A^c) = 1 - H(A)$, we must have

$$G(A), G(A^c) \sim \mathsf{Dirichlet}(\alpha H(A), \alpha(1 - H(A))).$$

This Dirichlet Distribution with two components is a Beta distribution for $G(A)$, so

$$E(G(A)) = \frac{\alpha H(A)}{\alpha H(A) + \alpha(1 - H(A))}$$

and hence $E(G(A)) = H(A)$. [EOP]

Proposition 8.12. *If $G \sim \Pi(\alpha, H)$ and $\theta \sim G$ then marginally $\theta \sim H$.*

Proof: The statement $\theta \sim H$ means $\Pr(\theta \in A) = H(A)$ for $A \in \mathcal{B}$, so check this:

$$\begin{aligned} \Pr(\theta \in A) &= E_G(E_{\theta|G}(\mathbb{I}_{\theta \in A}|G)) \\ &= E(G(A)), \end{aligned}$$

so $\Pr(\theta \in A) = H(A)$ for all $A \in \mathcal{B}$ by Proposition 8.11. It follows that $\theta \sim H$. [EOP]

Remark 8.13. If $G \sim \Pi(\alpha, H)$ and $\theta \sim G$ then for measurable $B \subseteq A$,

$$\Pr(\theta \in B | \theta \in A) \ = \ \frac{\Pr(\theta \in B)}{\Pr(\theta \in A)} \ = \ \frac{H(B)}{H(A)}$$

If $H(d\theta) = h(\theta)d\theta$, so distribution H has density h, then marginally $\theta|\theta \in A$ has density $h(\theta|\theta \in A)$.

8.2.5 DP generative model and predictive distributions

Suppose we make a sequence of draws $\theta = (\theta_1, ..., \theta_n)$ from the same "observation model" G. Our aim in this and the next section will be to compute $d\pi(\theta)$, the joint marginal distribution of $\theta_1, ..., \theta_n$ which we get by integrating out the common random measure G. As in the previous section we want to work with the marginal probabilities $\Pr(\theta \in A)$. However this time things are a bit more complicated as we want to consider $n > 1$ and in this case two $\theta's$ could be exactly equal.

First we give a process realising the sequence of draws $\theta = (\theta_1, ..., \theta_n)$:

Definition 8.14. The generative model for $\theta = (\theta_1, ..., \theta_n)$ is given by the following process:

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

Denote by $d\pi(\theta)$ the marginal distribution of θ generated in this way.

Remark 8.15. Here G is a "parameter" like θ, but we want to work with the marginal $d\pi(\theta)$ not the joint for G and θ. We will use the identity

$$d\pi(\theta) = d\pi(\theta_n|\theta_{1:n-1})d\pi(\theta_{n-1}|\theta_{1:n-2})...d\pi(\theta_1) \tag{8.2}$$

and calculate each term in the product. We know $d\pi(\theta_1) = H(d\theta_1)$ (this was the marginal distribution we covered in Proposition 8.12) so for $i = 2, ..., n$ the predictive distributions $d\pi(\theta_i|\theta_{1:i-1})$ are needed.

Proposition 8.16. *For $j = 0, ..., n-1$ we have $d\pi(\theta_{j+1}|\theta_{1:j}) = \tilde{H}_j(d\theta_{j+1})$ where*

$$\tilde{H}_j(d\theta_{j+1}) = \frac{\alpha H(d\theta_{j+1}) + \sum_{i=1}^{j} \delta_{\theta_i}(d\theta_{j+1})}{\alpha + j}. \tag{8.3}$$

Remark 8.17. Since we assume $H(d\theta_i) = h(\theta_i)d\theta_i$ we can write Equation 8.2 in terms of "densities",

$$\pi(\theta) = \pi(\theta_n|\theta_{1:n-1})\pi(\theta_{n-1}|\theta_{1:n-2})...\pi(\theta_1)$$

so from Proposition 8.16,

$$\pi(\theta) = \prod_{j=0}^{n-1} \frac{\alpha h(\theta_{j+1}) + \sum_{i=1}^{j} \delta_{\theta_i}(\theta_j)}{\alpha + j},$$

with the sum equal zero when $j = 0$. We might wonder how useful this expression will actually be!

Remark 8.18. Proposition 8.16 follows immediately from the following proposition, using Proposition 8.12, so the next proposition is what we actually have to show.

Proposition 8.19. *If $\theta_1, ..., \theta_n \sim G$ with $G \sim \Pi(\alpha, H)$ then*

$$G|\theta_{1:n} \sim \Pi\left(\tilde{\alpha}_n, \tilde{H}_n\right)$$

with $\tilde{\alpha}_n = \alpha + n$ and

$$\tilde{H}_n = \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}. \tag{8.4}$$

Proof: see PS4 using induction with Propostion 8.21 below. [EOP!]

Proof of Proposition 8.16: since $G|\theta_{1:n-1}$ is a DP by Proposition 8.19, set $\tilde{G} = G|\theta_{1:n-1}$ with $\tilde{G} \sim \Pi(\tilde{\alpha}_{n-1}, \tilde{H}_{n-1})$. Now $\theta_n \sim \tilde{G}$ so marginally $\theta_n \sim \tilde{H}_{n-1}$ by Proposition 8.12. This is just Equation 8.3 in Proposition 8.16. [EOP]

Remark 8.20. We now give the first step of the induction. You can use the following proposition to carry out the induction step to prove Proposition 8.19 in PS4.

Proposition 8.21. *(step from $d\pi(\theta_1)$ to $d\pi(\theta_2|\theta_1)$) Suppose $G \sim \Pi(\alpha, H)$ and $\theta_1 \sim G$. The conditional distribution of $G|\theta_1$ is*

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

Furthermore,

$$d\pi(\theta_2|\theta_1) = \frac{\alpha H(d\theta_2) + \delta_{\theta_1}(d\theta_2)}{\alpha+1}$$

in Equation 8.2.

Proof: let $A_1, ..., A_r$ be a partition of Ω. Suppose $\theta_1 \in A_j$ and let $G(A_1) = g_1, ..., G(A_r) = g_r$ be a realisation of G with $\sum_i g_i = 1$ and $g_i \geq 0,\ i = 1, ..., r$. Let $g = (g_1, ..., g_r)$. Let

$$f(g) = \mathsf{Dirichlet}(g; \alpha H_1, ...\alpha H_r),$$

where $H_j \equiv H(A_j)$, denote the (Dirichlet) density of g for our given fixed partition $A_1, ..., A_r$. We want $f(g|\theta_1)$ since this is the density of $G(A_1), ..., G(A_r)|\theta_1$. By Bayes rule we have

$$f(g|\theta_1) \ \propto\ \pi(\theta_1|g_1, ..., g_r)\ f(g). \tag{8.5}$$

We are interested in the g-dependence here. Since $\theta_1 \in A_j$, and recalling $G(A_j) = g_j$

$$\begin{aligned}\pi(\theta_1|g) &= \pi(\theta_1, \theta_1 \in A_j|g)\\ &= \pi(\theta_1|\theta_1 \in A_j, g)\,\pi(\theta_1 \in A_j|g)\\ &= h(\theta_1|\theta_1 \in A_j)\,g_j,\end{aligned}$$

when $H(d\theta_1) = h(\theta_1)d\theta_1$ in our DP by Remark 8.13. Here θ_1 is independent of $G(A_1), ..., G(A_r)$ given $\theta_1 \in A_j$ as $G(A_1), ..., G(A_r)$ give the overall probability masses assigned to sets $A_1, ..., A_r$ and contain no information about distributions within sets. Dropping the expression above for $\pi(\theta_1|g)$ into Equation 8.5,

$$\begin{aligned}f(g|\theta_1) &\propto h(\theta_1|\theta_1 \in A_j) \times g_j \times g_1^{\alpha H_1 - 1} \times ... \times g_r^{\alpha H_r - 1}\\ &\propto g_1^{\alpha H_1 - 1 + \mathbb{I}_{\theta_1 \in A_1}} \times ... \times g_r^{\alpha H_r - 1 + \mathbb{I}_{\theta_1 \in A_r}},\end{aligned}$$

dropping the constant $h(\theta_1|\theta_1 \in A_j)$ which is independent of $g_1, ..., g_r$, giving

$$G(A_1), ..., G(A_r)|\theta_1 \sim \mathsf{Dirichlet}(\alpha H_1 + \mathbb{I}_{\theta_1 \in A_1}, ..., \alpha H_r + \mathbb{I}_{\theta_1 \in A_r}).$$

It follows that $G|\theta_1 \sim \Pi(\tilde{H}_1, \tilde{\alpha}_1)$ if we choose $\tilde{\alpha}_1$ and $\tilde{H}_1$ so that

$$\tilde{\alpha}_1 \tilde{H}_1(A_j) = \alpha H(A_j) + \mathbb{I}_{\theta_1 \in A_j}, \quad \text{for } j = 1, ..., r.$$

We must take $\tilde{\alpha}_1 = \alpha + 1$ and base distribution

$$\tilde{H}_1 = \frac{\alpha H + \delta_{\theta_1}}{\alpha+1}$$

so that this works for every H-measurable partition $A_1, ..., A_r$ and is normalised, so indeed

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

For the second part of the claim, the "posterior predictive distribution" $\theta_2|\theta_1$ is $\tilde{H}_1$ as $G|\theta_1 \sim \Pi(\tilde{\alpha}_1, \tilde{H}_1)$ and $\theta_2 \sim G|\theta_1$ so marginally $\theta_2 \sim \tilde{H}_1$ by Proposition 8.12. [EOP]

Remark 8.22. We now interpret this result. The updated base distribution for $G|\theta_1$ is

$$\tilde{H}_1(d\theta_2) = \frac{\alpha}{\alpha+1}h(\theta_2)d\theta_2 + \frac{1}{\alpha+1}\delta_{\theta_1}(d\theta_2).$$

This is a mixture. In order to simulate (θ_1, θ_2) marginally we can either simulate $G \sim \Pi(\alpha, H)$ and then $\theta_1, \theta_2 \sim G$ or we can simulate $\theta_1 \sim H$ and then $\theta_2|\theta_1 \sim \tilde{H}_1$. In order to simulate $\theta_2|\theta_1 \sim \tilde{H}_1$ we simulate $\theta_2 \sim h$ with probability $\alpha/(\alpha+1)$ and otherwise (*ie* with probability $1/(\alpha+1)$) we set $\theta = \theta_1$. Notice the atom at $\theta = \theta_1$ so we may have $\theta_2 = \theta_1$.

The generative process we wrote down involving G can now be replaced with the marginal process.

Proposition 8.23. *Consider the marginal generative process*

1. $\theta_1 \sim H$
2. *for* $j = 1, ..., n-1$
   (a) *With probability* $\alpha/(\alpha+j)$ *simulate* $\theta_{j+1} \sim H$.
   (b) *Otherwise simulate* $\theta_{j+1} \sim U\{\theta_1, ..., \theta_j\}$.

The distribution of $\theta = (\theta_1, ..., \theta_n)$ *is unchanged from Definition 8.14. No* G *in sight!*

Proof: we have seen that $\theta_{j+1}|\theta_{1:j} \sim \tilde{H}_j$ with

$$\tilde{H}_j = \frac{\alpha}{\alpha+j}H + \frac{1}{\alpha+j}\sum_{i=1}^{j}\delta_{\theta_i}$$

so this just simulates $\theta_{j+1}|\theta_{1:j} \sim \tilde{H}_j$ for $j = 1, ..., n-1$. But $d\pi(\theta_{j+1}|\theta_{1:j}) = \tilde{H}_j(d\theta_{j+1})$ by Proposition 8.16 and so by Equation 8.2 we have $\theta \sim d\pi(\theta)$. [EOP]

Remark 8.24. Sequential simulation of $(\theta_1, ..., \theta_n)$ may give repeated θ-values (if we choose one of the "old" θ's). Our new notation in the following section removes replicates.

8.2.6 Sequential simulation and repeated values

The marginal distribution $d\pi(\theta)$ on the LHS of Eqn 8.2 is not very helpful as it is a product of sums of distributions. However we can make a change of variables that makes it much easier to work with. We then stick to working with the new variables.

Definition 8.25. Let $\theta = (\theta_1, ..., \theta_n)$ be generated as in Proposition 8.23. Denote by $\theta^* = (\theta_1^*, ..., \theta_K^*)$ the $1 \le K \le n$ unique θ-values in $(\theta_1, ..., \theta_n)$. Let n_k give the number of times θ_k^* appears in $(\theta_1, ..., \theta_n)$. For $k = 1, ..., K$ let

$$S_k = \{i \in \{1, \dots n\} : \ \theta_i = \theta_k^*\},$$

so these are the indices of the θ-values matching θ_k^*. We have a partition $S = \{S_1, ..., S_K\}$ of $1, ..., n$ with $n_k = |S_k|$ and a new set of variables (θ^*, S) with which to represent the distribution $d\pi(\theta)$. For $k = 1, ..., K$ we can call S_k a "cluster".

Remark 8.26. This may be obvious, but just in case: If $\sigma \in \mathcal{P}_K$ is a permutation of $1, ..., K$ then we do not distinguish the two partitions $S = (S_1, ..., S_K)$ and $S = (S_{\sigma_1}, ..., S_{\sigma_K})$. We can adopt the convention that the k-indices to S_k, $k = 1, ..., K$ are ordered by the least elements in the sets, so $\min(S_k) < \min(S_{k'})$ if and only if $k < k'$. This gives us a unique cluster labeling.

Remark 8.27. The mapping

$$\theta^* = \theta^*(\theta),\ S = S(\theta)$$

is invertible: for $i = 1, ..., n$ let $k_i = \{k : i \in S_k\}$; since S is a partition, k_i is unique; set $\theta_i = \theta^*_{k_i}$ yielding back $\theta = \theta(\theta^*, S)$. For any S a partition of $\{1, ..., n\}$ and $\theta^* \in \Omega^K$ with $K = K(S)$ the number of sets in S, we may write $\theta = \theta(\theta^*, S)$ with $\theta = \theta(\theta^*(\theta), S(\theta))$.

Remark 8.28. Since θ^*_k appears n_k times in the sum over i in $\tilde{H}_n$ (Equation 8.4), we can write

$$\tilde{H}_n = \frac{\alpha}{\alpha + n} H + \frac{1}{\alpha + n} \sum_{k=1}^{K} n_k \delta_{\theta^*_k}.$$

In order to simulate this we sample H with probability $\alpha/(\alpha + n)$ and otherwise we choose a θ^*_k with a weight proportional to n_k.

Remark 8.29. Here is the marginal generative process for (θ^*, S):

1. $\theta^*_1 \sim H$; set $K = 1$, $S_1 = \{1\}$ and $S = \{S_1\}$.
   # $\theta_1 = \theta^*_1$ is of course unique and goes in a cluster by itself
2. for $j = 1, ..., n - 1$,
   with probability $\alpha/(\alpha + j)$ do (a) and otherwise do (b):
   (a) simulate $\theta^*_{K+1} \sim H$; set $S_{K+1} = \{j + 1\}$, $S \leftarrow S \cup \{S_{K+1}\}$ and $K \leftarrow K + 1$.
   # if we generate a "new" $\theta_{j+1} = \theta^*_{K+1}$ then it starts a new cluster
   (b) for $k = 1, ..., K$ set $n_k = |S_k|$; simulate $k \sim (n_1, ..., n_K)/j$; set $S_k \leftarrow S_k \cup \{j + 1\}$.
   # choose $\theta_{j+1} = \theta^*_k$ weighted by n_k, $k = 1, ..., K$ and it goes in an old cluster.

We have reorganised the process so each distinct parameter is simulated just once. If we set $\theta = \theta(\theta^*, S)$ then $\theta \sim d\pi(\theta)$ as in Definition 8.14.

8.2.7 The joint distribution of θ^*, S

We observed that $d\pi(\theta)$ was a mess. What does $d\pi(\theta^*, S)$ look like in the new variables? We calculate it in this section and the next.

First we think about the parameter space. Recall that the base distribution H has probability space $(\Omega, \mathcal{B}, H)$. When we look at (θ^*, S), output from the marginal generating process in Remark 8.29, we see we can realise any partition S of $\{1, ..., n\}$ and then any $\theta^* \in \Omega^K$ where $K = K(S)$ is the number of clusters in S.

Definition 8.30. Let $[n] = \{1, ..., n\}$ and let $\Xi_{[n]}$ be the set of all partitions of $[n]$. The distribution $d\pi(\theta^*, S)$ is defined on the space

$$\Omega^* = \bigcup_{S \in \Xi_{[n]}} \Omega^{K(S)} \times \{S\}.$$

Proposition 8.31. *Let $(\theta^*, S) \sim d\pi(\theta^*, S)$ be determined by the process in Remark 8.29. Its distribution is*

$$d\pi(\theta^*, S) = \pi_S(S) d\pi(\theta^*|S), \quad (\theta^*, S) \in \Omega^*$$

with $\pi_S(S)$ a probability distribution over partitions and

$$d\pi(\theta^*|S) = \prod_{k=1}^{K} H(d\theta^*_k)$$

when $S = (S_1, ..., S_K)$ so $K = K(S)$ above, and $H(d\theta^*_k) = h(\theta^*_k) d\theta^*_k$.

Proof: the first equation follows by conditioning on the discrete random variable S and the expression for $d\pi(\theta^*|S)$ follows from the marginal generating process in Remark 8.29, where the $\theta^*_k,\ k = 1, ..., K$-values we generate at step (a) are independent samples from H. [EOP]

Remark 8.32. We work out $\pi_S(S),\ S \in \Xi_{[n]}$ in the next section, where call it $P_{\alpha,[n]}(S)$.

Remark 8.33. The dimension of $\theta^* \in \Omega^K$ is random, as S is random, so $K = K(S)$ is random. We have in fact $\dim(\Omega^K) = pK$, as we take $\Omega = \mathbb{R}^p$ and $H(d\theta^*_k) = h(\theta^*_k)d\theta^*_k$ with $d\theta^*_k$ the element of volume in $\mathbb{R}^p$.

Remark 8.34. If $A \subset \Omega^*$ is a set, chosen so the integrals below exist, and $(\theta^*, S) \sim d\pi(\theta^*, S)$ then

$$\Pr(\ (\theta^*, S) \in A) = \sum_{S \in \Xi_{[n]}} \int_{\Omega^{K(S)}} \mathbb{I}_{(\theta^*,S)\in A} \left[P_{\alpha,[n]}(S) \prod_{k=1}^{K(S)} h(\theta^*_k) \right] d\theta^*_1, ..., d\theta^*_{K(S)}$$

so this is how we do integration in Ω^*. We use Monte-Carlo to evaluate these integrals!

8.2.8 The Chinese Restaurant Process

We now calculate $P_{\alpha,[n]}(S)$ (ie, $\pi_S(S)$). The sequential simulation of parameters in Remark 8.29 is analogous to restaurant seating! The authors of the CRP paper dined regularly in Chinese restaurants and were impressed with their seemingly infinite capacity - a table could always be found!

Definition 8.35. (*Chinese Restaurant Process (CRP)*)

1. There is $j = 1$ one customer in the restaurant seated at table $k = 1$. After the first customer arrives, there are $n^1_k = 1$ people seated at table $k = 1$, and $K_1 = 1$ tables are occupied.
2. for $j = 1, ..., n-1$
   (a) the $j+1$'st arrival chooses new table $K_j + 1$ with probability $\alpha/(\alpha + j)$ and table k with probability $n^j_k/(\alpha + j)$.
   (b) after customer $j + 1$ arrives, there are n^{j+1}_k people seated at table k, and K_{j+1} tables are occupied.

After all n customers are seated, the CRP has shared n customers over $K = K_n$ tables. For $k = 1, ..., K$, set S_k lists customers at table k and $n_k = n^j_k$ with $j = n$ gives the final table counts, so $n_k = |S_k|$. Let $\pi_S(S) = P_{\alpha,[n]}(S)$ give the probability to realise partition S in this process.

Remark 8.36. If at the end we put an independent parameter $\theta^*_k \sim H$ on table $k = 1, ..., K$ (a single dish, which is shared!) then this is the same as the (θ^*, S) algorithm in Remark 8.29 above: the θ^*'s are independent draws from H and we can realise these once we know how many clusters there are in S. Together, $\theta(\theta^*, S) \sim G$ with $G \sim \Pi(\alpha, H)$ by Remark 8.29.

Exercise 8.37. (see PS4) Show from the CRP that

$$E(K) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}$$

Proposition 8.38. *The CRP in Defintion 8.35 realises partition $S \in \Xi_{[n]}$ with probability*

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

Remark 8.39. Here is some intuition from an example. Suppose table assignment-sequence is

$$T = (1, 1, 2, 1, 2, 3, 3, 2, 2, 4)$$

for $n = 10$ customers so $S = \{\{1, 2, 4\}, \{3, 5, 8, 9\}, \{6, 7\}, \{10\}\}$. Table assignment T and partition S are 1 to 1, so multiplying together the probabilities for the events leading to T,

$$\begin{aligned} P_{\alpha,[n]}(S) &= 1 \times \frac{1}{\alpha+1} \times \frac{\alpha}{\alpha+2} \times \frac{2}{\alpha+3} \times \frac{1}{\alpha+4} \times \frac{\alpha}{\alpha+5} \times \frac{1}{\alpha+6} \times \frac{2}{\alpha+7} \times \frac{3}{\alpha+8} \times \frac{\alpha}{\alpha+9} \\ &= \alpha^3 \; 2! \; 3! \; 1! \; 0! \prod_{i=2}^{10} \frac{1}{\alpha+i-1} \\ &= \alpha^4 \; \Gamma(3) \; \Gamma(4) \; \Gamma(2) \; \Gamma(1) \prod_{i=1}^{10} \frac{1}{\alpha+i-1} \\ &= \alpha^K \left[\prod_{k=1}^{K} \Gamma(n_k)\right] \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \end{aligned}$$

with $K = 4$ and $n = 10$.

Exercise 8.40. Use $\Gamma(x+1) = x\,\Gamma(x)$ for $x > 1$ to show $\Gamma(\alpha+n) = \Gamma(\alpha)\prod_{i=1}^{n}(\alpha+i-1)$ for $\alpha > 0$.

Proof of Proposition 8.38: for $i = 2, ..., n$ 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 chose table k with probability $\alpha/(\alpha+i_1-1)$ (and we already accounted for the denominator).

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 with probability

$$\frac{j-1}{\alpha+i_j-1} = \frac{i_j-1}{\alpha+i_j-1} \times \frac{j-1}{i_j-1}$$

so the numerator factor from table k 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

$$\begin{aligned} P_{\alpha,[n]}(S) &= \alpha^{K-1} \prod_{k=1}^{K} (n_k-1)! \prod_{i=2}^{n} \frac{1}{\alpha+i-1} \\ &= \alpha^K \prod_{k=1}^{K} (n_k-1)! \prod_{i=1}^{n} \frac{1}{\alpha+i-1} \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^K \prod_{k=1}^{K} \Gamma(n_k) \end{aligned}$$

where we extended the product in the denominator down to one in the first step and in the second step we used Exercise 8.40. [EOP]

Remark 8.41. Notice that we must have

$$\sum_{S \in \Xi_{[n]}} P_{\alpha,[n]}(S) \;=\; 1$$

since $P_{\alpha,[n]}(S)$ is a probability mass function.

Exercise 8.42. (see PS4 and note this is a quite different use of permutation to Remark 8.26) Let $\mathcal{P}_n$ be the set of all permutations of $\{1,...,n\}$. For $\sigma \in \mathcal{P}_n$ let $S(\sigma)$ be the partition obtained by permuting the customer labels in S according to σ. For example if $S = \{\{1,2\},\{3\}\}$ and $\sigma = (3,1,2)$ then $S(\sigma) = \{\{\sigma_1,\sigma_2\},\{\sigma_3\}\} = \{\{3,1\},\{2\}\}$. Show that $P_{\alpha,[n]}(S) = P_{\alpha,[n]}(S(\sigma))$ (CRP outcomes don't depend on customer arrival order). *Hint: n_k doesnt change.*

Exercise 8.43. (see PS4) Let $S \sim P_{\alpha,[n]}$ and

$$S^{-i} = (S_1^{-i},...,S_{K^{-i}}^{-i})$$

be the partition with $i \in \{1,...,n\}$ removed. For example if $S = \{\{1,2\},\{3\}\}$ then $S^{-3} = \{\{1,2\}\}$ (so we may drop a whole cluster).

Let $P(S^{-i})$ give the distribution of S^{-i} (so run the CRP on $1,...,n$ and then remove i). Here $K^{-i} = K-1$ if we create an empty cluster when we remove i and otherwise $K^{-i} = K$. Let $P_{\alpha,[n]\setminus\{i\}}(S')$, $S' \in \Xi_{[n]\setminus\{i\}}$ give the probability to realise S' if i is removed from the list of customers at the start (so run the CRP on customer labels $[n]\setminus\{i\}$).

Show that

$$P(S^{-i}) = P_{\alpha,[n]\setminus\{i\}}(S^{-i})$$

(which we can write $S^{-i} \sim P_{\alpha,[n]\setminus\{i\}}$) and

$$\Pr(i \in S_k | S^{-i}) = P_{\alpha,[n]}(S)/P_{\alpha,[n]\setminus\{i\}}(S^{-i})$$

where S is obtained from S^{-i} by adding i to partition set S_k in S in the second equation.
Hint: Write the result of Exercise 8.42 in the form $P_{\alpha,[n]}(S) = P_{\alpha,\sigma}(S)$, where the σ subscript indicates the customer order (we are just permuting labels), and make i the last customer to arrive.

8.3 Inference for a Dirichlet process mixture

Suppose our data are n independent samples $y_i \sim f(y_i|\theta_i), i = 1,...,n$ from a mixture with an unknown number of components. The likelihood for $\theta = (\theta_1,...,\theta_n)$ is

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta_i).$$

There seems to be a parameter for every observation, but the θ's are equal within clusters. In a DP model for a mixture $y_i \sim f(y_i|\theta_i)$ our prior for $\theta = (\theta_1,...,\theta_n)$ is

$$\theta_1,...,\theta_n \sim G \quad \text{with} \quad G \sim \Pi(\alpha, H).$$

In (θ^*, S) notation, the observation model is $f(y|\theta) = f(y|\theta(\theta^*,S))$ or equivalently $f(y|\theta^*,S)$. For $i \in S_k$ we have $\theta_i = \theta_k^*$ so

$$f(y|\theta^*,S) = \prod_{k=1}^{K} f(y_{S_k}|\theta_k^*)$$

where $y_{S_k} = (y_i)_{i\in S_k}$ and $f(y_{S_k}|\theta_k^*) = \prod_{i\in S_k} f(y_i|\theta_k^*)$.

Theorem 8.44. *The posterior for θ under a DP prior $\theta \sim G$ with $G \sim \Pi(\alpha,H)$ is given by*

$$\pi(\theta^*, S|y) \propto f(y|\theta^*,S)\,\pi(\theta^*|S)\,P_{\alpha,[n]}(S), \tag{8.6}$$

with $\theta = \theta(\theta^*,S)$ and $(\theta^*,S) \in \Omega^*$ given in Definition 8.30.

Proof: apply Bayes rule and Proposition 8.31 with $\pi_S = P_{\alpha,[n]}$ by Proposition 8.38. [EOP]

Remark 8.45. I called this a theorem because it is really the point and outcome of Chapter 8. Equation 8.6 is a Very Useful Relation. It takes us straight to an expression for the posterior in a general DP-process mixture! We can just write this straight down once we are clear on the observation model.

Remark 8.46. In detail, from Proposition 8.31 and Proposition 8.38,

$$\pi(\theta^*, S|y) \propto \prod_{k=1}^{K} [\, \alpha\, \Gamma(n_k)\, h(\theta_k^*)\, f(y_{S_k}|\theta_k^*)\,], \tag{8.7}$$

where $S = (S_1, ..., S_K)$ so $K = |S|$ and $n_k = |S_k|,\ k = 1, ..., K$.

Remark 8.47. This is a model averaging setup in which π_M in Chapter 6 is $\pi_S = P_{\alpha,[n]}$ here, so S plays the role of a model index. The joint posterior distribution of model and parameter is $\pi(\theta^*, S|y)$. The model space indexed by $S \in \Xi_{[n]}$ is finite (but large). The number of components K in θ^* is a random variable.

8.3.1 Normal mixture for the Galaxy data

Recall the Galaxy radial velocity data which we saw in Section 7.4. It is natural to model this via a mixture of normals. However, we do not know the number of components in the mixture and we don't know which observation is drawn from which mixture component. Our setup here differs from Section 7.4 as we will take explicit cluster labels rather than weights. We will have $y_i \sim N(y_i; \mu_i, \sigma_i^2)$ where μ_i and σ_i are the parameters for the cluster to which i belongs.

In terms of our (θ^*, S) notation, each component of the mixture has an unknown mean and variance, $\theta_k^* = (\mu_k^*, \sigma_k^{*2})$. Our base distribution H gives a prior for the components with density

$$h(\theta_k^*) = h_\mu(\mu_k^*) h_\sigma(\sigma_k^{*2}).$$

If $S = (S_1, ..., S_K)$ is a partition of $[n] = \{1, 2, ..., n\}$ with $n = 82$, and $i \in S_{k_i}$ then

$$y_i|S, \mu^*, \sigma^* \ \sim\ N(\mu_{k_i}^*, {\sigma_{k_i}^*}^2).$$

This determines the likelihood. Our priors are

$$h_\mu(\mu_k^*) = N(\mu_k^*; \mu_0, \sigma_0^2)$$

and

$$h_\sigma(\sigma_k^{*2}) = \text{I}\Gamma(\sigma_k^{*2}; \alpha_0, \beta_0).$$

with $\mu_0 = 20, \sigma_0 = 10, \alpha_0 = 2$ and $\beta_0 = 1/9$ fixed hyper-parameters, so from Equation 8.7,

$$\begin{aligned}
\pi(S, \mu^*, \sigma^*|y) \quad &\propto \quad f(y; \mu^*, \sigma^*, S)\pi(\mu^*, \sigma^*|S)P_{\alpha,[n]}(S) \\
&\propto \quad \prod_{k=1}^{K} \prod_{i \in S_k} N(y_i; \mu_k^*, \sigma_k^{*2}) \\
&\times \quad \prod_{k=1}^{K} N(\mu_k^*; \mu_0, \sigma_0^2)\, \text{I}\Gamma(\sigma_k^{*2}; \alpha_0, \beta_0) \\
&\times \quad \alpha^K \prod_{k=1}^{K} \Gamma(n_k).
\end{aligned}$$

We dropped the denominator in the expression for $P_{\alpha,[n]}(S)$ as it does not depend on S. Comparing with Equation 7.11 for the posterior we wrote down in Section 7.4, there is no $n!$ - or $K!$ - that is because $S = (S_1, ..., S_K)$ is ordered by the labelling convention in Remark 8.26). Any given partition of the data $(y_1, ..., y_n)$ into clusters has a unique S. The posteriors are not simply related as the priors are qualitatively different. Before we had a Poisson prior over the number of clusters, while here the prior for K is determined by $P_{\alpha,[n]}$, and isn't Poisson.

Remark 8.48. We take Normal/inv-Gamma for the μ^*/σ^*-prior to keep things simple and conjugate, so that we can can Gibbs-sample μ^*, σ^*. If we just did straightforward MH-MCMC on μ^* and σ^* that wouldn't be necessary.

Conjugate priors are popular in this field as it allows us to integrate out $\theta^* = (\mu^*, \sigma^*)$ completely and just sample the discrete distribution $\pi(S|y)$. This is the "collapsed Gibbs sampler". It is efficient. If our purpose is clustering, S is all we need anyway. The downside is we cant model μ^* and σ^* with freedom.

Remark 8.49. We took $\alpha = 1$ for the purpose of this example. This controls the prior distribution on the number of clusters. I used simulation (of the CRP) to check this distribution was sensible. The prior mean is

$$E(K) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}.$$

which is about $E(K) = 5$ here. We would take α a bit larger if our prior elicitation favored larger values. When we did reversible jump our prior mean for the number of components was about ten. It is sometimes straightforward to impose a hyper-prior on α and infer it along with everything else using Metropolis-Hastings.

8.3.2 Gibbs sampler for the mixture parameters μ^*, σ^*

Iterate through the parameters sampling them conditionally. The conditional posterior distribution for μ_k^* given everything else is

$$\pi(\mu_k^*|\mu_{-k}^*, \sigma^*, y) \propto N(\mu_k^*; \mu_0, \sigma_0^2) \prod_{i \in S_k} N(y_i; \mu_k^*, \sigma_k^{*2}).$$

We can complete the square and find $\mu_k^*|\sigma_k^*, y \sim N(a, b)$ with

$$a = b\left(\frac{n_k \bar{y}_k}{\sigma_k^{*2}} + \frac{\mu_0}{\sigma_0^2}\right), \qquad b = \left(\frac{n_k}{\sigma_k^{*2}} + \frac{1}{\sigma_0^2}\right)^{-1},$$

where $n_k = |S_k|$ and $\bar{y}_k = n_k^{-1} \sum_{i \in S_k} y_i$.

A similar calculation gives $\sigma^{*2}|\mu_k^*, y \sim \text{I}\Gamma(c, d)$ with

$$c = \alpha_0 + n_k/2, \qquad d = \beta_0 + \frac{1}{2} \sum_{i \in S_k} (y_i - \mu_k^*)^2.$$

8.3.3 Gibbs sampler for the partition

We need an update that operates on $S = (S_1, ..., S_K)$. The update must be irreducible on the space of partitions $\Xi_{[n]}$. The idea is to pick an entry $i \in [n]$ at random and move it to a randomly chosen

cluster. The new cluster could be cluster $K+1$, so we add a cluster and increase dimension. When we remove i from its cluster k_i we may empty that cluster, so we remove that cluster and decrease dimension. We actually use a proposal with an acceptance probability equal one, so effectively a Gibbs sampler. The easiest way to set this up in a general way and verify it is correct is to use reversible jump.

Here is some motivation. Consider the sequence of pairs $(y_j, \theta_j), j = 1, 2, ..., n$. The generative model for our data is

1. Simulate $G \sim \Pi(\alpha, H)$
2. For $j = 1, ..., n$ simulate $\theta_j \sim G$ and $y_j \sim f(\cdot|\theta_j)$.

For our galaxy data $\theta_j = (\mu_j, \sigma_j)$. We can write this using the CRP as follows:

1. $\theta_1^* \sim H$; set $K = 1$, $S_1 = \{1\}$ and $S = \{S_1\}$. Set $\theta_1 = \theta_1^*$ and simulate $y_1 \sim f(\cdot|\theta_1)$.
2. for $j = 1, ..., n-1$,
   set $n_k = |S_k|$, $k \in [K]$ and choose $k' \sim p$, $p \propto (n_1, ..., n_K, \alpha)$.
   If $k' = K+1$ then do (a) otherwise do (b):
   (a) simulate $\theta_{K+1}^* \sim H$; set $\theta_{j+1} = \theta_{K+1}^*$ and simulate $y_{j+1} \sim f(\cdot|\theta_{j+1})$; set $S_{K+1} = \{j+1\}$, $S \leftarrow S \cup \{S_{K+1}\}$ and $K \leftarrow K+1$.
   (b) set $S_{k'} \leftarrow S_{k'} \cup \{j+1\}$; set $\theta_{j+1} = \theta_{k'}^*$ and simulate $y_{j+1} \sim f(\cdot|\theta_{j+1})$.

The (θ_j, y_j)'s are exchangeable so we may make i the last arrival in the CRP when we generate y, θ^*, S. If we remove i from S at the end then we know S^{-i} and y_i. If removing i created an empty cluster in S then the last arrival i must have created cluster $k_i = K$ in S which is cluster $k_i = K^{-1} + 1$ in S^{-i}. In this case we already know $\theta_{K^{-i}+1}^* = \theta_K^*$. If removing i doesnt leave an empty cluster then i didnt add a cluster on arrival, so we dont know anything about $\theta_{K^{-i}+1}^*$, except that $\theta_{K^{-i}+1}^* \sim H$, so we simulate it. In both cases the choice of cluster for i is conditioned on all of $S^{-1}, (\theta_1^*, ..., \theta_{K^{-i}}^*, \theta_{K^{-i}+1}^*)$ and y_i where in the singleton case we set $\theta_{K^{-i}+1}^* = \theta_{k_i}^*$ and otherwise we have to simulate $\theta_{K^{-i}+1}^* \sim H$. Referring to Equations 8.8 and 8.9 below, if k' is the new random cluster we choose for i then the conditional probabilities for the assignment of i to a cluster are

$$\Pr(k' = k|\theta_1^*, ..., \theta_{K^{-i}+1}^*, y_i, S^{-i}) \propto \begin{cases} \vec{q}_k(S, k_i) & \text{in (8.8) with } \theta_{K^{-i}+1}^* \sim H \text{ if } n_{k_i} > 1; \\ \overleftarrow{q}_k(S, k_i) & \text{in (8.9) with } \theta_{K^{-i}+1}^* = \theta_{k_i}^* \text{ if } n_{k_i} = 1. \end{cases}$$

Proposition 8.50. *The acceptance probabilities for the proposals in Figure 18 are all equal one.*

Proof. The cases where $k' = k_i$, no change, are fairly clear. There are three distinct update types, where the dimension goes up, down, or stays the same. It goes up when $n_{k_i} > 1$ and we choose $k' = K+1$ using the cluster selection distribution $\vec{p}(S, k_i)$ just as above. When we are reversing the move the current state is (θ', S') and we must select i to move and propose to move it into S'_{k_i}, deleting S'_{K+1} and restoring S_{k_i}. We will be using the cluster selection distribution $\overleftarrow{p}(S', K+1)$. We have,

$$\begin{aligned} \alpha(\theta', S'|\theta^*, S) &= \min\left\{1, \frac{\pi(\theta', S'|y)}{\pi(\theta, S|y)} \times \frac{n^{-1}\overleftarrow{p}_{k_i}(S', K+1)}{n^{-1}h(\theta'_{K+1})\vec{p}_{K+1}(S, k_i)}\right\} \\ &= \min\left\{1, \frac{f(y_i|\theta'_{K+1})h(\theta'_{K+1})\alpha^{K+1}\Gamma(n_{K+1})\Gamma(n_{k_i}-1)}{f(y_i|\theta^*_{k_i})\alpha^K\Gamma(n_{k_i})} \times \frac{\overleftarrow{q}_{k_i}(S', K+1)}{h(\theta'_{K+1})\vec{q}_{K+1}(S, k_i)}\right\} \\ &= \min\left\{1, \frac{f(y_i|\theta'_{K+1})h(\theta'_{K+1})\alpha}{f(y_i|\theta^*_{k_i})(n_{k_i}-1)} \times \frac{(n_{k_i}-1)f(y_i|\theta^*_{k_i})}{h(\theta'_{K+1})\alpha f(y_i|\theta'_{K+1})}\right\}, \end{aligned}$$

Let $X_t = (\theta^*, S)$.

1. Choose $i \sim U\{1, ..., n\}$ and suppose $i \in S_{k_i}$.
2. If $n_{k_i} > 1$ - removing i wont empty S_{k_i}.
   (a) Simulate $\theta'_{K+1} \sim H$
   (b) let
   $$\vec{q}_k(S, k_i) = \begin{cases} (n_{k_i} - 1) f(y_i|\theta^*_{k_i}) & \text{if } k = k_i, \\ \alpha f(y_i|\theta'_{K+1}) & \text{if } k = K+1, \\ n_k f(y_i|\theta^*_k) & \text{for } k \in [K] \setminus \{k_i\}. \end{cases}, \tag{8.8}$$
   (c) new cluster $k' \sim (\vec{p}_k)_{k=1,...,K+1}$ where $\vec{p}_k = \vec{q}_k / \sum_{j=1}^{K+1} \vec{q}_j$.
      i. (no change) If $k' = k_i$ then $(\theta', S') = (\theta^*, S)$.
      ii. (increase dim) If $k' = K+1$ then $\theta' = (\theta^*, \theta'_{K+1})$, $S'_{k_i} = S_{k_i} \setminus \{i\}$, $S'_{K+1} = \{i\}$ $S'_k = S_k,\ k \in [K] \setminus \{k_i\}$
      iii. (fixed dim) If $k' \neq k_i, K+1$ then $\theta' = \theta^*$ and $S'_{k_i} = S_{k_i} \setminus \{i\}$, $S'_{k'} = S_{k'} \cup \{i\}$ and $S'_k = S_k,\ k \in [K] \setminus \{k_i, k'\}$.
3. Else if $n_{k_i} = 1$ - removing i will empty S_{k_i}.
   (a) let
   $$\overleftarrow{q}_k(S, k_i) = \begin{cases} \alpha f(y_i|\theta^*_k) & \text{if } k = k_i \\ n_k f(y_i|\theta^*_k) & \text{for } k \in [K] \setminus \{k_i\}. \end{cases}, \tag{8.9}$$
   (b) new cluster $k' \sim (\overleftarrow{p}_k)_{k=1,...,K}$ where $\overleftarrow{p}_k = \overleftarrow{q}_k / \sum_{j=1}^{K} \overleftarrow{q}_j$.
      i. (no change) If $k' = k_i$ then $(\theta', S') = (\theta^*, S)$.
      ii. (decrease dim) If $k' \neq k_i$ then $S'_{k_i} = \emptyset$, $S'_{k'} = S_{k'} \cup \{i\}$ and $S'_k = S_k,\ k \in [K] \setminus \{k_i, k'\}$.
      Now remove $S'_{k_i} = \emptyset$: $S' \leftarrow S' \setminus S'_{k_i}$ and $\theta' = \theta^*_{-k_i}$.
4. $X_{t+1} = (\theta', S')$.

We may need to resort the cluster labels to meet our labeling convention $\min(S_k) < \min(S_{k'}) \Rightarrow k < k'$ in step 2(c)ii,iii and 3(b)ii.

Figure 18: Gibbs sampler for the random partition S. I reformatted this, otherwise unchanged.

which equals one. Most of this is just careful accounting, but the key step is canceling the denominators in $\overleftarrow{p}_{k_i}(S', K+1)$ and $\vec{p}_{K+1}(S, k_i)$. These are equal,

$$\begin{aligned} \sum_{j=1}^{K(S')} \overleftarrow{q}_j(S', K+1) &= n'_1 f(y_i|\theta'_1) + ... + n'_{k_i} f(y_i|\theta'_{k_i}) + ... + \alpha f(y_i|\theta'_{K(S')}) \\ &= n_1 f(y_i|\theta'_1) + ... + (n_{k_i} - 1) f(y_i|\theta^*_{k_i}) + ... + \alpha f(y_i|\theta'_{K(S)+1}) \\ &= \sum_{j=1}^{K(S)+1} \vec{q}_j(S, k_i). \end{aligned}$$

Notice that in the reverse move we are emptying S'_{K+1} so it contributes a term $\alpha f(y_i|\theta'_{K+1})$ per case 1 of Equation 8.9 (identifying k_i there with $K+1$, the label of the cluster we are emptying). Also, since i was removed from S_{k_i}, we have $n'_{k_i} = n_{k_i} - 1$. In the forward move the new candidate cluster S'_{K+1} gives a term $\alpha f(y_i|\theta'_{K+1})$ per case 2 of Equation 8.8.

The case where $n_{k_i} = 1$ and we select $k' \neq k_i$, so we delete a cluster, is the reverse of the move above. If you would like to check that I suggest you use the labeling *before* the empty set was removed and the partitions relabelled.

The case where $n_{k_i} > 1$ and we dont select $k' = k_i$ or $k' = K+1$ is interesting. We are including

S'_{K+1} in the list of candidate clusters, so we generate $\theta'_{K+1} \sim h$ in both directions and use $\vec{p}$ (as $n'_{k'} = n_{k'} + 1$ must be bigger than one). We have

$$\begin{aligned}
\alpha(\theta', S'|\theta^*, S) &= \min\left\{1,\ \frac{\pi(\theta', S'|y)}{\pi(\theta, S|y)} \ \times\ \frac{n^{-1}h(\theta'_{K+1})\vec{p}_{k_i}(S', k')}{n^{-1}h(\theta'_{K+1})\vec{p}_{k'}(S, k_i)}\right\} \\
&= \min\left\{1,\ \frac{f(y_i|\theta^*_{k'})\Gamma(n_{k'}+1)\Gamma(n_{k_i}-1)}{f(y_i|\theta^*_{k_i})\Gamma(n_{k'})\Gamma(n_{k_i})} \ \times\ \frac{\vec{q}_{k_i}(S', k')}{\vec{q}_{k'}(S, k_i)}\right\} \\
&= \min\left\{1,\ \frac{f(y_i|\theta^*_{k'})n_{k'}}{f(y_i|\theta^*_{k_i})(n_{k_i}-1)} \ \times\ \frac{(n_{k_i}-1)f(y_i|\theta^*_{k_i})}{n_{k'}f(y_i|\theta^*_{k'})}\right\},
\end{aligned}$$

so again, the acceptance probability is one. Checking the denominators of $\vec{p}_{k_i}(S', k')$ and $\vec{p}_{k'}(S, k_i)$,

$$\begin{aligned}
\sum_{j=1}^{K(S')+1} \vec{q}_j(S', k') &= n'_1 f(y_i|\theta'_1) + ... + n'_{k_i} f(y_i|\theta'_{k_i}) + ... + (n'_{k'} - 1)f(y_i|\theta'_{k'}) + ... + \alpha f(y_i|\theta'_{K(S')+1}) \\
&= n_1 f(y_i|\theta^*_1) + ... + (n_{k_i} - 1)f(y_i|\theta'_{k_i}) + ... + (n_k + 1 - 1)f(y_i|\theta^*_{k'}) + ... + \alpha f(y_i|\theta'_{K(S)+1}) \\
&= \sum_{j=1}^{K(S)+1} \vec{q}_j(S, k_i)
\end{aligned}$$

so they cancel again.

It may be unclear what the role of θ'_{K+1} is in this update. In the moves adding and deleting clusters it is part of what we called $g(u)$ in the RJ-MCMC algorithm (and $u = \theta'_{K+1}$ so the Jacobian equals one). In this fixed dimension update θ'_{K+1} behaves as part of what we called $\rho_{m,m'}$ - we select an update from a continuously indexed set of possible updates, each of which satisfies DB. This is a fixed dimension update so DB is easily checked, so you can check for yourself it works. □

Remark 8.51. This is implemented for our mixture model application where $\theta^*_k = (\mu^*_k, \sigma^{*2}_k)$,

$$h(\theta^*_k) = N(\mu^*_k; \mu_0, \sigma^2_0)\, \text{I}\Gamma(\sigma^{*2}_k; \alpha_0, \beta_0)$$

and $f(y_j|\theta^*_k) = N(y_i; \mu^*_k, \sigma^{*2}_k)$, $k = 1, ..., K$ and $i = 1, ..., n$, in the code for this lecture.

8.3.4 Results for the Galaxy Radial Velocity data DP-mixture

The R-code and further detail of the algorithm are available with these notes. We ran the code and generated samples $(\theta^{*,(t)}, S^{(t)}), t = 1, 2, ..., T$ from the joint posterior distribution over the parameters and partition with $\theta^{*,(t)} = (\mu^{*,(t)}, \sigma^{*,(t)})$, $t = 1, ..., T$.

The figure below shows an estimate $\widehat{p(y'|y)}$ of the posterior predictive distribution $p(y'|y)$ (black line) at each point $y' \in \mathbb{R}$ on the x-axis (notice $y \in \mathbb{R}^n$ is a different animal). Since

$$p(y'|y) = \sum_{S \in \Xi_{[n]}} \int_{\Omega^{K(S)}} p(y'|\theta^*, S)\pi(\theta^*, S|y) d\theta^*_1, ..., d\theta^*_{K(S)},$$

we use the natural estimate (with $p(y'|\theta^{*,(t)}, S^{(t)}) = f(y'|\theta^{*,(t)}, S^{(t)})$ here)

$$\widehat{p(y'|y)} = \frac{1}{T}\sum_{t=1}^{T} f(y'|\theta^{*,(t)}, S^{(t)}).$$

Now, dropping the "(t)" superscript,

$$f(y'|\theta^*, S) = \int_\Omega p(y', \theta'|\theta^*, S) d\theta',$$

so using the conditional independence we can see in the generative model we get,

$$= \int_\Omega f(y'|\theta') p(\theta'|\theta^*, S) d\theta'$$

and taking $p(\theta'|\theta^*, S)$ from Equation 8.3 in θ^*, S notation,

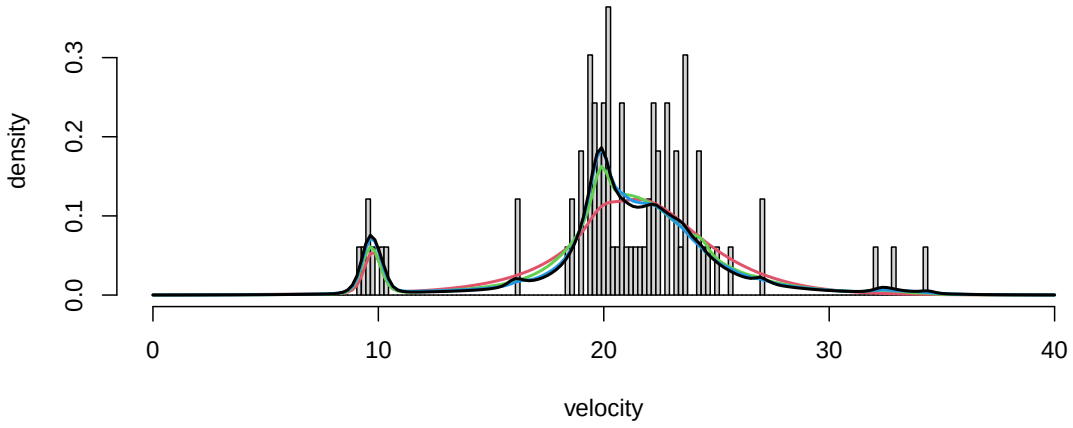
$$= \int_\Omega f(y'|\theta') \left[\frac{\alpha h(\theta') + \sum_{k=1}^K n_k \delta_{\theta_k^*}(\theta')}{\alpha + n} \right] d\theta'$$

$$= \frac{\alpha}{\alpha + n} p(y') + \sum_{k=1}^K f(y'|\theta_k^*) \frac{n_k}{\alpha + n}.$$

An unbiased estimate for $p(y') = E_{\theta'}(f(y'|\theta'))$ is given by $f(y'|\theta_{K+1}^*)$ with $\theta_{K+1}^* \sim h(\cdot)$ (the $K+1$ index because y', θ' are in a new cluster). In the conjugate-prior setting we dont need to estimate $p(y')$. I prefer a general setup, so my code forms an estimate of $p(y')$ using a fresh set of prior samples, $\theta_{K+1}^{*,(t)} \sim h(\cdot)$ iid for $t = 1, ..., T$,

$$\widehat{p(y'|y)} = \frac{1}{T} \sum_{t=1}^T \left[\frac{\alpha}{\alpha + n} f(y'|\theta_{K+1}^{*,(t)}) + \sum_{k=1}^{K^{(t)}} f(y'|\theta_k^{*,(t)}) \frac{|S_k^{(t)}|}{\alpha + n} \right]$$

with $K^{(t)} = K(S^{(t)})$.



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. We estimate the posterior predictive distribution (black) and the predictive distribution conditioned on K = 3,4 and 5 components (red, green and blue) and plot these PPD's over the data. The fit seems reasonable for the model averaged predictive distribution in black, and for 4 or more clusters. The posterior distribution of K gives us the posterior distribution of the unknown number of clusters.

The code explores other visualisations of the output, such as the posterior distribution of the dimension-parameter $2K$.

8.4 Appendices

8.4.1 Appendix for Section 8.2.3: The DP as the limit of the Multinomial DP

[The material in this Appendix is outside the course. It is a neat application of agglomeration.]

The DP is also obtained in a concrete construction as the limit $M \to \infty$ of the multinomial Dirichlet process. Let $A_1, ..., A_r$ be any fixed H-measurable partition of Ω. The sample space of $G_M(A_1), ..., G_M(A_r)$ is $\Lambda = \{g \in (0,1)^r : \sum_{i=1}^r g_i = 1\}$. Let $G_M(A_i) = g_i, i = 1, ..., r$ be a realisation and write $g = (g_1, ..., g_r)$.

Let $f_M(g), g \in \Lambda$ denote the joint density of $G_M(A_1), ..., G_M(A_r)$. We will see that this density exists for each M.

Proposition 8.52. *For each $g \in \Lambda$,*

$$\lim_{M\to\infty} f_M(g) = \mathit{Dirichlet}(g; \alpha H(A_1), ..., \alpha H(A_r))$$

so $G_M(A_1), ..., G_M(A_r) \xrightarrow{D} \mathit{Dirichlet}(\alpha H(A_1), ..., \alpha H(A_r))$ and in this sense the distribution of G_M converges to $DP(\alpha, H)$.

Proof: For $i = 1, ..., r$ let $H_i = H(A_i)$, and let $N_i = \sum_{j=1}^{M} \mathbb{I}_{\theta_j \in A_i}$ count the atoms in A_i. Since $\Pr(\theta_j \in A_i) = H_i$, we have

$$N_1, ..., N_r \sim \text{Multinomial}(M, H_1, ..., H_r). \tag{8.10}$$

For $i = 1, ..., r$, $G_M(A_i)$ is the sum of the N_i Dirichlet-distributed weights w_j for indices j such that $\theta_j \in A_i$. If we condition on $N_1, ..., N_r$ then we are summing a fixed number of these Dirichlet-distributed weights in each set, so by the agglomerative property of Dirichlet distributions, the conditional density at $G_M(A_1) = g_1, ..., G_M(A_r) = g_r$ wrt Lebesgue measure dg of Λ is

$$f_M(g|N_1, ..., N_r) = \text{Dirichlet}(g; \alpha N_1/M, ..., \alpha N_r/M).$$

This is random, as the counts $N_1, ..., N_r$ are random. We observe that

$$f_M(g) = E_{N_1,...,N_r}(f_M(g|N_1, ..., N_r))$$

and so this density wrt dg exists. However, $N_1/M, ..., N_r/M \xrightarrow{P} H_1, ..., H_r$ from Eqn. 8.10, so

$$f_M(g|N_1, ..., N_r) \xrightarrow{P} \text{Dirichlet}(g; \alpha H_1, ..., \alpha H_r)$$

at each $g \in \Lambda$ by the continuous mapping theorem. Since the random conditional density at g converges to a constant (not depending on $N_1, ..., N_r$), the two limits $\lim_{M\to\infty} f_M(g|N_1, ..., N_r)$ and $\lim_{M\to\infty} f_M(g)$ must be equal so

$$\lim_{M\to\infty} f_M(g) = \text{Dirichlet}(g; \alpha H_1, ..., \alpha H_r).$$

EOP