Bayes Methods - SC7 lecture notes, MT25

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.*

The following plan will be roughly correct, though the map to lectures may be $\pm 1$.

Problem sheet 0 - Some background you might like to look at in Week 1.
Problem sheet 1 - Lectures 1-4 on prior elicitation, de Finetti and utility theory, Sec 1–4.2
Problem sheet 2 - Lectures 5-8 on the Savage Axioms and MCMC, Sec 4.3–6
Problem sheet 3 - Lectures 9-12 on ABC, Model Averaging, Reversible Jump MCMC, Sec 7–9.3.1
Problem sheet 4 - Lectures 13-16 give a RJ example and the Dirichlet Process, Sec 9.4–10

*This version: 08/11/2025. Revisions (in blue).*

# Contents

## 6 Model selection: estimating the marginal likelihood 65

## 7 Likelihood-free methods: Approximate Bayesian Computation 71

## 8 Model averaging 79

## 9 Reversible-Jump MCMC 86

# 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 are familiar, so this is a whistle-stop tour.

The course aims to get you to a point where you can carry out subjective Bayesian inference in practice. There are three sides to this: statistical modelling, statistical inference and statistical computing. Statistical modelling is really about making a map between mathematical and statistical structures and the real world processes that generated the data: as the physicists say, "the elements of the model should be in one to one correspondence with elements of reality". The foundations of coherent Bayesian inference are in decision theory. When this theory was first developed it wasn't really feasible to do what the math was telling us to do as the quantities of interest were intractable integrals over high dimensional spaces. However, yesterday's intractable integral is today's five minute task and a consequence of the availability of increasingly powerful computers is that we now can and should implement coherent inference. We want to make efficient use of this resource and that's where scaleable algorithms for statistical computing enter.

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 on github. Some may find this an aid to learning, and it may be useful if you need to make your own applications, but the R and coding is not examined in the written exams in any way.

The topics below are chosen to represent certain important themes in modern Bayesian research, to some extent for their immediate potential for use in application, but also to illustrate the subject as a whole: fundamentals (Chapters 3 and 4), computational methods (Chapters 5, 6, 7 and 9) and different varieties of model (Chapters 8 and 10). 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 when choosing the prior. 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.

## 1.2 Bayesian inference

Statistical inference has a "pipeline". For example for a hypothesis test carried out in a frequentist setting, 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 9.

### 1.2.1 Prior, Observation model, posterior

Some unknown real-world quantity $\Theta$ takes values in a parameter space $\Omega$. Typically $\Omega$ is a $p$-dimensional subset of $R^p$ below. Let $S \subseteq \Omega$ be a subset of possible values for $\Theta$. Is $\Theta$ 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 $\Omega$ 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 $\Theta$, 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 $\sigma$-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 $\sigma$-algebra of subsets of $\Omega$, $d\pi(\theta)$ is a general probability measure on $\Omega$ and $d\theta$ is Lebesque volume measure in $\Omega$, 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 $\theta$ 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 $\theta$.

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

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

is the posterior probability the unknown true $\Theta$ 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 $\Theta, Y$ 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 $\Theta$ then realises observations $Y|\Theta$ and returns $Y = y_{obs}$. The conditional distribution of $\Theta|Y = y_{obs}$ is proportional to the joint model $p(y_{obs}|\theta)\pi(\theta)$. We can turn this round and use it to simulate the conditional.

*Exercise* 1.2. Suppose the data $Y$ and $\Theta$ 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 $\theta$-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 $\Theta'|Y' = y_{obs}$ 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 of 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.

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

---

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.

formalising knowledge about the world. Scientists often 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*[2]

1. Is the parameter $\theta$ generated by some process we can model? If so then the distribution over $\theta$ determined by the process *is* the prior. See Section 1.4 for an example.
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. See Section 1.3.2
4. How reliable is the information you are using to build a prior? If it is uncertain, you may wish to take as your prior a mixture distribution over priors. This leads to model averaging (Section 8) and non-parametric Bayes (Section 10).
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. The parameter dimension may be known, but vary from one data set to another (for example in the hierarchical model $\theta \sim \pi(\cdot)$, $\phi_i \sim \pi(\cdot|\theta)$, $y_i \sim p(\cdot|\phi_i)$, $i = 1, \ldots, n$). When we specify a family of priors, one for each possible dimension of $\theta$, we need the prior distributions to be *marginally consistent*. We explain this in Section 3.
7. Sometimes the dimension of the parameter vector is unknown! 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! See Sections 8, 9 and 10.
8. 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. See Section 1.3.2 for an example.
9. It isnt necessary 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.

The observation model is typically easier to model: we often have many iid observations $y_1, \ldots, y_n$ and we can use these to learn the density $p(y_i|\theta)$. However these observations depend on just one realisation of $\theta$ and we dont even get to observe it! This is why modeling the prior distribution of $\theta$ is so much harder than that for $y$, even when $\theta$ is a physically meaningful quantity.

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

---

[2]Here is a recent paper on PE by some knowledgable people. Might be useful to get a different perspective, but not part of the course. Petrus Mikkola et al. "Prior Knowledge Elicitation: The Past, Present, and Future", Bayesian Analysis, 19(4) 1129-1161 (2024).

**Prior standard deviation 3**     **Prior standard deviation 100**
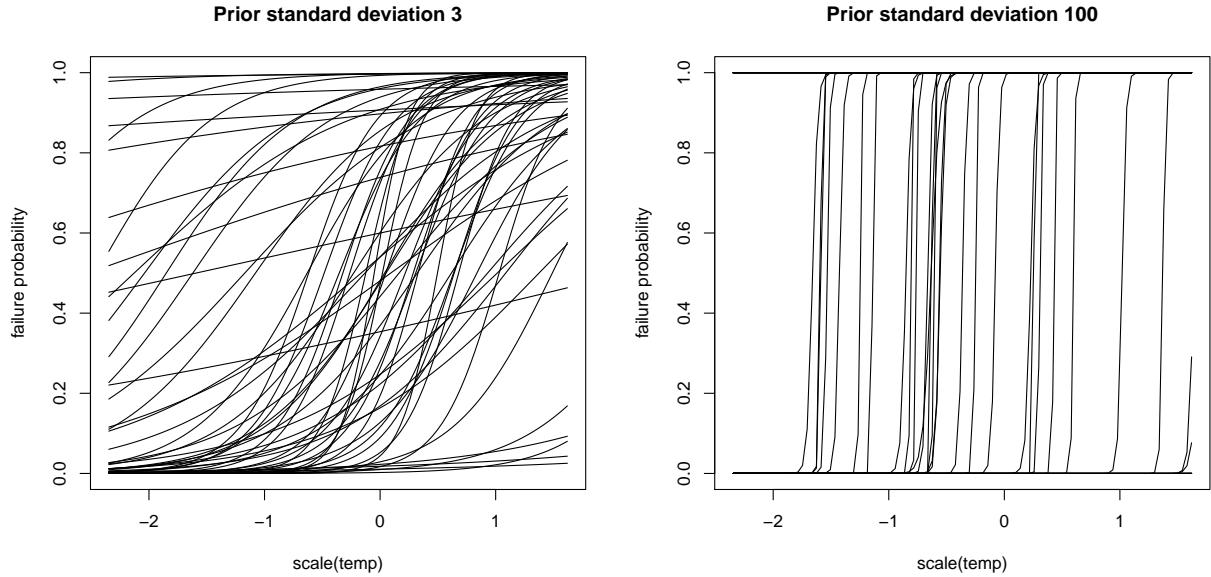
Figure 1: Samples of the function $\mu(\beta_1 + \beta_2 x)$, the probability for O-ring failure, drawn from the prior and plotted against $x$, the scaled temperature: (Left) prior standard deviation of $\beta_1, \beta_2$ equal 3; (Right) prior standard deviation 100. Only increasing $\mu(\eta)$ are plotted for clarity.

with an obligation to check it is representative of the information you have.

### 1.3.2 Prior for effect sizes in a model for O-ring data

Here is an example of how a prior that seems uninformative can be very informative. We return to it as a model selection example in Section 6.3 and model averaging in Example 8.6 in Section 8.

In the Challenger O-ring data[3] $(y_i, x_i), i = 1, 2, ..., n$, the response $y_i \in \{0, 1\}$ indicates O-ring failure and the covariate $x_i \in R$ is the centred and scaled temperature on the day the ring was tested (so $x_i$ is of order one). The observation model is $y_i \sim Bern(\mu(\beta_1 + \beta_2 x_i))$, $i = 1, ..., n$. The linear predictor is $\eta = \beta_1 + \beta_2 x$ and use the logistic link function $\mu(\eta) = (1 + \exp(-\eta))^{-1}$. Suppose we take a simple prior with $\beta_1, \beta_2 \sim N(0, v)$ independent for some fixed $v > 0$.

You will commonly see people taking $v$ very large, and this is often presented as a "non-informative" choice. Before seeing the data we have some prior knowledge of $\mu(\eta)$, the probability for O-ring failure. We know they can but don't 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 $\beta_1, \beta_2 \sim N(0, v)$ 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 1, 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]see S. Dalal, E. Fowlkes and B. Hoadley (1989) *Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure*, Journal of the American Statistical Association, 84: 945-957, and the R-code for Lecture 8

### 1.3.3 Decision Theory for Bayesian estimation

*This only gets a mention in Lecture 1 - we will review it in about Lecture 4.*

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 $\Theta$. This will be informed by the data and prior so we must have $\delta = \delta(y)$. In the language of decision theory, $\delta$ 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 $\Theta$-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

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

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

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

A Bayes estimator $\delta^{\pi}$ for $\theta$ minimises the Bayes risk

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

This is not straightforward as an estimator $\delta$ 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

$$\rho(\pi, \delta|y) = E_{\Theta|Y=y}(L(\Theta, \delta(y)))$$
$$= \int_{\Omega} L(\theta, \delta(y)) \pi(\theta|y) d\theta \tag{1.1}$$

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 $\delta$ when the truth is $\Theta$ 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. (now do exercise 2 in Problem sheet 0). Appendix D below shows that Bayes estimators are in some sense "optimal". The result is often summarised as "every admissible estimator is a Bayes estimator".

### 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-$\alpha$ Highest Posterior Density (HPD) credible set $C_\alpha$ 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 1.3 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 *this* credible set is $1 - \alpha$" which would be incorrect for a CI. See Appendix C for an exercise exploring the relation between HPD sets and Decision Theory.

*Exercise* 1.3. (not so easy) 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 $c_\alpha$ 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 $C_\alpha$ 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 $C_\alpha$ is an HPD set if $c_\alpha = F^{-1}(\alpha)$. Now $q^{\{\lfloor \alpha T \rfloor\}}$ is the empirical $\alpha$-quantile of $Q$ so by the hint it is a consistent estimator for $c_\alpha$. 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. ♣

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.

### 1.3.5 Model selection

Suppose we are considering a discrete set $\mathcal{M}$ of models indexed by integers $m = 0, 1, 2, ....$ 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 $\Theta, Y$, 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 $m \in \mathcal{M}$. Some thoughts about how to define a prior over a large but finite set of models are given in Cox, D. *Principles of Statistical Inference* (2006) in Section 5.15. 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.

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 0. We return to it in detail in Section 8.

### 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.

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 8. 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 Chapters 8 and 9, 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. 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 $\theta$ (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. Under regularity conditions, the Laplace approximation (accurate at large data sample size $n$) gives
$$\log(p(y|m)) = \ell(\hat{\theta}_{ML}; y) + \log(\pi(\hat{\theta}_{ML})) + \log|\Sigma|^{1/2} + \frac{p}{2}\log(2\pi/n) + O(1/n) \qquad (1.2)$$
where $\hat{\theta}_{ML}$ is the MLE and $\Sigma$ is the observed unit Fisher information.[4] This decreases with increasing $p$ at large $n$ so the marginal likelihood 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 $\theta$ *everywhere* in $\Omega_m$ (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 $\theta$) 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 $\Theta$ 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.[5]

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 hard to estimate (as a function of $\theta$, $p(y|\theta)$ is typically concentrated in $\Omega$, $\pi$ is diffuse) and this simple computational obstacle has been one of the principal obstacles to more widespread use of Bayes methods. We discuss some practical schemes in Section 6.

## 1.4   Lecture 2: 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.

---

[4]See Section 6.2 for further detail.
[5]This leads to the Lindley paradox.

```
#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 $\sigma_i$ (which is given as part of the data) and standard deviation $\sigma_c(\theta_i)^2$ in the calibration map $\mu$ 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 2 shows $p(y_6|\theta_6)$ on the $y$-axis as a function of $\theta_6$ 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:

$$\pi_u(\theta) = (U - L)^{-n} \prod_{i=1}^{n} \mathbb{I}(L \leq \theta_i \leq U).$$
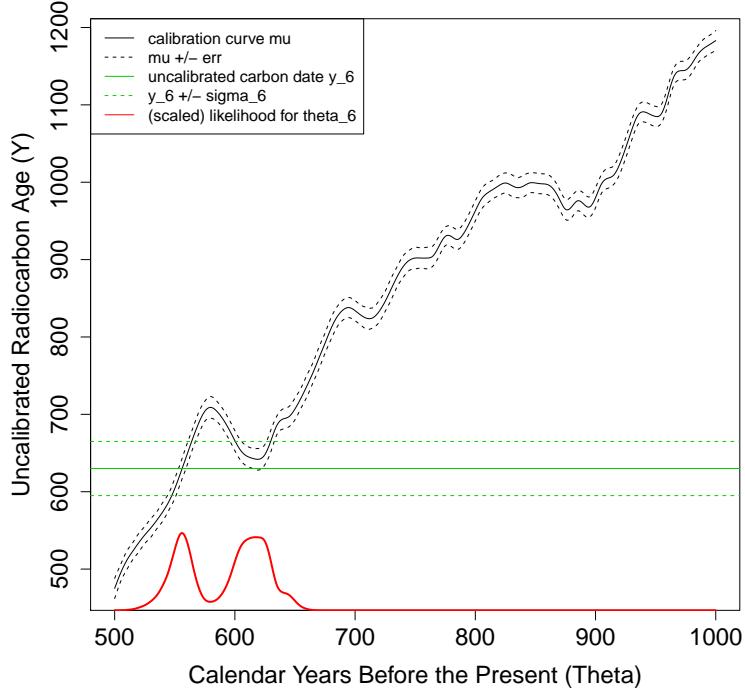
Figure 2: Calibration curve $\mu$ with likelihood $p(y_6|\theta_6)$ (in red). The solid green line is the observed value $y_6$ (with dashed lines at $y_6 \pm \sigma_6$). The solid black line gives the calibration curve $\mu(x)$ (with dashed lines at $\mu(x) \pm \sigma_c(x)$). The red curve, representing 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 $\theta^- = \min(\theta)$ and $\theta^+ = \max(\theta)$, the most recent and earliest sample dates, and estimate the span using

$$S_u = \theta^+ - \theta^-.$$

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 $\pi_u$ 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 $\theta^= - \theta^-$-values (see Figure 3 at right). We can calculate it.

*Exercise* 1.4. 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: See Appendix A. ♣

*Exercise* 1.5. Make a change of variables from $(\theta^-, \theta^+)$ to $(\theta^-, s_u)$ with $s_u = \theta^+ - \theta^-$ and Jacobian
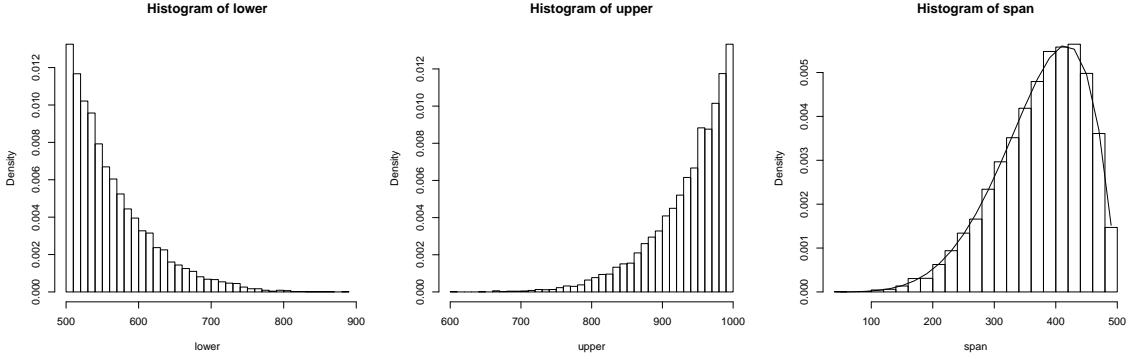
Figure 3: Marginal priors $\pi_u(\theta^-)$ (left), $\pi_u(\theta^+)$ (mid), $\pi_u(S_u)$ (right). We have no basis in fact for favoring large $S_u$ values over smaller values.

equal one and integrate over $\theta^-$ 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 $\theta^-$ 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 $\theta^-$ and gives $\pi_{S_u}(s_u)$ above. ♣

We can check the distribution of $S_u$ (and my calculation) using histograms (see Figure 3) from prior simulations. Let's interpret these graphs: in the uniform prior $\pi_u$, 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 $\pi_u$ 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 $\theta$ 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 $\theta$

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 $\psi_1$ 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 $\lambda dt$ - that is, the dates $\theta$ are a realisation of a Poisson process with constant rate $\lambda$ 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 $\pi_u$. 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 $\psi$-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 (See Appendix B for some background on Poisson processes)*

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* 1.6. 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 uniform on $[0, U - L]$.

Ans: make a change of variables from $(\psi_1, \psi_2)$ to $(\psi_1, s)$ in the density and integrate $\psi_1$ 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 $\psi_1$, $\psi_2$ and $S_s = \psi_2 - \psi_1$ in Figure 4. These marginal priors on $\psi_1$, $\psi_2$ and $S_s$ in Figure 4 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,

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

and

$$\pi_s(\theta, \psi|y) \propto p(y|\theta)\pi_s(\theta, \psi).$$
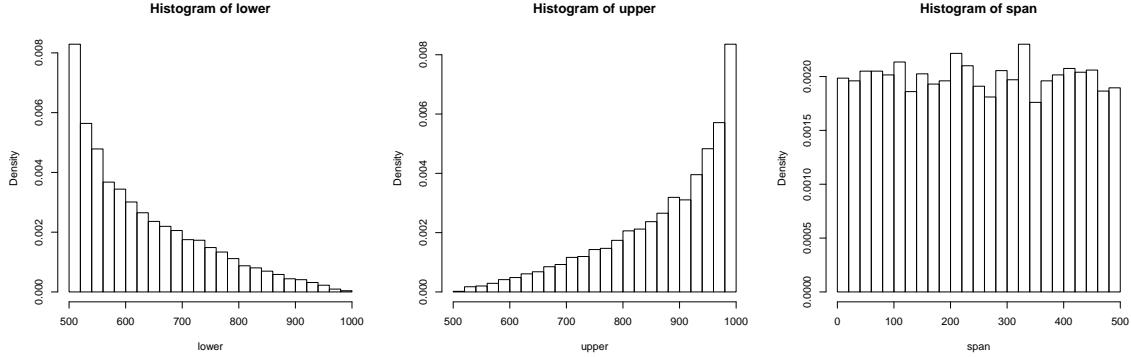
Figure 4: Marginal priors $\pi_s(\psi_1)$ (left), $\pi_s(\psi_2)$ (mid), $\pi_{S_s}(s)$ (right).

The two posterior distributions have the same likelihood, but different priors $\pi_u$ and $\pi_s$, and parameters spaces $\Omega_u$ and $\Omega_s$. 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 $\theta_{(1)}^{(t)}, \theta_{(n)}^{(t)}$ and span $S_u^{(t)} = \theta_{(n)}^{(t)} - \theta_{(1)}^{(t)}$ 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 $S_s^{(t)} = \psi_2^{(t)} - \psi_1^{(t)}$ 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/$\pi_u$): HPD set [70,160]        Model 2 (shrink/$\pi_s$): 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 $\pi_u$ 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.[6] 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 $\pi_s$ is clearly favoured.

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.

---

[6]We will see how this is done in Chapter 6.

**Posterior span, shringage prior**     **Posterior span, uniform prior**
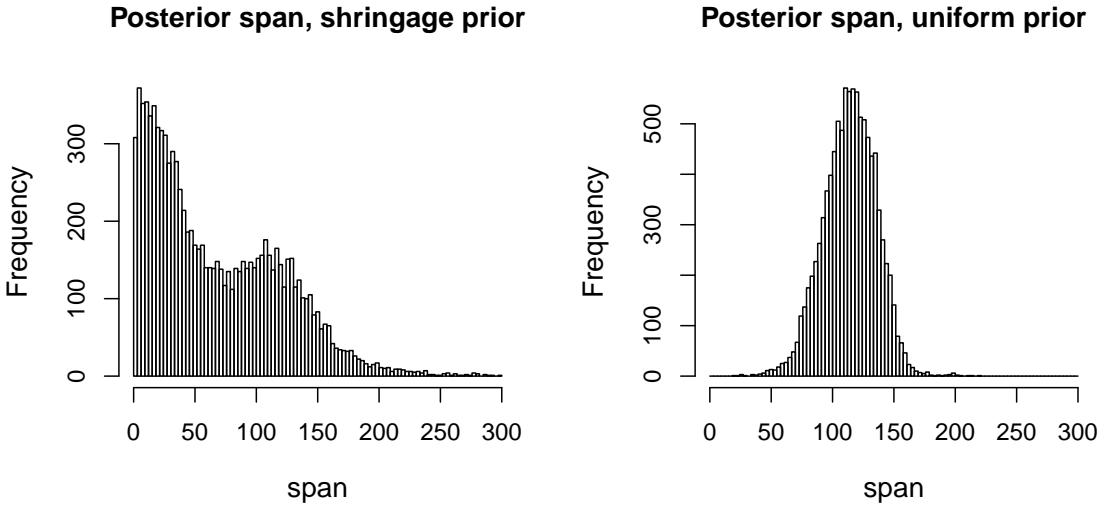
Figure 5: (Left) Posterior for the occupation span under model 2, the uniform span prior $\pi_s$. (right) posterior for the occupation span under model 1 the uniform age prior $\pi_u$.

### 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 $\pi_s$ better represented actual prior knowledge than $\pi_u$ and was strongly favored by the data ($B_{s,u} \simeq 200$).

The process-model based prior $\pi_s$ 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.

## 1.5 Appendices

### 1.5.1 Appendix A: Solution to Exercise 1.4

Here is a solution to Exercise 1.4. Let $\xi_{1,2}$ be the event that $\theta_1 < \theta_3, \ldots, \theta_n < \theta_2$ so that $\theta^+ = \theta_2$ and $\theta^- = \theta_1$ everywhere in $\xi_{1,2}$. Conditioning on $\xi_{1,2}$, for $\theta$ satisfying this condition,

$$\pi_u(\theta|\xi_{1,2}) = \pi_u(\theta)/\pi_u(\xi_{1,2})$$
$$= (U - L)^{-n} n(n-1),$$

so

$$\pi_u(\theta^+, \theta^-|\xi_{1,2}) = \int_{\theta^-}^{\theta^+} \cdots \int_{\theta^-}^{\theta^+} \pi_u(\theta|\xi_{1,2}) d\theta_{3:n}$$
$$= (U - L)^{-n} n(n-1)(\theta^+ - \theta^-)^{n-2}$$
$$= \pi_u(\theta^+, \theta^-)$$

as the conditional doesn't depend on $(i,j)$ in $\xi_{i,j}$ so $\pi_u(\theta^+, \theta^-) = \sum_{(i,j)} \pi_u(\theta^+, \theta^-|\xi_{1,2})\pi_u(\xi_{1,2}) = \pi_u(\theta^+, \theta^-|\xi_{1,2})$. Make a COV to $s,t$ with $s = \theta^+ - \theta^-$ and $t = \theta^+ + \theta^-$. The Jacobian factor

$|\partial(s,t)/\partial(\theta^+, \theta^-)|^{-1} = 1/2$ so

$$\pi_u(s,t) = (U-L)^{-n} n(n-1) s^{n-2}/2,$$

so integrating over $L + (L+s) < t < U + (U-s)$ to get the marginal in $s$,

$$\pi_u(s) = \int_{2L+s}^{2U-s} \pi_u(s,t) dt$$
$$= (U-L)^{-n} n(n-1) s^{n-2} (U-L-s),$$

as claimed. Notice the undesirable power dependence on $s$.

### 1.5.2 Appendix B: Some notes on the Poisson process

We make use of Poisson processes at a few points in this course. You may have seen some of this in earlier courses. There is a shift in emphasis compared to the usual setup - we are interested in the joint distribution of the event locations $X = (X_1, \ldots, X_N)$ in an interval and not so much in the continuous time stochastic process that determines this distribution (only insofar as it motivates the joint distribution). One advantage of this view is that it extends straightforwardly to dimension greater than one (and we call it a Poisson Point Process in $R^p$ with intensity $\lambda$ rather than a Poisson process in $R$ with rate $\lambda$).

The headline results are that the number of events $N_A = |X \cap A|$ in any bounded open set $A \subset R^p$ is Poisson $\lambda |A|$, where $|A| = \int_A dz$ is the volume measure of the set, and the joint distribution of the points $X_i \in A$, $i = 1, \ldots, n$, given the number $N_A = n$ is iid uniform in $A$. This is intuitively obvious: the probability there is a point in a small element of volume $\delta \subset A$ is $\lambda |\delta| + o(|\delta|)$ independently in each disjoint set $\delta$ (same as the process in $R$), so the number of occupied cells in $A$ is approximately $N_A \sim \text{Binomial}(N_A; |A|/|\delta|, \lambda|\delta|)$ which converges to $\text{Poisson}(N_A; \lambda|A|)$ as $|\delta| \to 0$. Since the cells $\delta$ are occupied independently it is fairly clear that when we condition on the number of occupied cells, each occupied cell is uniformly distributed over $A$.

Here is how it works in $R$, the setup relevant for the radiocarbon dating example in Section 1.4. Let a rate $\lambda > 0$ be given and let $X_i$, $i = 1, 2, \ldots$ (with $X_0 = 0$) be the arrival times of a Poisson Process, with inter-arrival times $T_i = X_i - X_{i-1}$ (iid). It is a Poisson process with rate $\lambda$ so $T_i \sim \text{Exp}(\lambda)$, $i = 1, 2, \ldots$ are iid. Let $X = (X_1, X_2, \ldots) \cap (0, U)$ be events in $(0, U)$. The dimension of $X = (X_1, \ldots, X_N)$ is random (number of events before $U$). Let $x = (x_1, \ldots, x_n)$ be a realisation with $N = n$. Let $\Omega_n = \{x \in R^n : 0 < x_1 < x_2 < \cdots < x_n < U\}$ (so the points are distinguishable) with $\Omega_0 = \emptyset$. The state space for $X$ is

$$\Omega = \bigcup_{n=0}^{\infty} \Omega_n.$$

The distribution of $X$ has density $\pi_X(x)$ at $x \in \Omega$. Let us calculate this density. Write down the density for $T_1, T_2, \ldots$ and make a COV to $x$ (with Jacobian equal one). Suppose $T = (T_1, \ldots, T_N)$ and $t = (t_1, \ldots, t_n)$ is a realisation with $\sum_{i=1}^n t_i < U$. The density for $T$ is

$$\pi_T(t) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n t_i\right) \times \exp\left(-\lambda(U - \sum_{i=1}^n t_i)\right)$$

as there is no event in the interval $(x_n, U)$ of length $U - x_n$, so

$$\pi_X(x) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i - x_{i-1}\right) \times \exp(-\lambda(U - x_n))$$
$$= \lambda^n \exp(-\lambda U), \ x \in \Omega_n.$$

We gave the density for some $x \in \Omega_n$ but it is normalised over $\Omega$ (see below).

Now $\Pr(N = n) = \Pr(X \in \Omega_n)$ (and call this $\pi_N(n)$, $n = 0, 1, 2, \ldots$ say) so

$$\pi_N(n) = \int_{\Omega_n} \pi_X(x) dx$$
$$= \int_0^U dx_n \int_0^{x_n} d x_{n-1} \cdots \int_0^{x_2} \lambda^n \exp(-\lambda U) dx_1$$
$$= \lambda^n \exp(-\lambda U) \times U^n / n!$$
$$= \text{Poisson}(n; \lambda U).$$

So marginally $N \sim \text{Poisson}(\lambda U)$.

Conditioning on $N = n$ the distribution of $X|N = n$ is uniform and doesnt depend on $\lambda$, as we now verify. For $x \in \Omega_n$,

$$\pi_X(x|X \in \Omega_n) = \pi_X(x) / \Pr(X \in \Omega_n)$$
$$= \lambda^n \exp(-\lambda U) / \text{Poisson}(n; \lambda U)$$
$$= n! / U^n.$$

This doesn't depend on $x$ or $\lambda$ so $X|N = n$ is uniform on $\Omega_n$.

Notice that $\pi_X(x)$ integrates to one over $\Omega$. This is an example where the measure (ie the rule assigning probability mass to sets) is a mixture of counting measure and volume measure.

$$\int_\Omega \pi_X(x) dx = \sum_{n=0}^\infty \int_{\Omega_n} \pi_X(x_{1:n}) dx_{1:n}$$
$$= \sum_{n=0}^\infty \text{Poisson}(n; \lambda U)$$
$$= 1.$$

### 1.5.3 Appendix C: HPD sets from Decision Theory

*Exercise* 1.7. (If you haven't seen Decision Theory then read Section 1.3.3 and do problem 2 on Problem Sheet 0) The HPD set is a Bayes estimator. Since it is a subset of $\Omega$, the action $\delta$ must be a subset of $\Omega$, 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 $\delta$. 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 $\Delta$ by any set with (equal) least volume. This is achieved by an HPD set $C_\alpha$. 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. ♣

### 1.5.4 Appendix D: Admissiblity

The material in this Appendix (C) is not examinable. I include it as one of the arguments used to motivate Bayesian Inference.

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

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

and for at least one $\theta_0$,

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

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

Estimators that seem reasonable (recall James-Stein beats MLE 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 1.8.** *(Proposition 2.4.22 in the textbook CR-TBC): If prior $\pi$ is strictly positive on $\Omega$ with finite Bayes risk, and the risk, $\mathcal{R}(\theta, \delta)$, is a continuous function of $\theta$, then Bayes estimator $\delta^\pi$ is admissible.*

*Proof.* Suppose the opposite. For some $\delta'$, $\mathcal{R}(\theta, \delta^\pi) \geq \mathcal{R}(\theta, \delta')$ for each $\theta$, and there exists $\theta'$ and an open neighborhood $C'$ of $\theta'$ 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. $\qquad\square$

# 2    Marginal Consistency

A family of probability distributions is marginally consistent if every marginal of every distribution in the family is also in the family. We start with $N$ random variables $X_1, \ldots, X_N$ and take a subset $X_{a_1}, \ldots, X_{a_n}$ of size $n$. The distribution of the subset should be given as their marginal in the joint distribution of the original $N$ variables.

How could this possibly not hold? It follows from the axioms of probability! Marginal consistency is an issue for prior elicitation. In some applications, when we change the set of variables we are modeling (the number or the selection), we change their distribution. This should be straightforward, not deserving a chapter in these notes. However, people often forget that this issue even exists, and sometimes when working with complex distributions write down models which violate marginal consistency unintentionally. This tends to happen in relatively complex models for random spatial processes and in models for random graphs. It is related to the Kolmogorov extension theorem, which we wont discuss, but I invite you to check out for further reading.

## 2.1   Lecture 3: Illustration

Let $X_i \in \{0, 1\}$ be the indicator that student $i = 1, \ldots, n$ was born on the same day of the year as Bruno de Finetti (13th June) and let $X = (X_1, \ldots, X_n)$. The distribution of $X_1$ (or the joint distribution of any subset of these indicators) doesn't depend on $n$. This will be a marginally consistent example which we use to illustrate notation.

Let $p_{1:n}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$ give the joint distribution of $X$. We write $X \sim p_{1:n}$. There are some pretty obvious concrete models we might take for $p_{1:n}$. Now suppose a new student arrives so there are $n+1$ students. The first $n$ students haven't changed. Let $Y_i \in \{0, 1\}$ be the indicator that student $i = 1, \ldots, n+1$ was born on 13th June with $Y = (Y_1, \ldots, Y_{n+1})$. Let $p_{1:n+1}(x_1, \ldots, x_n, x_{n+1}) = P(Y_1 = x_1, \ldots, Y_n = x_n, Y_{n+1} = x_{n+1})$ give the joint distribution of $Y$. There is no reason to change our model for birth dates of students $1, \ldots, n$ just because we have another student, so we require the joint distributions $(X_1, \ldots, X_n) \sim (Y_1, \ldots, Y_n)$ to match, the same as saying that $p_{1:n}$ is the marginal of $p_{1:n+1}$ and $X_1, \ldots, X_n$ are *the same random variables* as $Y_1, \ldots, Y_n$. By the countable additivity Axiom of Probability (AoP),

$$p_1(x_1, \ldots, x_n) = p_{1:n+1}(x_1, \ldots, x_n, 0) + p_{1:n+1}(x_1, \ldots, x_n, 1).$$

This is an example of *marginal consistency*. When would we ever depart from this setup?

Sometimes the random variables change depending on the particular subset of objects they model. Let $\tilde{X}_i \in \{0, 1\}$ be the indicator that student $i = 1, \ldots, n$ asks a question in the lecture on de Finetti's Theorem, with $\tilde{p}_{1:n}(x_1, \ldots, x_n) = P(\tilde{X}_1 = x_1, \ldots, \tilde{X}_n = x_n)$ so $\tilde{X} \sim \tilde{p}_{1:n}$ for $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)$. Suppose students are less likely to ask questions as the class size grows. This seems plausible from a modeling perspective. When we add a student, the marginal distributions of the indicator variables change, so this will give an example which is not marginally consistent.

In slightly more detail, let $\tilde{Y}_i \in \{0, 1\}$ be the indicator that student $i = 1, \ldots, n+1$ asks a question when there are $n + 1$ students, with $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_{n+1})$ and $\tilde{Y} \sim \tilde{p}_{1:n+1}$. Since the behaviour of student $i$ changes when $n$ changes, we do not impose $(\tilde{X}_1, \ldots, \tilde{X}_n) \sim (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ in our model, so *when $n$ changes the random variables themselves change*. In the family of models for different class sizes

$$\tilde{p}_{1:n}(x_1, \ldots, x_n) \neq \tilde{p}_{1:n+1}(x_1, \ldots, x_n, 0) + \tilde{p}_{1:n+1}(x_1, \ldots, x_n, 1).$$

This time the distributions are not marginally consistent.

## 2.2 Definition

To be concrete, and because this is the relevant case for the next chapter, we write this down for binary variables. It is clear enough how this works in general (for eg, for continuous variables, sums become integrals) and I leave it to you to make the generalisation.

Let $\mathcal{S}_{[n]}$ be the set of all nonempty subsets of $[n]$ and suppose $s = (s_1, ..., s_m)$ for some $s \in \mathcal{S}_{[n]}$. In this and the next chapter subscripts indicate the variables present so

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

and,

$$p_s(x_{s_1}, ..., x_{s_m}) = P(X_{s_1}^s = x_{s_1}, ..., X_{s_n}^s = x_{s_m}).$$

Notice the superscript $[n]$ on $X^{[n]} = (X_1^{[n]}, ..., X_n^{[n]})$ and $s$ on $X^s = (X_{s_1}^s, ..., X_{s_m}^s)$. The point here is that $X^{[n]} \sim p_{1:n}$ and $X^s \sim p_s$ but the random variables $(X_{s_1}^s, ..., X_{s_m}^s)$ need not have the same distribution as the corresponding random variables $X_s^{[n]} = (X_{s_1}^{[n]}, ..., X_{s_m}^{[n]})$.

**Definition 2.1.** (binary case) Suppose $x_i \in \mathcal{X}$, $i = 1, ..., n$ with $\mathcal{X} = \{0, 1\}$. Let

$$\mathcal{F}_{[n]} = \bigcup_{s \in \mathcal{S}_{[n]}} \{p_s(\cdot)\}$$

be a given family of probability mass functions (PMFs) $p_s : \mathcal{X}^m \to [0, 1]$. Let $r = [n] \setminus s$ be the complement of $s$ and

$$q_s(x_{s_1}, ..., x_{s_m}) = \sum_{x_{r_1} \in \mathcal{X}} \cdots \sum_{x_{r_{n-m}} \in \mathcal{X}} p_{1:n}(x_1, ..., x_n) \tag{2.1}$$

be the marginal PMF for the random variables $X_s^{([n])}$ if $X^{[n]} \sim p_{1:n}$. The family $\mathcal{F}_{[n]}$ is marginally consistent if and only if $p_s(x_s) = q_s(x_s)$ for every $x_s \in \mathcal{X}^m$ and all $s \in \mathcal{S}_{[n]}$. $\diamond$

*Remark* 2.2. Marginally consistent families of distributions are sometimes called "projective". ✠

*Example* 2.3. If $\mathcal{F}_{[n]}$ is marginally consistent then in particular

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

(choose $s = (1, ..., n-1)$ in Eqn. 2.1 so $r = (n)$ and just one sum on the RHS of Eqn. 2.1). ♠

If $p_s(x_s) = q_s(x_s)$ for all $x_s$ and every $s$ then $X^{(s)} \sim X_s^{([n])}$ and we can just take $X = X^{([n])}$ and drop the superscript.

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

$$p_s(x_1, ..., x_m) = P(X_{s_1} = x_1, ..., X_{s_m} = x_m)$$

for each $s \in \mathcal{S}_{[n]}$ then the probability mass functions $p_s$ are marginally consistent by construction.

However, if we *start* with a collection of arbitrary probability mass functions $p_s$, $s \in \mathcal{S}_{[n]}$ there may not exist a set of random variables $X_{[n]}$ with those marginals. If the probability mass functions are not marginally consistent there cant be such a set: if there were then $p_s$ would have to be consistent by the AoP.

## 2.3 Examples

*Example* 2.4. Here is an example (an Ising model on a complete graph) where the probability mass functions are defined for every $n \geq 1$ but no sequence of random variable 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

$$p_{1:n}(x) = c_n 2^{-k(x)(n-k(x))}, \tag{2.2}$$

with $c_n$ a normalising constant with $c_3 = 2/7$ and $c_4 = 8/27$. Consider the sequence of probability mass functions $p_{1:n}$, $n \geq 1$. Does a sequence $X_1, X_2, X_3, ...$ with PMF's $P(X_1 = x_1, ..., X_n = x_n) = p_{1:n}(x_1, ..., x_n)$ exist? If it did then these PMF's would be marginally consistent.

This doesn't hold. For example, for MC 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 both $(X_1, ..., X_4) \sim p_{1:4}$ and $(X_1, ..., X_3) \sim p_{1:3}$. ♠

One way to prove that a family of distributions is marginally consistent is to give a generative model $\theta \sim \pi(\cdot)$ and $X_i \sim p(\cdot|\theta)$ iid for $i = 1, 2, 3, \ldots$ that satisfies $X_1, \ldots, X_n \sim p_{1:n}$ for each $n$. The joint is

$$p_{1:n}(x) = \int p(x|\theta)\pi(\theta)d\theta$$

with $p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$ and these (ie, the $p_{1:n}$) are marginally consistent (by the AOP, or by doing the sums over $x_i = 0, 1$, $i \in r$ under the integral). The trick is to spot the generative model.

*Exercise* 2.5. Let $x = (x_1, ..., x_n)$ with $x \in \{0, 1\}^n$. Show that the distributions

$$p_{1:n}(x) = \frac{k(x)!(n - k(x))!}{(n+1)!}, \quad n \geq 1 \tag{2.3}$$

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

$$\begin{aligned}
p_{1:n}(x) &= \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \int_0^1 \theta^{k(x)}(1-\theta)^{n-k(x)}d\theta \\
&= \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}d\theta. 
\end{aligned} \tag{2.4}$$

This is the same as the distribution of $X_1, \ldots, X_n$ in the generative model $\theta \sim U(0, 1)$, $X_i \sim$ Bernoulli$(\theta)$, $i = 1, 2, 3, ....$ Since this generates an infinite sequence of random variables with the stated marginals, it follows by the AoP that $p_{1:n}$ are marginally consistent. This is a Polya urn (Section 3.2.1) with $b = 1$ black balls, $w = 1$ white and $A = 1$ added at each step. ♣

# 3   Exchangeability

In this chapter and the next we are concerned with the existence of priors (this chapter) and the definition of probability in terms of "preference" or "subjective" probability (next chapter). The Kolmogorov axioms of probability define probability as a mathematical system. However, there is a question of how the objects in this system relate to real-world events. In Frequentist inference this connection is made by defining the probability of an event as the proportion of successes in an infinite sequence of trials. If the distribution of the data is defined in this way then it satisfies the axioms. If we are OK to assume that we could in principle make arbitrarily many observations $y_1, y_2, \ldots$ then some true data-distribution $p^*(y)$ exists and it is the job of statistical modelling to find a distribution $p(y|\theta)$ which approximates it reasonably well. However, in Bayesian inference we need a prior distribution for the observation model parameters. There may only ever be one realisation of $\theta$ so it's not clear how its probability distribution is defined, even in principle.

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 a generative model $p(y|\theta)\pi(\theta)$ for the data and parameter exists.

In preparing this section I found the following references 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.*

## 3.1   Exchangeability and Infinite Exchangeable Sequences

### 3.1.1   Exchangeability in finite sequences

**Definition 3.1.** Consider a finite sequence $X_i \in \mathcal{X}, i = 1, \ldots, n$ of $n \geq 1$ random variables. The random variables are *exchangeable* if their joint distribution is unchanged by permutation of the indices, so
$$(X_1, X_2, \ldots, X_n) \sim (X_{\sigma_1}, X_{\sigma_2}, \ldots, X_{\sigma_n})$$
for every permutation $\sigma \in \mathcal{P}_{[n]}$, $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ of the numbers $[n] = \{1, 2, \ldots, n\}$.   $\Diamond$

Any realisation of $(X_1, X_2, \ldots, X_n)$ is equally likely to be seen in any order. For example, suppose $X_1, \ldots, X_n$ are discrete with PMF $p_{1:n}(x_1, \ldots, x_n)$ on $\mathcal{X}^n$ so
$$P(X_1 = x_1, \ldots, X_n = x_n) = p_{1:n}(x_1, \ldots, x_n).$$
The subscript $1:n = (1, 2, \ldots, n)$ on $p_{1:n}$ emphasises the variables present and the order in which they are taken so
$$P(X_{\sigma_1} = x_1, \ldots, X_{\sigma_n} = x_n) = p_\sigma(x_1, \ldots, x_n)$$
and
$$P(X_1 = x_{\sigma_1}, \ldots, X_n = x_{\sigma_n}) = p_{1:n}(x_{\sigma_1}, \ldots, x_{\sigma_n}).$$
The distributions are marginally consistent so we write $X$ and dont need a superscript $X^{[n]}$.

Random variables $X_1, \ldots, X_n$ are then *exchangeable if and only if* for every $(x_1, \ldots, x_n) \in \mathcal{X}^n$,
$$p_{1:n}(x_1, \ldots, x_n) = p_{1:n}(x_{\sigma_1}, \ldots, x_{\sigma_n}), \quad \text{for each } \sigma \in \mathcal{P}_{[n]}, \tag{3.1}$$

as (3.1) is what $X_{1:n} \sim X_\sigma$ means for discrete variables in Definition 3.1. The same condition works for continuous $X_1, ..., X_n$ with joint density $p_{1:n}(x_1, ..., x_n)$ on $\mathcal{X}^n$.

*Exercise* 3.2. For $i = 1, ..., n$ let $X_i \in \{green = 0, red = 1\}$ be the colors of $n$ balls sampled without replacement from an urn containing $N \geq n$ balls, exactly $K$ of which are red. Let $x = (x_1, ..., x_n)$ for $x \in \{0, 1\}^n$ and $k(x) = \sum_i x_i$. Show that $(X_1, ..., X_n)$ are exchangeable but not independent.

ANS: Calculate $p_{1:n}(x)$ and verify Equation 3.1. Let $x_{<i} = (x_1, ..., x_{i-1})$. Since the balls are drawn consecutively we use the identity

$$p_{1:n}(x) = P(X_1 = x_1) P(X_2 = x_2 | X_1 = x_1) ... P(X_n = x_n | X_{<n} = x_{<n}).$$

Suppose $k(x) = k$ and 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 $P(X_{i_a} = x_{i_a} | X_{<i_a} = x_{<i_a}) = (K - a + 1)/(N - i_a + 1)$ in the probability for the sequence. Gathering up the factors,

$$p_{1:n}(x_1, ..., x_n) = \prod_{a=1}^{k} \frac{\textcolor{red}{K - (a-1)}}{\textcolor{blue}{N - (i_a - 1)}} \times \prod_{b=1}^{n-k} \frac{\textcolor{green}{N - K - (b-1)}}{\textcolor{blue}{N - (j_b - 1)}}$$
$$\frac{\textcolor{red}{K!}}{\textcolor{red}{(K-k)!}} \times \frac{\textcolor{green}{(N-K)!}}{\textcolor{green}{(N - K - (n-k))!}} \times \frac{\textcolor{blue}{(N-n)!}}{\textcolor{blue}{N!}},$$

where color indicates the factors in the first line that give the corresponding factor in the second line. Since this only depends on $x$ through $k(x)$, and $k(x)$ 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. However, the variables are not independent since for example,

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

but

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

*Exercise* 3.3. 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. $\qquad \clubsuit$

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

### 3.1.2 Infinite Exchangeable sequences

**Definition 3.4.** 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 $n \geq 1$. $\qquad \diamondsuit$

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* 3.5. 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 indices on the left we just shuffle the order of factors in the product. $\qquad \spadesuit$

*Exercise* 3.6. Let $p_{1:n}$ be given by Eqn. 2.3. Show that there is an IES with these marginals. ANS: we saw that if $\theta \sim U(0,1)$ and $X_i \sim$ Bernoulli$(\theta)$, $i = 1, 2, 3, ...$ then $(X_1, \ldots, X_n) \sim p_{1:n}$. We see from Eqn. 2.4 that the distributions are exchangeable for each $n \geq 1$ so we have an IES. ♣

### 3.1.3 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 $\Sigma_{i,i}^{(n)} = 1$, $i = 1, ..., n$ and for $0 \leq \rho < 1$ and $i, j = 1, ..., n$, $i \neq j$, equal positive covariances $\Sigma_{i,j}^{(n)} = \rho$. If $x = (x_1, ..., x_n)$, write $p_{1:n}(x_1, ..., x_n) = p_{1:n}(x)$ and suppose

$$p_{1:n}(x) = N(x; 0_n, \Sigma^{(n)}) \tag{3.2}$$

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 3.2 for every $n$.

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

$$X_i = \theta + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, 1 - \rho), \quad \text{iid for } i = 1, 2, 3.... \tag{3.3}$$

These $X$'s are distributed as in Equation 3.2. They are jointly normal as they are linear combinations of normal random variables. 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 $n$. It's clear that $p_{1:n}(x) = p_{1:n}(x_\sigma)$ for all $x \in \mathbb{R}^n$, $\sigma \in \mathcal{P}_n$ as the $X$'s are independent given $\theta$, so any realisation is equally likely to appear in any order. Verify this from the joint density. Integrating over $\theta$,

$$p_{1:n}(x_1, ..., x_n) = \int_{-\infty}^{\infty} p_{1:n}(x_1, ..., x_n | \theta) \pi(\theta) \, d\theta$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{n} N(x_i; \theta, 1 - \rho) N(\theta; 0, \rho) \, d\theta \tag{3.4}$$

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

*Exercise* 3.7. The covariance matrix $\Sigma^{(n)}$ is positive definite for every $n \geq 1$. Show that the parameter $\rho$ 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 $\lambda_i = 1 - \rho$, $i = 2, ..., n$ and

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

Hence show $-1/(n-1) < \rho < 1$ 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 $n \geq 1$. ♣

In this section we started with a probability distribution $p_{1:n}$ specified for every $n \geq 1$, and showed it determined an IES. We didnt just check invariance of $p_{1:n}$ under permutation of its arguments as that only tells us that for each $n \geq 1$ the random variables $(X_1^{([n])}, \ldots, X_n^{([n])}) \sim p_{1:n}$ are exchangeable - we also need marginal consistency. We need to show that there actually exists a set of random variables that works for all $n$ and a clear way to do that is to give a generative model for the variables as in Equation 3.3.

*Example* 3.8. Let $p_{1:n}$, $n \geq 1$ be given by Eqn. 2.2. Do these distributions define an IES? It is clear from the form of $p_{1:n}$ that the random variables $(X_1^{([n])}, \ldots, X_n^{([n])}) \sim p_{1:n}$ are exchangeable for every $n \geq 1$ ($p_{1:n}$ only depends on $\sum_i x_i$). However, we showed in Example 2.4 that there is no infinite sequence of random variables $(X_1, \ldots, X_n)$ distributed like $p_{1:n}$ for every $n \geq 1$, so there is no IES. This is easy - an IES of random variables is marginally consistent, because the joint distributions of the variables satisfy the axioms of probability. So if we start with a family of distributions which are not marginally consistent, it wont admit an IES. ♠

## 3.2  de Finetti's Theorem

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

*Example* 3.9. We saw in the hierarchical normal in Section 3.1.3 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 we could write the multivariate normal as an integral using Eqn. 3.4. Our infinite exchangeable sequence was actually 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 3.10.** *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 distribution $F(\theta)$ on $[0,1]$ such that*

$$p_{1:n}(x_1, ..., x_n) = \int_0^1 \left[ \prod_{i=1}^n p(x_i|\theta) \right] dF(\theta) \tag{3.5}$$

*with*

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

*It further holds that the conditioned distribution is Bernoulli,*

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

*with $p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$.*

We can interpret the theorem as telling us that a generative model for $\theta$ and the $X$'s exists. 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 that a CDF $F$ and observation model $p(\cdot|\theta)$ 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 $\theta$.

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 the space $\Omega$ of the random variable $\Theta$.

### 3.2.1   The Polya urn

This is a process generating an infinite exchangeable sequence and illustrates de Finetti in action. Consider an urn with $b$ black and $w$ white balls. We sample a ball and return it to the urn, adding $A$ balls of same color. For $i = 1, 2, 3, ...$ let $X_i = 1$ if the $i$th 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* 3.11. Show for the Polya urn with $k(x) = \sum_{i=1}^n x_i$ that

$$p_{1:n}(x_1, ..., x_n) = \frac{\prod_{i=0}^{k(x)-1}(b+iA) \prod_{j=0}^{n-k(x)-1}(w+jA)}{\prod_{i=0}^{n-1} b + w + iA} \tag{3.6}$$

$$= \int_0^1 \prod_i \theta^{x_i}(1-\theta)^{1-x_i} \mathsf{Beta}(\theta; b/A, w/A)\, d\theta. \tag{3.7}$$

and use this result to show that $X_1, X_2, X_3, ...$ is an infinite exchangeable sequence.
Hint: to get Equation 3.6 use the same approach as in Example 3.2. Next, to show equality, work from both ends: write Equation 3.6 in terms of $\Gamma$-functions using $x\Gamma(x) = \Gamma(x+1)$; do the integral in Equation 3.7 as you know how to normalise a Beta-density; observe that the two expressions are equal. We get an IES because we have the de Finetti representation in Equation 3.7.   ♣

### 3.2.2   Proof of de Finetti's Theorem

Equation 3.5 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 3.5 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 3.5 as $N \to \infty$.

**Proposition 3.12.** *Let*

$$S_n = X_1 + X_2 ... + X_n, \quad 1 \le n \le N$$

*and suppose $S_n = r$ and $S_N = s$. If $x_1 = 1, x_2 = 1, ..., x_r = 1, x_{r+1} = 0, x_{r+2} = 0, ..., x_n = 0$ then*

$$\Pr(S_n = r) = \binom{n}{r} p_{1:n}(x_1, ..., x_n) \tag{3.8}$$

*and*

$$\Pr(S_n = r | S_N = s) = \binom{s}{r} \binom{N-s}{n-r} / \binom{N}{n}. \tag{3.9}$$

*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 3.8. For Equation 3.9, 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. $\qquad \square$

When $S_n = r$ the sequence $X_1, ..., X_N$ contains at least $n - r$ 0's (in $X_1, ..., X_n$) so

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

$$= \sum_{s=r}^{N-(n-r)} \Pr(S_n = r | S_N/N = \theta(s)) \Pr(S_N/N = \theta(s)), \tag{3.10}$$

where $\theta(s) \equiv s/N$. This is true for every $N \ge n$. If we take the limit $N \to \infty$ on the RHS in Equation 3.10 with $n$ and $r$ fixed, the sum will converge to an integral and give us an integral like Equation 3.5. 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

$$F_N(\theta) = \Pr(\Theta_N \le \theta),$$

$$= \sum_{s=0}^{N} \Pr(S_N = N\theta(s)) \mathbb{I}_{\theta(s) \le \theta}, \tag{3.11}$$

for $\theta \in [0, 1]$. Give a "density" $f_N$ for the CDF $F_N$ and express $\Pr(S_n = r)$ in terms of $f_N$.

**Definition 3.13.** For $0 \le \tau \le 1$ denote by $\delta_\tau(\theta)$ the Dirac delta-function giving the density of a distribution $\delta_\tau$ 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

$$\int_0^1 g(\theta) \delta_\tau(\theta) \, d\theta = g(\tau). \qquad \diamond$$

As we take $\theta$ increasing past $\theta(s)$ for some $s$ in $0, 1, ..., N$ in Equation 3.11, 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* 3.14. 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 3.15.** *Let $dF_N(\theta) \equiv f_N(\theta) d\theta$. For $0 \le r \le N$ it holds that*

$$\Pr(S_n = r) = \int_{r/N}^{1-(n-r)/N} \Pr(S_n = r | S_N = N\theta)\, dF_N(\theta). \tag{3.12}$$

*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 $g_N$,

$$\int_{r/N}^{1-(n-r)/N} \Pr(S_n = r | S_N = N\theta) dF_N(\theta) = \int_0^1 \mathbb{I}_{r \le N\theta \le N-(n-r)} \Pr(S_n = r | S_N = N\theta) f_N(\theta) d\theta$$

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

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

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

and now substituting for $g_N$ and using the indicator function to trim the sum,

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

and the RHS is $\Pr(S_n = r)$ by Eqn. 3.10. □

**Proposition 3.16.** *There exists a distribution $F(\theta)$ such that*

$$\Pr(S_n = r) = \binom{n}{r} \int_0^1 \theta^r (1 - \theta)^{n-r} dF(\theta), \tag{3.13}$$

*for every $r, n$ satisfying $0 \le r \le n$.*

*Proof.* The expression we have for $\Pr(S_n = r)$ in Proposition 3.15 holds for every $N \ge n$. Since $\Pr(S_n = r)$ does not depend on $N$, the RHS of Equation 3.12 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 $\theta$, the hypergeometric probability $\Pr(S_n = r | S_N = N\theta)$ converges uniformly as a function of $\theta$ 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 $N_1 \le N_2 \le N_3 \le ...$ of $N$-values. On the subsequence,

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

where $\Theta \sim F$ and $\Theta = \lim_{N\to\infty} N^{-1}S_N$. □

Equating the RHS of Equations 3.8 and 3.13 for $\Pr(S_n = r)$ and cancelling $n$ choose $r$ we obtain

$$p_{1:n}(x_1, ..., x_n) = \int_0^1 \theta^r (1-\theta)^{n-r} dF(\theta) \tag{3.14}$$

where $r = \sum_{i=1}^n x_i$, the result of de Finetti given in Equation 3.5.

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 3.14 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.

## 3.3 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

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

for the data. These distributions exist by de Finetti. Here $F$ is the distribution giving the true generative model for $\theta$, and we should use this as our prior. It could be called "nature's prior". Our own prior $\pi$ may not coincide with the true generative model $F$ but at least we a modeling something that exists. 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. At this level the modelling tasks in Frequentist and Bayesian inference are similar.

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

$$p(x_{m+1:n} | x_{1:m}) = p(x_{1:n})/p(x_{1:m})$$
$$= \int p(x_{m+1:n} | \theta) \frac{p(x_{1:m} | \theta) dF(\theta)}{p(x_{1:m})}$$
$$= \int p(x_{m+1:n} | \theta) d\tilde{F}(\theta).$$

After we get the data $X_1 = x_1, ... X_m = x_m$ there exists a generative model

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

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.

# 4 The Savage Axioms

Consider the following conversation.

Student: What is subjective probability?

Teacher: It expresses the strengths of preferences you hold for different sets of outcomes.

Student: What is preference?

Teacher: Your order $A \succ B$ express the belief that $\theta \in A$ is more likely to hold than $\theta \in B$.

Student: What do you mean by "more likely to hold"?

Teacher: It's subjectively more probable.

Student: Weren't you defining subjective probability?

The Savage Axioms break this loop. 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 $\pi$ satisfying the Axioms of Probability and expressing those preferences, so $\pi(A) > \pi(B)$ etc.

The Savage Axioms and de Finetti's theorem make a pair, just as the Axioms of Probability and the Frequentist definition of probability make a pair. The Axioms of Probability define the "algebraic" properties of a probability space. The Frequentist definition of probability (proportion of successes in an infinite sequence of trials) shows us that a data distribution $p^*(\cdot)$ exists for our problem. It is our job to make an observation model $p(\cdot|\theta)$ that satisfies the AoP and approximates $p^*$.

In the same way, the Savage Axioms (SA) just define the algebraic properties of preference relations, they don't say how preference is realised in the world. If preference relations satisfy the SA then a unique prior satisfying the AoP exists. de Finetti's Theorem shows us that a generative model $\psi \sim F$, $y \sim p^*(\cdot|\psi)$ exists for our problem. It is our job to write down a set of preference relations satisfying the SA and giving $\pi(\cdot)$ approximating $F$.

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 Lecture 4: Utility theory

Review the material in Section 1.3.3 on Decision theory before continuing.

### 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 $\theta$. In some settings $\theta$ is a parameter and $\delta$ is an estimator (so, each action is a distinct function of the data $\hat{\theta} = \delta(y)$) but $\delta$ might select a lottery and then $\theta$ is the outcome, win or lose. When we come to the Savage Axioms, the action will be a set $\delta \in \mathcal{B}_\Omega$ and the reward will be $r(\theta, \delta) = \mathbb{I}_{\theta \in \delta}$. We assume rewards, $r \in \mathbb{R}$ are bounded so $\mathbb{R} = [r_{\min}, r_{\max}]$ or $\mathbb{R} = \{r_{\min}, r_{\min} + 1, ..., r_{\max}\}$. Let $U(r)$ denote the utility of reward $r$. In the literature we are referencing (principally Maurice DeGroot, "Optimal Statistical Decisions" (1970)), 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 1.3.3,

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

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, in the parameter-estimation setting, $\theta \sim \pi(\cdot)$ and $y \sim p(\cdot|\theta)$. Conditioned on 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

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

If there is no data then the action doesn't depend on data (we would predict $\theta$ 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 $P_\delta$ 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 $P_{\delta,y}$ and $P_{\delta',y}$ over the same space of rewards. The action selects the reward distribution. When discussing two reward distributions, with fixed data or no data, we may drop the $\delta$-subscript and write $P, P'$ etc.

The *expected utility* of the action $\delta$ 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 1.3.3:

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

The 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 $\delta$ 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 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 $\theta$ 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

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

and the utilities are $U(0) = 0$ and $U(1) = u$ with $u > 0$. 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 $\phi$ then our prior for $\theta$ is

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

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

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

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

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 $R \sim P(\cdot)$ we are choosing between two reward distributions $P$ and $P'$ where $P'$ puts probability one on the reward $R' = r_0$. For a strictly concave utility and continuous $P$ we take the average reward, since $E_P(U(R)) < U(E_P(R)) = U(r_0) = E_{P'}(U(R'))$ by Jensen's inequality so $E(U(r)) < U(r_0)$. ♠

## 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 $\theta$ exists and $p(y)$ is finite (so the posterior is proper). If this prior exists and is unique we have *coherent belief*. Since everything hinges on the existence of the prior, we focus on the reward distribution defined in Equation 4.3. If coherent inference is possible without data then we assume it is possible with 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. We need to assign prior probabilities $\pi(A)$ to sets $A \in \mathcal{B}_\Omega$, where $\mathcal{B}_\Omega$ is a $\sigma$-field of $\Omega$ containing all the sets of interest, in such a way that $(\Omega, \mathcal{B}_\Omega, \pi)$ is a probability space. If we can do that then the prior distribution $\pi$ clearly exists.

Consider two sets, $A, B \in \mathcal{B}_\Omega$ 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 $\delta = A$ under our prior is

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

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

$$E_\theta(U(r(\theta, \delta))) = P_\delta(0)U(0) + P_\delta(1)U(1)$$
$$= u\pi(A),$$

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

**Definition 4.3.** A preference order $\succeq$ is a binary relation over pairs of elements $A, B \in \mathcal{B}_\Omega$. By definition $A \sim A$ and if $A \succeq B$ then either $A \succ B$ or $A \sim B$. We don't really need $\prec$ or $\preceq$ but it's convenient so add $A \succ B \Leftrightarrow B \prec A$ (by definition) and if $A \preceq B$ then either $A \prec B$ or $A \sim B$.[7] $\diamond$

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

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

for every pair of sets $A, B \in \mathcal{B}_\Omega$ and similarly for the $\prec$ and $\sim$ relations. We say that the probability distribution $\pi(A)$, $A \in \mathcal{B}_\Omega$ *expresses the preference order* in this case. If our preferences $\succeq$ over 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 $\delta, \delta'$. Define a second set of order relations $\succeq_U$ satisfying

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

and

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

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

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

and similarly for all pairs of reward distributions $P_\delta, P_{\delta'}$ generated by actions $\delta, \delta' \in \mathcal{D}$. If our preferences $\succ_U$ on reward distributions satisfy the Savage axioms for utility (given below) then a utility function exists and is unique. If the prior $\pi$ 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 $\pi$ 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.

---

[7]If you are familiar with general order relations you might be wondering why we left off the usual non-reflexive and transitive conditions. They will be added or implied by the axioms.

|        | Option 1 | Option 2 |
|        | color$_{urn}$ | color$_{urn}$ |
|--------|----------|----------|
| 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 can bet 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, $E(U|\text{choose } r_B) = u(1 - E_\pi(\phi))$, as in Example 4.1, 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 prior for $\phi$ doesn't 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, £750K.

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

$$E(U|A) = u \qquad\qquad E(U|B) = 0.1 + 0.89u$$
$$E(U|C) = 0.11u \qquad\qquad E(U|D) = 0.1$$

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   Lecture 5: 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 a utility function 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 $\sigma$-field of sets in $S$ satisfying $S \in \mathcal{S}$, $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S}$ and any set formed by taking countable unions and intersections of sets in $\mathcal{S}$ is also 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) $\pi(A) \geq 0$, $A \in \mathcal{S}$; (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 preference

We begin with the Savage axioms for preference. Our main interest will be on Axioms 1-3 and existence of the prior following from Axioms 1-5.

**Theorem 4.4.** *Let $S$ be a sample space and let $\mathcal{S}$ be a $\sigma$-field of sets in $S$. Let a system $\succeq$ 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 4.4.1.*

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

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

Remark: if every pair of sets $A, B \in \mathcal{S}$ is ordered by preference or equal in preference, how can we have a $\succeq$ 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 $x = y$ 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 $A_i \succeq B_i, i = 1, 2$ then

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

If in addition either $A_1 \succ B_1$ or $A_2 \succ B_2$ then $A_1 \cup A_2 \succ B_1 \cup B_2$.

Remark: If $A = A_1 \cup A_2$ can happen two mutually exclusive ways, and each is more likely than corresponding events leading to $B = B_1 \cup B_2$ then $A$ is more likely then $B$.

**Axiom 3**. If $A \in \mathcal{S}$ then $\emptyset \preceq A$. Furthermore $\emptyset \prec S$.

Remark: $\emptyset \prec S$ 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 (PS2) from Axioms 1-3 that the order is transitive (if $A \preceq B$ and $B \preceq C$ then $A \preceq C$) and $A \preceq B \Rightarrow A^c \succeq B^c$. 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* 4.5. 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 standard preferences preferred $r_A$ to $b_B$ (so $A_1 \succ B_1$) and $b_A$ to $r_B$ (so $A_2 \succ B_2$). By Axiom 2 we have

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

but $A_2 = A_1^c$ and $B_2 = B_1^c$ so $S \succ S$, which contradicts Axiom 1, as $S \sim S$ by Definition 4.3. ♠

Axioms 4 and 5 and the proof of Theorem 4.4 are outside the scope of the course. I include some comments on these as an Appendix in Section 4.4.1.

### 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 $\pi$ 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* 4.6. Let $A, B, D$ be sets in $\mathcal{S}$. Show that if $A \cap D = B \cap D = \emptyset$ then $A \cup D \succ B \cup D$ if and only if $A \succ B$. ANS: see PS2. ♣

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* 4.7. 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 $A \preceq B$ (in preference order). To see this suppose $A \succ B$. By the exercise we can remove $A$ from both sides to get $\emptyset \succ B \setminus A$ 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 $\pi$ expresses this system of preferences so $A \preceq B$. ♠

### 4.3.4 Axioms of utility

Let $\mathcal{P} = \{P_\delta\}_{\delta \in \mathcal{D}}$ denote a set of reward distributions $P \in \mathcal{P}$ over a common space of bounded rewards $r \in \mathbb{R}$. 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 $P, P' \in \mathcal{P}$,

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

based on expected utility. If we *start* with a set of preference relations $\succeq_U$ 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. 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!

**Theorem 4.8.** *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 4.4.2.*

Proof (outside the syllabus). See remarks in Appendix 4.4.2.

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: $P \succ_U P'$ or $P \prec_U P'$ or $P \sim_U P'$.

Axiom 7: If $P \succeq_U P'$ and $P' \succeq_U P''$ then $P \succeq_U P''$.

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

Axiom 8: let $0 < \alpha < 1$ and $P, P', P'' \in \mathcal{P}$ be given. Then $P' \succ_U P''$ if and only if

$$\alpha P' + (1 - \alpha)P \succ_U \alpha P'' + (1 - \alpha)P.$$

*Example* 4.9. 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

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

Since $P$ is the same for options (A) and (B) and we prefer (A) we have $P' \succ_U P''$ by Axiom 8. Similarly, $\tilde{P}$ is the same for (C) and (D) so a preference for (D) leads to $P' \prec_U P''$, a contradiction. It seems Allais posed this example to Savage himself, who preferred (A) to (B) and (D) to (C)! ♠

Axioms 9 and 10 and the proof of Theorem 4.8 are outside the scope of the course. I include some comments on these as an Appendix in Section 4.4.2.

### 4.3.5 Conclusions

What should we 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. When we derive a prior from a simple physical model for the process generating $\theta$ (as in Section 1.4) we largely avoid this sort of consideration and there is usually 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.

We often have a candidate prior with hyper-parameters, and we choose the hyper-parameters to make the prior representative of prior knowledge. For a more interesting question in practice try "is this prior representative of available information?". Also, we may satisfy the SAs with a bad prior. Hence Prof David Cox's criticism of the Savage Axioms: "what's missing is the truth".

## 4.4 Appendices

### 4.4.1 Appendix for Section 4.3.2 - Axioms of Preference

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 $A_i \succeq B$ for each $i = 1, 2, ...$ then $\bigcap_{i=1}^{\infty} A_i \succeq B$.

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**. Let $X \sim U(0, 1)$. For each $p \in [0, 1]$ there is $B \in \mathcal{S}$ satisfying $B \sim \{X \leq p\}$.

The trick here is to add all the sets of events $\{X \leq p\}$, $p \in [0, 1]$ with known probabilities to the set of possible outcomes. 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 $\sigma$-algebra. We can then fix $\pi(A), A \in \mathcal{S}$ by finding a matching event with known probability. Given $A$, we show there exists $B_p \in \mathcal{S}$ 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.

### 4.4.2 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 $P \succ P'$.

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 feasible. 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 $P \succeq P' \Leftrightarrow E(U|P) \geq E(U|P')$.

# 5 Markov chain Monte Carlo Methods

## 5.1 Lecture 6: MCMC

### 5.1.1 Introduction

MCMC is a family of algorithms for simulating $X_0, X_1, X_2, ....$ so that $X_t \xrightarrow{D} p$, (meaning $X_t$ converges to $p$ in distribution) for a user-defined probability distribution $p$. When we come to use this $p$ will be $\pi$, 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 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. I will quote theory for the case that $\Omega$, 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 $\theta$, $\pi(\theta)$ and $\pi(\theta|y)$ using finite precision arithmetic so we are really working with finite $\Omega$ anyway. I sometimes refer to this as the *computer measure*.

### 5.1.2 Markov chains

Let $\{X_t\}_{t=0}^{\infty}$ be a homogeneous Markov chain of random variables on $\Omega$ with starting distribution $X_0 \sim p^{(0)}$ and transition probability 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 let $p_i^{(t)} = \mathbb{P}(X_t = i)$, with $p^{(t)} = (p_i^{(t)})_{i \in \Omega}$ 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 $P_{i,j}^{(n)} > 0$. A state $i \in \Omega$ is *aperiodic* if $P_{i,i}^{(n)}$ is non zero for all sufficiently large $n$ and the chain is aperiodic if all its states are aperiodic. If the chain is irreducible and any state is aperiodic then the chain is aperiodic[8].

### 5.1.3 The Stationary Distribution and Detailed Balance

In discussing Markov chains we will work with a generic "target" distribution $p = (p_i)_{i \in \Omega}$ (with $p_i = \Pr(X = i)$ if $X \sim p$, and $p$ 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 $\pi(\theta|y)$, $\theta \in \Omega$.

---

[8]See Norris, *Markov Chains* CUP, (2012).

The probability mass function (PMF) $p_i, i \in \Omega$, $\sum_{i \in \Omega} p_i = 1$ is a *stationary distribution* of $P$ if $pP = p$. By the partition theorem for probability (PTP),

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

which is $p_j^{(1)} = [p^{(0)}P]_j$ so if the chain starts with $p^{(0)} = p$ and $pP = p$ then $p_j^{(1)} = p_j$ also. Iterating, $p^{(t)} = p$ for each $t = 1, 2, \ldots$ in the chain, so the distribution of $X_t \sim p$ 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 $X_t \xrightarrow{D} p$. The convergence theorem for finite irreducible Markov chains tells us that if $\Omega$ is finite, if $pP = p$, and if $P$ is irreducible and aperiodic, then indeed $X_t \xrightarrow{D} p$. However, checking $pP = p$ is hard, as we have to sum over all $\Omega$ to evaluate $pP$.

**Definition 5.1.** *Detailed Balance (DB, 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. $\diamond$

*Exercise* 5.2. 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*.

### 5.1.4 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 $X_t \xrightarrow{D} p$, then $X_t \sim p^{(t)}$ with $p^{(t)} \simeq p$ at large $t$, so when we look at our Markov chain $X_0, X_1, \ldots, X_T$, "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 5.3.** *(Ergodic Theorem) If $\{X_t\}_{t=0}^{\infty}$ is an irreducible and aperiodic Markov chain on a finite space of states $\Omega$ satisfying detailed balance with respect to the probability distribution $p$, then as $T \to \infty$*

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

*for any bounded function $f : \Omega \to R$. The convergence is almost sure (a.s.). 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*]

### 5.1.5 The Metropolis-Hastings Algorithm

Suppose we need samples from a pmf $p_i, i \in \Omega$ with $\Omega$ 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 $P_{i,j} = P(X_{t+1} = j | X_t = i)$ and hence the transition matrix $P$. The algorithm is constructed to ensure the chain targets $p$.

We simulate a random walk $X_0, X_1, X_2, \dots$ in $\Omega$ by accepting or rejecting proposals from a simple irreducible transition matrix $Q_{i,j}, i, j \in \Omega$ which we 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 5.4.** *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 $t = 1, 2, 3, \dots T$ to simulate the samples we need. $\Diamond$

**Lemma 5.5.** *If the Markov chain realised by the algorithm given in Definition 5.4 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 $\Omega$ is finite, it is sufficient to show that the transition matrix determined by the random MCMC update satisfies detailed balance with $p$, by Theorem 5.3, 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
$$P_{i,j} = q(j|i)\alpha(j|i).$$

Now check DB:

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

and we are done. $\square$

### 5.1.6 Example: Simulating the hypergeometric distribution

The hypergeometric distribution $\text{HyperGeom}(k; K, N, n)$ with parameters $K = 10, N = 20, n = 10$ gives the probability for $k$ successes in $n$ draws (without replacement) from a population of size $N$ containing $K$ successes. If $p_k = \text{HyperGeom}(k; K, N, n)$ then for $k \in \Omega$,

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

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

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

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 to 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 $B^- \le j \le B^+$ and

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

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 $\Omega$ for any $u$ in the MCMC step 2.

Step 3: check the chain is irreducible and aperiodic. This can be seen since $q$ allows us to visit any state in $\Omega$ 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 6.

### 5.1.7 Notation for the continuous case

In this section we give notation for the continuous case and write down detailed balance. We give the extension (used often in practice) to the case where we choose a MH-update at random from a set of candidates.
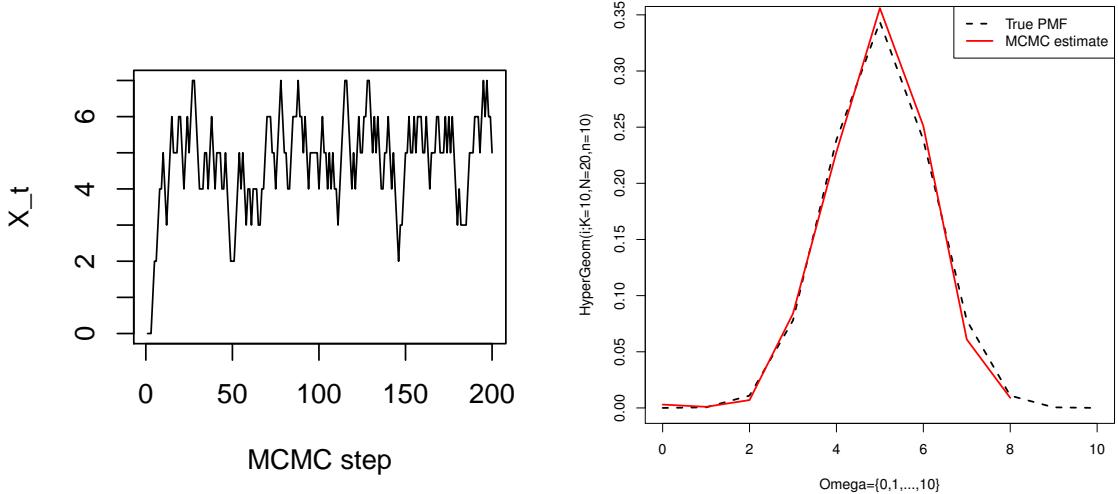
Figure 6: Summary output of the R-implementation given in the text. Left: $x$-axis is step counter $t = 1, 2, 3...200$. The $y$-axis is Markov chain state $X_t$ targeting HyperGeom($K = 10, N = 20, n = 10$). Right: histogram of $X_1, X_2, ..., X_n$ for $T = 1000$.

If $\Omega$ is an open subset of $R^p$ then the random variable $\theta \in \Omega$ is continuous. If its density is $p(\theta)$ then the distribution $p(d\theta) = p(\theta)d\theta$ is defined for sets $A \in \mathcal{B}_\Omega$ so $p(A) = \int_A p(\theta)d\theta$ and the probability space is $(\Omega, \mathcal{B}_\Omega, p)$. In a simple setup we have a proposal distribution $q(d\theta'|\theta)$ with density $q(\theta'|\theta)$. In the proposal distribution $q(A|\theta) = \int_A q(d\theta'|\theta)$. The acceptance probability is

$$\alpha(\theta'|\theta) = \min \left\{ 1, \frac{p(\theta')q(\theta|\theta')}{p(\theta)q(\theta'|\theta)} \right\}. \tag{5.1}$$

With these substitutions, the Metropolis-Hastings algorithm is otherwise unchanged. 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 $\theta$ at the current step.

**Proposition 5.6.** *The transition kernel for Metropolis-Hasting MCMC is*

$$K(\theta, d\theta') = \alpha(\theta'|\theta)q(d\theta'|\theta) + c(\theta)\delta_\theta(d\theta'), \tag{5.2}$$

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

$$\Pr(X_{t+1} \in A | X_t = \theta) = c(\theta)\mathbb{I}_{\theta \in A} + \int_A \alpha(\theta'|\theta)q(d\theta'|\theta)$$

$$= \int_A K(\theta, d\theta'),$$

Page 51

where the second line follows by substituting the proposed form for $K$ given in Eqn. 5.2 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. $\square$

*Remark* 5.7. The term involving $c(\theta)$ is the probability for rejection and ensures that when we integrate $K(\theta, d\theta')$ over $\Omega$ we get one. We previously only considered the case where $\theta' \neq \theta$. ✠

**Definition 5.8.** 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 $\theta' \in B$,
$$\int_B p(d\theta') \int_A K(\theta', d\theta) = \int_A p(d\theta) \int_B K(\theta, d\theta'). \tag{5.3}$$
must hold for every pair of sets $A, B \in \mathcal{B}_\Omega$. $\diamond$

*Remark* 5.9. If this holds then the process is stationary. Take $A = \Omega$ in Eqn. 5.3 to obtain
$$p(B) = \int_\Omega K(\theta, B)p(d\theta), \quad \text{for all } B \in \mathcal{B}_\Omega. \tag{✠}$$

**Proposition 5.10.** *Detailed balance holds for $K$ and $p$ in the MH algorithm if and only if*
$$\int_B p(d\theta') \int_A q(d\theta|\theta')\alpha(\theta|\theta') = \int_A p(d\theta) \int_B q(d\theta'|\theta)\alpha(\theta'|\theta). \tag{5.4}$$
*for every pair of sets $A, B \in \mathcal{B}_\Omega$.*

*Proof.* Substitute Eqn. 5.2 into Eqn. 5.3. We get Eqn. 5.4 because the terms involving $c$ will cancel: on the RHS we get
$$\int_A p(d\theta) \int_B c(\theta)\delta_\theta(d\theta') = \int_A p(d\theta)c(\theta)\mathbb{I}_{\theta \in B},$$
$$= \int_{A \cap B} p(d\theta)c(\theta)$$
from this term and on the LHS we get $\int_{B \cap A} p(d\theta')c(\theta')$. These are equal, so the contribution cancels and we are left with Eqn. 5.4. We can reverse this reasoning for "iff". $\square$

*Remark* 5.11. In terms of probability densities, detailed balance in Eqn. 5.4 is
$$p(\theta')q(\theta|\theta')\alpha(\theta|\theta') = p(\theta)q(\theta'|\theta)\alpha(\theta'|\theta), \tag{5.5}$$
and this may readily be verified for the acceptance probability in Eqn. 5.1. ✠

### 5.1.8 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 $\mathbb{R}^2$,
$$p(\theta) = (2\pi)^{-1} \left( 0.5e^{-(\theta-\mu_1)\Sigma_1^{-1}(\theta-\mu_1)/2} + 0.5e^{-(\theta-\mu_2)\Sigma_2^{-1}(\theta-\mu_2)/2} \right)$$
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

$$\theta'_i \sim U(\theta_i - a, \theta_i + a), \text{ independently for } i = 1 \text{ and } i = 2,$$

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 $q(\theta'|\theta) = q(\theta|\theta') = 1/4a^2$ (only needed up to a 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

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

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.

See code `First-MCMC-example.R` accompanying the lecture for an implementation. Try implementing your own MCMC and 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 7 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 7 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 $p(\theta^{(t)}) \gg p(\theta^{(t+1)})$ (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 is 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.

### 5.1.9 Mixing updates for 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 $\xi_k$ and then accept or reject.

If $\theta = (\theta_1, ..., \theta_p)$ and we target $p(\theta)$, $\theta \in \Omega$ with $p$ at all large, we commonly update one variable at a time. Smaller changes to the state will typically have a higher acceptance probability (if the target $p(\theta)$ is a smooth function of $\theta$, 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 $\Omega$ if the change to the state is always small, so there is a trade off.

To be concrete we suppose the $k$'th update acts on the $k$'th component of $\theta$, so $N = p$ and $\theta'_j = \theta_j$, $j \neq k$. The proposal distribution is $q_k(d\theta'_k|\theta) = q(\theta'_k|\theta)d\theta'_k$. We further suppose (for simplicity) that $\theta_k \in \Omega_k$, $k = 1, \ldots, p$ and $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_p$ so that the space for $\theta_k$ doesnt
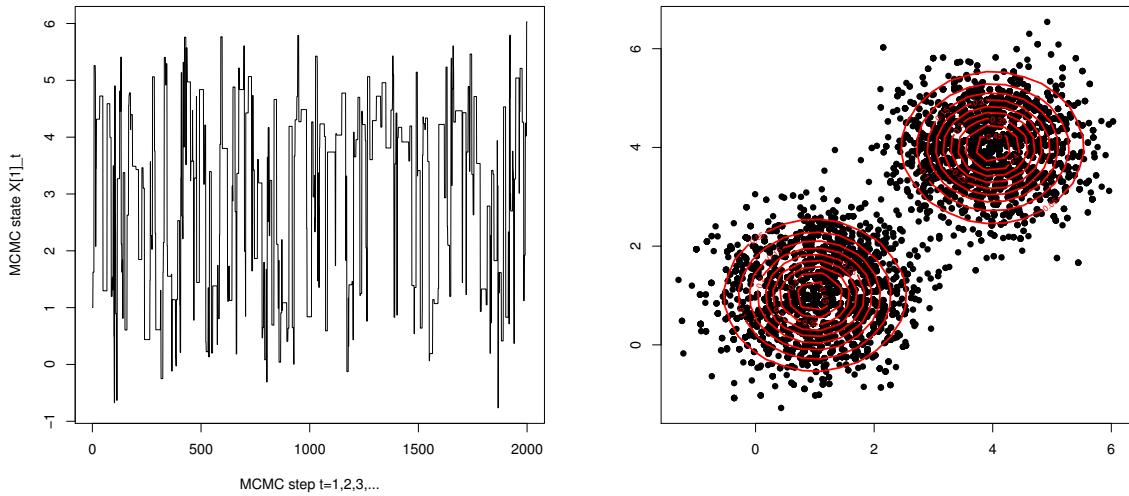
Figure 7: MCMC targeting bivariate normal mixture: (Left) MCMC trace of $\theta_1^{(t)}$ against $t$ for $t = 1, ..., 2000$; (Right) scatter plot of sampled parameter vectors $(\theta_1^{(t)}, \theta_2^{(t)})$, $t = 1, ..., 2000$.

depend on $\theta_{-k}$ (or take $p(\theta) = 0$ for $\theta \in R^p$ but not in $\Omega$). The algorithm given below is sometimes called "Metropolis within Gibbs", because it updates one component at a time, so the $k$'th update targets $p(\theta_k | \theta_{-k})$.

**Definition 5.12.** *Metropolis Hastings MCMC with a mixtures of updates*: the following algorithm simulates a Markov chain targeting $p(\cdot)$. For $k = 1, ..., p$ let $\theta' = (\theta_1, \ldots, \theta_{k-1}, \theta'_k, \theta_{k+1}, \ldots, \theta_p)$ and let $q_k(\theta'_k | \theta)$ be a *proposal probability density* satisfying

$$q_k(\theta'_k | \theta) > 0 \Leftrightarrow q_k(\theta_k | \theta') > 0 \qquad \text{for } k = 1, \ldots, p, \text{ with } \theta'_{-k} = \theta_{-k}.$$

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

1. Draw $k \sim \text{Multinon}(\xi_1, ..., \xi_p)$ and $\theta'_k \sim q_k(\cdot | \theta)$. Set $\theta'_{-k} = \theta_{-k}$ and draw $u \sim U[0, 1]$.

2. If $u \leq \alpha_k(\theta' | \theta)$ where

$$\alpha_k(\theta' | \theta) = \min \left\{ 1, \frac{p(\theta'_k | \theta'_{-k}) \, q_k(\theta_k | \theta')}{p(\theta_k | \theta_{-k}) \, q_k(\theta'_k | \theta)} \right\} \tag{5.6}$$

(and then as $\theta'_{-k} = \theta_{-k}$ so the factors $p(\theta'_{-k})$ and $p(\theta_{-k})$ cancel)

$$= \min \left\{ 1, \frac{p(\theta') \, q_k(\theta_k | \theta')}{p(\theta) \, q_k(\theta'_k | \theta)} \right\} \tag{5.7}$$

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. $\diamond$

This targets $p(\theta)$ (if it is irreducible etc). We are choosing a kernel at random from $K_1, ..., K_p$ and using it to update the state. The individual kernels are, from Eqn. 5.2,

$$K_k(\theta, d\theta'_k) = \alpha_k(\theta' | \theta) q_k(d\theta'_k | \theta) + c_k(\theta) \delta_{\theta_k}(d\theta'_k), \tag{5.8}$$

with $c_k(\theta) = `1 - \int_{\Omega_k} \alpha_k(\theta' | \theta) q_k(d\theta'_k | \theta)$.

*Exercise* 5.13. Show that

$$p(d\theta_k|\theta_{-k})K_k(\theta, d\theta'_k) = p(d\theta'_k|\theta'_{-k})K_k(\theta', d\theta_k)$$

so each kernel $K_k$ satisfies DB wrt $p(\theta_k|\theta_{-k})$.
ANS: this is really immediate because we see from Eqn. 5.6 that $K_k$ in Eqn. 5.8 is a standard Metropolis Hastings kernel targeting $p(\cdot|\theta_{-k})$, so Proposition 5.10 holds. ♣

*Exercise* 5.14. Show that if $X_t \sim p(\cdot)$ then $X_{t+1} \sim p(\cdot)$ so the process is stationary wrt $p(\cdot)$.
ANS: If $X_t \sim p(\cdot)$ and $(X_{t+1}|k) \sim p_k(\cdot)$ is the unknown distribution of $X_{t+1}$ given we update component $k$ then $p_k(\theta'_k|\theta_{-k}) = p(\theta'_k|\theta_{-k})$ ($p(\theta'_k|\theta_{-k})$ is the target in Equation 5.6 so it's stationary). Also, $p_k(\theta_{-k}) = p(\theta_{-k})$ (as $\theta_{-k}$ didnt change) so $p_k(\theta') = p(\theta'_k|\theta_{-k})p(\theta_{-k}) = p(\theta')$. Since the conditional distribution of $X_{t+1}|k$ doesnt depend on $k$ we have $(X_{t+1}|k) \sim X_{t+1}$ and so $X_{t+1} \sim p(\cdot)$ also, and hence the process is stationary with respect to $p(\cdot)$. ♣

I went a bit around the houses to show stationarity in the second exercise. That was to avoid some notationally awkward but conceptually simple measures. The overall kernel is

$$dK(\theta, \theta') = \sum_{k=1}^{p} \xi_k K_k(\theta, d\theta'_k).$$

We write $dK(\theta, \theta')$ rather than $K(\theta, d\theta')$ because the measure of the mixture is only supported on states $\theta'$ that differ from $\theta$ at exactly one component. When we write $K(\theta, d\theta')$ we imply that we have a distribution which is absolutely continuous with respect to $d\theta'$, volume measure on $R^p$, which is not the case. However, it is intuitively easy to see that the mixture kernel $dK$ satisfies DB, because each component does. We can safely update the state with any randomly chosen stationary update we like. A general form of DB

$$dp(\theta)dK(\theta, \theta') = dp(\theta')dK(\theta', \theta)$$

will hold, because it holds between each pair of terms with the same index $k$ on the left and right sides of DB. The same idea works much more generally if we have $N$ updates each satisfying DB, not just updating individual components of $\theta$ but groups of variables etc.

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 of parameters to update). There may be many different proposals connecting two states $\theta, \theta'$. 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 where it is related to the Hammersley-Clifford Theorem. We dont need that here as we simply assume irreducibility as a requirement on the mixture of kernels.

### 5.1.10   MH example: bivariate normals one variable at a time

Here is an example of mixing updates (one variable at a time) using the same bivariate normal target $p(\theta)$, $\theta \in \mathbb{R}^2$ as Section 5.1.8. Since we have $p = 2$ variables we choose a variable to update with probabilities $\xi = (\xi_1, \xi_2)$. I took $\xi_1 = 1/2$ so $\xi = (1/2, 1/2)$. If I choose to update component $k = 1$ then I take $\theta'_1 \sim U(\theta_1 - a, \theta_1 + a)$ (with $a > 0$ fixed) and set $\theta'_2 = \theta_2$. This gives

$$q_1(\theta'_1|\theta) = 1/2a \times \mathbb{I}_{\theta_1 - a < \theta'_1 < \theta_1 + a},$$

and this cancels in the Hastings ratio in Equation 5.7. If I choose $k = 2$ it's the same but with $1 \leftrightarrow 2$. The acceptance probability is unchanged.

The algorithm is as follows. Suppose $\theta^{(t)} = \theta$ with $\theta \in \mathbb{R}^2$.
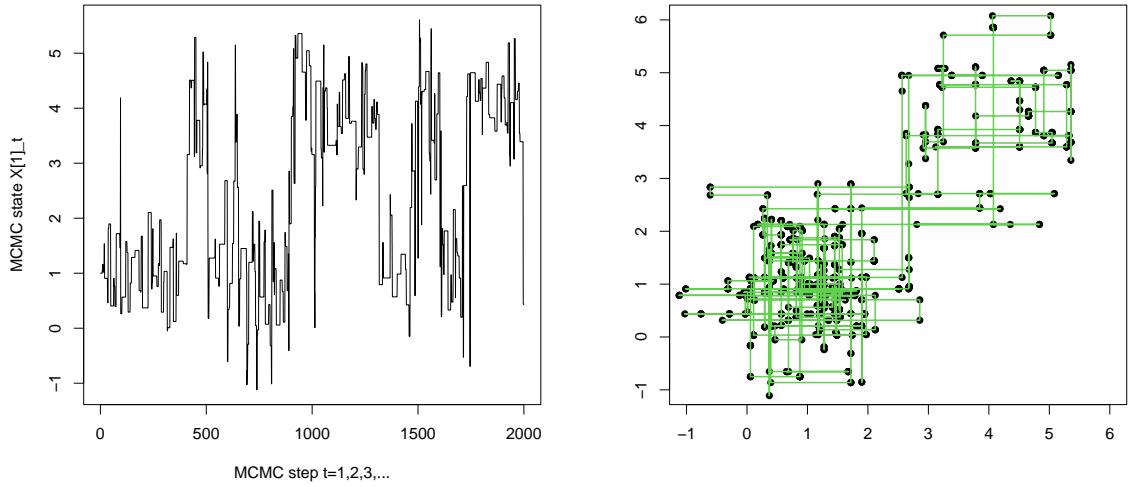
Figure 8: MCMC targeting a bivariate normal mixture updating one variable at a time: (Left) MCMC trace of $\theta_1^{(t)}$ against $t$ for $t = 1, ..., 2000$; (Right) scatter plot of sampled parameter vectors $(\theta_1^{(t)}, \theta_2^{(t)})$, $t = 1, ..., 2000$ with green line segments connecting adjacent points $\theta^{(t)}$ and $\theta^{(t+1)}$.

1. Draw $k \sim U\{1, 2\}$ and simulate $\theta_k' \sim U(\theta_k - a, \theta_k + a)$. Set $\theta_{3-k}' = \theta_{3-k}$ and draw $u \sim U[0, 1]$.

2. If $u \leq \alpha_k(\theta'|\theta)$ where

$$\alpha_k(\theta'|\theta) = \min\left\{1, \frac{p(\theta')}{p(\theta)}\right\}$$

then set $\theta^{(t+1)} = \theta'$, otherwise set $\theta^{(t+1)} = \theta$.

Initialise this with $\theta^{(0)} = \mu_1$ (arbitrary start state).

See online code `First-MCMC-example.R` for an implementation. Simulated output is shown in Figure 8. We can see the path on the right makes axis-parallel steps. This makes it less probable (at any given step) that the chain will jump from one mode to the other. If the chain is close to one of the means so $\theta \simeq \mu_1$ or $\theta \simeq \mu_2$ then a large axis-parallel step will give a candidate state $\theta'$ in a region where $p(\theta')$ is very small. That means $p(\theta')/p(\theta)$ will be small and the proposal is likely to be rejected. Looking at the sequence of sampled values of $\theta_1$ on the left, we see it spends many updates in a mode before moving to the other mode. Notice the contrast with Figure 7 at left where the chain moves freely between the two modes. The chain in Figure 8 will still converge to the correct target (so if we run it long enough we will get a scatter plot like Figure 7 at right) but it takes more updates to generate samples which are representative of the target. We will return to this when we discuss convergence and mixing.

## 5.2 Lecture 7: The Gibbs sampler and data augmentation

### 5.2.1 The Gibbs sampler

Consider a $p$-dimension target distribution $p(\theta)$, $\theta \in \mathbb{R}^p$, with $\theta = (\theta_1, ..., \theta_p)$ as in Section 5.1.9. *Random scan Gibbs* is a multi-component Metropolis Hastings sampler of the form given in Section 5.1.9 which selects components $i = 1, ..., p$ for update with probability $\xi_i = 1/p$ and takes as

proposal density the conditional density, $q_i(\theta_i|\theta) = p(\theta_i|\theta_{-i})$ where

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

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

*Exercise* 5.15. 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 computer measure at least) then the chain is an ergodic chain targeting (the computer's approximation to) $p(\cdot)$. The description above assumes the parameters are taken one at a time, so the conditionals are univariate. If we can 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 $\theta_i$ 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 $p(\theta_1, \theta_2)$ and $X_t = (\theta_1, \theta_2)$.

Algorithm:

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

**Proposition 5.16.** *If $X_t \sim p(\cdot)$ then after these two steps we have a new correlated sample $X_{t+1} \sim p(\cdot)$, so the process is stationary wrt $p(\cdot)$.*

*Proof.* If the distribution of $X_{t+1} = (\theta'_1, \theta'_2)$ is some unknown distribution $X_{t+1} \sim P(\cdot)$ then

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

as $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta'_1)$ are normalised in the last step. The same approach works for $p > 2$. □

### 5.2.2   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 model 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,

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

This is an obstacle for MCMC as we cant easily calculate ratios $\pi(\theta'|y)/\pi(\theta|y)$. These are needed in the acceptance probability $\alpha(\theta'|\theta)$.

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

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

If we sample $\theta, z \sim \pi(\theta, z|y)$ and ignore $z$ then the distribution of $\theta$ is $\pi(\theta|y)$, just what we want. Now we must have MCMC updates for both $z$ and $\theta$. However the acceptance probability $\alpha(\theta', z'|\theta, z)$ now depends on the ratio $\pi(\theta', z'|y)/\pi(\theta, z|y)$, which is easily evaluated.

This idea - of treating missing data $z$ as simply another parameter like $\theta$, 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 $y = (y_1, ..., y_n)$ or covariate values 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.

### 5.2.3   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 $i$th 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 $\Phi$ is the cdf of $N(0, 1)$. If the prior for $\theta$ is $\pi(\theta)$ then the posterior for $\theta|y$ is

$$\pi(\theta|y) \propto \pi(\theta) \prod_{i=1}^{n} \Phi(\eta_i)^{y_i} (1 - \Phi(\eta_i))^{1-y_i} \tag{5.9}$$

with $\eta_i = \eta_i(\theta), i = 1, ..., n$. We cant calculate conditionals $\pi(\theta_j|\theta_{-j}, y)$ as $\theta$ appears inside $\Phi$ 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 $z \sim \pi(\cdot|\theta)$ with

$$z_i \sim N(\eta_i, 1), \quad \text{iid for i=1,...,n}$$

then $z_i = \eta_i + W_i$ with $W_i \sim N(0, 1)$ iid for $i = 1, ..., n$, so $\Pr(z_i > 0|\theta) = \Pr(W_i > -\eta_i|\theta) = \Phi(\eta_i)$ 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

$$\Pr(y_i = 1|\theta) = \Pr(z_i > 0|\theta)$$
$$= \Phi(\eta_i)$$

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

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

The marginal for $\theta$ is just $\pi(\theta|y)$ in Eqn. 5.9, since

$$\pi(\theta|y) \propto \pi(\theta) \prod_i \int_{-\infty}^{\infty} \pi(z_i|\theta)\mathbb{I}_{y_i = \mathbb{I}_{z_i > 0}} dz_i$$
$$= \pi(\theta) \prod_i \left\{ \begin{array}{ll} \Phi(\eta_i) & \text{if } y_i = 1 \\ 1 - \Phi(\eta_i) & \text{if } y_i = 0 \end{array} \right.$$
$$\propto \pi(\theta)p(y|\theta),$$

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 $\theta$-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[9] in $\theta$ 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

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

conditionally independent for $i = 1, \dots, n$. Here then is the Gibbs sampler targeting $\pi(\theta, z|y)$ by alternating between $\theta$-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 the online code and applied to a simple example.

## 5.3 Output analysis

See CJ Geyer "Practical Markov Chain Monte Carlo" Statistical Science (1992) and Sokal "Monte Carlo Methods in Statistical Mechanics", Springer (1996) for background (I have a copy).

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

---

[9]See PS2 and R code example.

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.

### 5.3.1 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 $p^{(0)} = p$, 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 9 shows output from two MCMC runs for the normal mixture example in Section 5.1.8 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.

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

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

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 $\text{var}(\bar{f}_n)$ 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
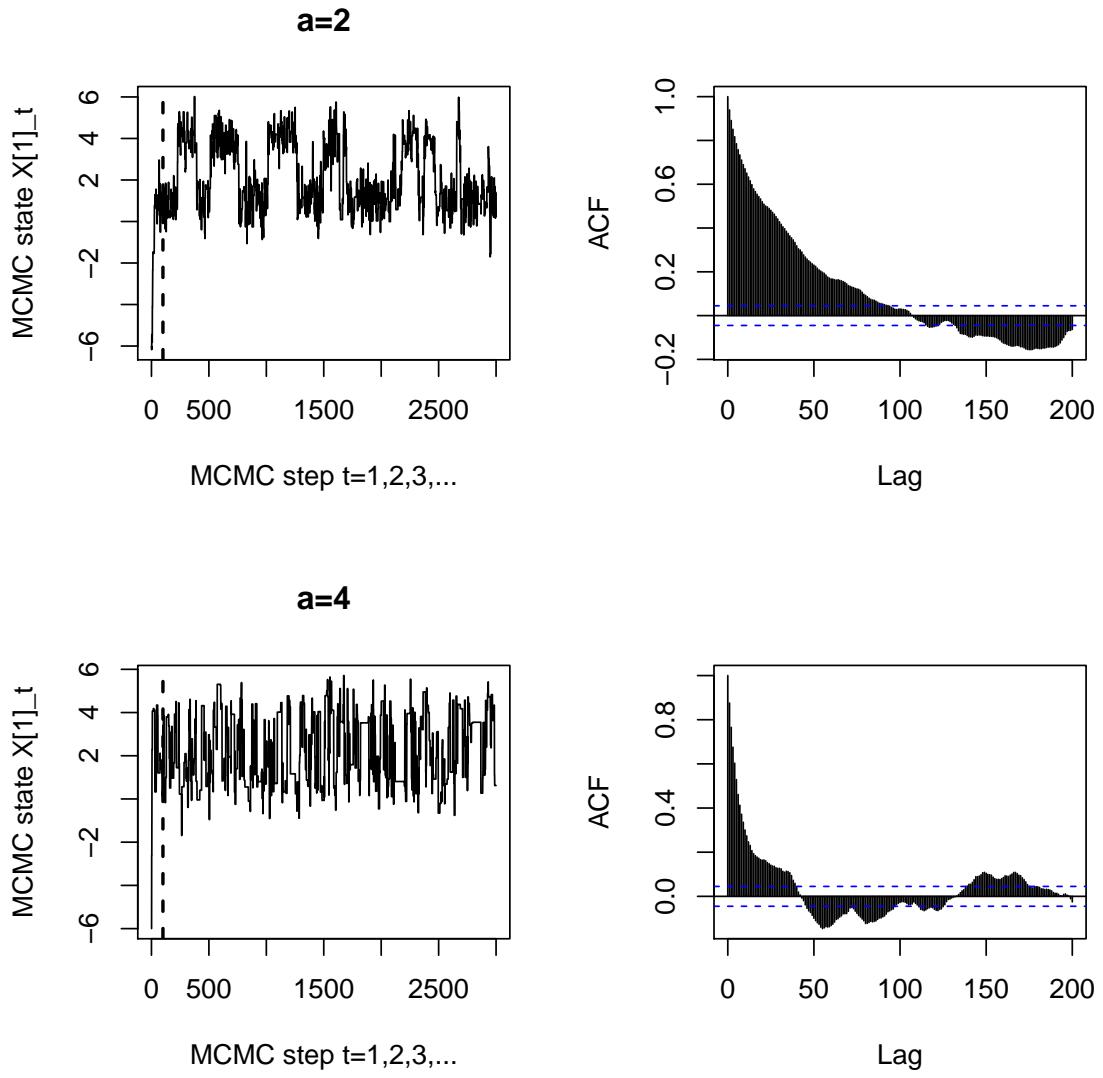
Figure 9: MCMC targeting the normal mixture in Section 5.1.8: (Left column) run traces for $\theta_1^{(t)}$, $t = 1, ..., 3000$ taking the jump size $a = 2$ (top) and $a = 4$ (bottom); (Right column) autocorrelation plots for the MCMC output sequences at left. Dotted lines in the ACF plots show asymptotic standard deviation estimates for the ACF estimates.

$\theta^{(k,t)}, \; k = 1, ..., K, \; t = 1, ..., n.$ These samples are correlated within a run but independent across runs and can be used to estimate $\text{var}(\overline{f}_n)$. We estimate $\overline{f}_{k,n} = n^{-1} \sum_t f(\theta^{(k,t)})$ as the average from each run, and then estimate $\text{var}(\overline{f}_n) \simeq \hat{\sigma}^2_{f,n}$ with

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

from the standard error across independent runs. Now $\hat{\sigma}^2_{f,n}$ 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 $\hat{\sigma}^2_{f,n} \simeq \hat{\sigma}^2_f / ESS$ so the ESS for a single run of length $n$ is approximately

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

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 $\text{var}(\overline{f}_n)$ given above wouldnt be sensible, as we incur a burn-in cost for each run. This leads to alot of discarded 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 $\sigma^2_f = \text{var}(f(X_t))$, $X_t \sim p$. This doesn't depend on $t$ because the chain is stationary. Express $\text{var}(\overline{f}_n)$ in terms of $\rho_s$. This gives insight and leads to an estimator for $\text{var}(\overline{f}_n)$, since we can estimate $\rho_s$ from the MCMC samples.

**Proposition 5.17.** *Let $f(X_t), \; t \geq 0$ be a stationary Markov Chain satisfying $\sum_{s=0}^{\infty} |\rho_s| < \infty$ and let $\tau_{f,n} = 1 + 2 \sum_{s=0}^{n-1} \rho_s$ (the* Integrated Autocorrelation Time*, IACT). If $\tau_f = \lim_{n \to \infty} \tau_{f,n}$ then*

$$\lim_{n \to \infty} n var(\overline{f}_n) = \sigma^2_f \tau_f,$$

*with $var(\overline{f}_n) \simeq \sigma^2_f \tau_{f,n} / n$ at large $n$.*

*Proof.* Following Fearnhead et al. Scalable Monte Carlo for Bayesian Learning CUP (2025),

$$
\begin{aligned}
\mathrm{var}(\overline{f}_n) &= n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{cov}(f(X_i), f(X_j)) \\
&= \sigma^2 n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{|i-j|} \\
&= \sigma^2 n^{-2} \left[ n\rho_0 + 2 \sum_{s=1}^{n-1} (n-s)\rho_s \right] \\
&= \sigma^2 n^{-1} \left[ 1 + 2 \sum_{s=1}^{\infty} \max\left\{ 0, 1 - \frac{s}{n} \right\} \rho_s \right]
\end{aligned}
$$

and so by dominated convergence

$$
\begin{aligned}
\lim_{n \to \infty} n\mathrm{var}(\overline{f}_n) &= \sigma^2 \left[ 1 + 2 \sum_{s=1}^{\infty} \rho_s \right] \\
&= \sigma^2 \tau_f.
\end{aligned}
\qquad \square
$$

Here

$$
ESS = n/\tau_f
$$

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.

We can estimate[10] $\gamma_s = \mathrm{cov}(f(X_i), f(X_{i+s}))$ using

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

and $\gamma_0 = \mathrm{var}(f(X_i))$ (as usual) from the sample output, and compute $\hat{\rho}_s = \hat{\gamma}_s / \hat{\gamma}_0$. This leads to an estimate of $\tau_f$,

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

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 $\rho_s$ 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, $\Gamma_s = \rho_s + \rho_{s+1}$ 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 $\hat{\Gamma}_s > 0$ and $\hat{\Gamma}_s < \hat{\Gamma}_{s-1}$. As $n$ grows, the variance of $\hat{\rho}_s$ decreases, and $\hat{\Gamma}_s$ converges to a positive monotone function of $s$, so these conditions are violated at increasingly large values of $s$ (random, but this is the trend) and so the truncation bound grows with $n$. Geyer shows that this leads to a consistent estimate for $\tau_f$ and hence the ESS.

---

[10]See Priestly 1981 "Spectral Analysis and Timeseries" Academic London, pp323 for the factor $1/n$ where we expect $1/(n-s)$.
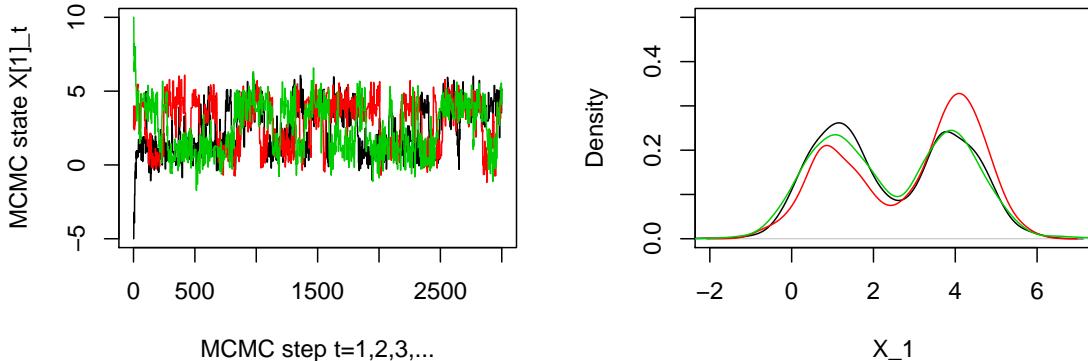
Figure 10: Two diagnostic plots for convergence checking: (Left) MCMC traces for selected statistics - in this case one of the components of the parameter vector - the log-likelihood is often a good choice; (Right) Histograms - or density plots - of a parameter of interest, which should be near identical. 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.

### 5.3.3 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 10 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.

# 6 Model selection: estimating the marginal likelihood

## 6.1 Lecture 8: Estimating the marginal likelihood using Monte Carlo

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, in order of increasing stability. I also give the Laplace estimator, which is biased, but (under regularity conditions) asymptotically consistent with increasing number of data points $n$ when $y = (y_1, \ldots, y_n)$. In the last decade this area has seen substantial research (seeking numericaly stable, efficient estimators). However, the new methods are often fairly complex, and restricted to cases where the parameter $\theta$ is continuous, or are designed for some specific class of posteriors. They are often built on the ideas below. The bridge estimator (below) and the Ratio Importance Sampler (see PS2) offer a good balance of power and simplicity.

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

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 $\theta$-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*: an importance sampling scheme using posterior samples.

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

$$\hat{p}_m^{(is)} = \frac{1}{T} \sum_t w_t \, p(y|\theta^{(t)},m)$$

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

$$E_{\theta^{(1:T)}|y,m}(\hat{p}_m^{(is)}) = T^{-1} \sum_t \int_\Omega w_t \, p(y|\theta^{(t)},m)\pi(\theta^{(t)}|y,m) \, d\theta^{(t)}$$

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

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

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

Now

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

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

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

and this is a consistent estimator by the continuous mapping theorem.

This "harmonic mean" estimator (ie, $\hat{p}_m^{(hm)}$) is not to be trusted. Though widely used, it is the *worst Monte Carlo method ever.* (Radford Neal's blog, 2008). 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. We can see in the normalised version, with weights $w_t$, that the target distribution ($\pi(\theta^{(t)}|m)$, numerator) has heavier tails than the IS proposal distribution ($\pi(\theta^{(t)}|y,m)$, denominator). This often leads to infinite variance weights.

*Bridge Estimate*: Bridge estimators generalise importance sampling and yield more stable estimates, as they minimise the mean squared error of the estimate. In the following we temporarily drop the model indicator $m$, so $p(y)$ is $p(y|m)$. We are working on just one model at a time so the posterior is $\pi(\theta|y) = \pi(\theta|y,m)$ etc.

**Proposition 6.1.**
$$p(y) = \frac{E_{\theta\sim\pi(\cdot)}(\pi(\theta)p(y|\theta)h(\theta))}{E_{\theta\sim\pi(\cdot|y)}(\pi(\theta)h(\theta))}$$

*where $h : \Omega \to R$ is chosen so that the expectations are finite and non-zero.*

*Exercise* 6.2. Verify this.
ANS: Replace the expectations with integrals, substitute $\pi(\theta|y) = \pi(\theta)p(y|\theta)/p(y)$ and cancel. ♣

*Exercise* 6.3. 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? ANS: $h(\theta) = 1/p(y|\theta)\pi(\theta)$. ♣

We can estimate the RHS of the identity in Proposition 6.1 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

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

This is consistent for $p(y)$ with growing effective sample size.

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 $E((\hat{p}^{(br)} - p(y))^2)/p(y)^2$ small (expectation is over Monte-Carlo sampling variation of $\hat{p}^{(br)}$).

*Exercise* 6.4. (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). ♣

Unfortunately, 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

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

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

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

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 the 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 6.5.** *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* 6.6. Verify this identity using the same procedure as Exercise 6.2. ♣

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

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

fairly stable and close to the state of the art for simple 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 samples 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.

## 6.2 The Laplace approximation to the marginal likelihood

This is an example of a "fixed" approximation. If the data set is $y = (y_1, \ldots, y_n)$ and $n$ is fixed (as it always is for any given data set), then there is no precision parameter we can change to control the accuracy of our estimate. Take it or leave it you might say. The estimators given in Section 6.1 all have a precision parameter of this sort: the number of MCMC samples $T$ we use to estimate expectations in Section 6.1; as $T$ increases we can drive the error to be as small as we like, subject to making longer and more time-consuming MCMC runs. It does hold that as $n \to \infty$ then, under regularity conditions, the Laplace estimator is asymptotically consistent[11]. The Laplace approximation can be used to approximate the posterior itself, or expectations over the posterior. Here we use it to approximate the marginal likelihood.

### 6.2.1 Laplace approximation

Suppose $\theta$ is a scalar and we have an integral $I_n$ to do of the form

$$I_n = \int_{-\infty}^{\infty} e^{-nh(\theta)} d\theta$$

for some $n \geq 1$ and $h : \mathbb{R} \to \mathbb{R}$ a given function of $\theta$. The integral may be dominated by values of $\theta$ where $h(\theta)$ is small. Let $h^{(k)}$ be the $k$'th derivative, $h^{(k)}(\theta) = \partial^k h(\theta)/\partial\theta^k$. Suppose $\hat{\theta}$ is a minimum so $h^{(1)}(\hat{\theta}) = 0$ and $h^{(2)}(\hat{\theta}) > 0$. Write $\hat{h} = h(\hat{\theta})$ and $\hat{h}^{(k)} = h^{(k)}(\hat{\theta})$ for short and let $\sigma^2 = 1/\hat{h}^{(2)}$.

**Proposition 6.7.** *If $h : \mathbb{R}^p \to \mathbb{R}$ is a given function of the parameter $\theta \in \mathbb{R}^p$ which is (1) $C^\infty$ (all derivatives exist) and (2) at least asymptotically independent of $n$ (so $h$ is $O(1)$ in $n$) and (3) has a unique minimum in $\mathbb{R}^p$ then, when $p = 1$,*

$$\int_{-\infty}^{\infty} e^{-nh(\theta)} d\theta = e^{-n\hat{h}} \sqrt{\frac{2\pi\sigma^2}{n}} \left( 1 + \frac{5[\hat{h}^{(3)}]^2\sigma^6 - 3\hat{h}^{(4)}\sigma^4}{24n} + O(n^{-2}) \right).$$

*The multivariate version $\theta = (\theta_1, ..., \theta_p)$ of this is*

$$\int_{\mathbb{R}^p} e^{-nh(\theta)} d\theta = e^{-nh(\hat{\theta})} (2\pi/n)^{p/2} |\Sigma|^{1/2} \left( 1 + O(1/n) \right)$$

*where*

$$\Sigma^{-1} = \left. \frac{\partial^2 h}{\partial\theta\partial\theta^T} \right|_{\theta=\hat{\theta}}$$

*is the Hessian of $h(\theta)$ at $\hat{\theta}$.*

*Proof.* Here is the plan: expand $h$ as a Taylor series to fourth order about $\hat{\theta}$, pull out the constant and quadratic terms, then further expand the exponential $\exp(f(\theta)) = 1 + f + f^2/2 + ...$ for the terms involving $\hat{h}^{(3)}$ and $\hat{h}^{(4)}$. This leaves normal moments $E_\theta((\theta - \hat{\theta})^k)$ for $k = 4, 6, 8$ which we can look up. I leave the multivariate case to you (just verify the $O(1/n)$ terms).

When we Taylor expand about $\hat{\theta}$ we get factors which we write $\delta = (\theta - \hat{\theta})$ and change the integration variable from $\theta$ to $\delta$. The quadratic term gives us a normal $\delta$ with "variance" $\sigma^2 = 1/\hat{h}^{(2)}$. After Taylor expanding, odd powers of $\delta$ integrate to zero over the normal density. The main thing to

[11]Tierney L, Kadane JB (1986). *Accurate Approximation for Posterior Moments and Marginal Densities*, Journal of the American Statistical Association, v81, pp82-86.

be careful with is tracking the $n$-dependence of terms and gathering everything that is $O(1/n^{3/2})$.

$$I_n = \int_{-\infty}^{\infty} \exp\left(-n\hat{h} - n\hat{h}^{(2)}\delta^2/2 - n\hat{h}^{(3)}\delta^3/6 - n\hat{h}^{(4)}\delta^4/24 - O(n\delta^5) - O(n\delta^6)\right) d\delta$$

$$= e^{-n\hat{h}}\sqrt{\frac{2\pi\sigma^2}{n}} \int_{-\infty}^{\infty} N(\delta; 0, \sigma^2/n) \left[1 - O(n\delta^3) - n\hat{h}^{(4)}\delta^4/24 - O(n\delta^5) - O(n\delta^6)\right.$$

$$\left. + n^2[\hat{h}^{(3)}]^2\delta^6/2 \cdot 36 + O(n^2\delta^7) + O(n^2\delta^8)\right] d\delta.$$

The moments of a normal are

$$E_{X\sim N(0,s^2)}(X^p) = \begin{cases} 0 & \text{if } p \text{ is odd} \\ s^p(p-1)!! & \text{if } p \text{ is even} \end{cases}$$

where $(p-1)!! = (p-1) \times (p-3) \times \cdots \times 3 \times 1$. We have $s = \sigma/\sqrt{n}$ so $E(\delta^4) = 3\sigma^4/n^2$, $E(\delta^6) = 15\sigma^6/n^3$ and $E(\delta^8) = O(1/n^4)$. This gives

$$I_n = e^{-n\hat{h}}\sqrt{\frac{2\pi\sigma^2}{n}} \left[1 - 3n\hat{h}^{(4)}\sigma^4/24n^2 + 15n^2[\hat{h}^{(3)}]^2\sigma^6/72n^3 + O(1/n^2)\right]$$

$$= e^{-n\hat{h}}\sqrt{\frac{2\pi\sigma^2}{n}} \left(1 + \frac{5[\hat{h}^{(3)}]^2\sigma^6 - 3\hat{h}^{(4)}\sigma^4}{24n} + O(n^{-2})\right). \qquad \square$$

### 6.2.2 Approximating the marginal likelihood and deriving the BIC

Suppose (for clarity, for example from de Finetti) the observations are conditionally independent, so $\ell(\theta; y) = \sum_i \ell(\theta; y_i)$, with $\theta \in \mathbb{R}^p$ and $y \in \mathbb{R}^n$. Dropping the model indicator and writing $\pi(\theta)$ and $p(y)$ etc instead of $\pi(\theta|m)$ and $p(y|m)$,

$$p(y) = \int_{\mathbb{R}^p} \exp(-n[-\bar{\ell}(\theta; y) - \log(\pi(\theta))/n]) d\theta.$$

where

$$\bar{\ell}(\theta; y) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; y_i).$$

Evaluate this using the Laplace approximation, with $h(\theta) = -\bar{\ell}(\theta) - \log(\pi(\theta))/n$. Notice that $h$ is $O(1)$ in $n$ when the data are conditionally independent given $\theta$. That's important for correct $n$-dependence of the terms in Proposition 6.7.

Apply Proposition 6.7 (multivariate case). Dropping terms down on the leading term by $O(1/n)$,

$$\int e^{-nh(\theta)} d\theta = e^{-nh(\hat{\theta})}(2\pi/n)^{p/2}|\Sigma|^{1/2}\left(1 + O(1/n)\right)$$

$$p(y) = L(\hat{\theta}; y)\pi(\hat{\theta})(2\pi/n)^{p/2}|\Sigma|^{1/2}\left(1 + O(1/n)\right)$$

where $L(\hat{\theta}; y) = \exp(\ell(\hat{\theta}; y))$ and $\hat{\theta} = \arg\max_\theta \pi(\theta|y)$ is the posterior mode. Here $\Sigma^{-1} \simeq J_0(\hat{\theta})$, the observed unit Fisher information (but evaluated at the posterior mode, $\hat{\theta}$, not the MLE $\hat{\theta}_{MLE}$).

$$\Sigma^{-1} = \left.\frac{\partial^2 h(\theta)}{\partial\theta\partial\theta^T}\right|_{\theta=\hat{\theta}}$$

$$= -\frac{1}{n}\left.\frac{\partial^2 \ell(\theta; y)}{\partial\theta\partial\theta^T}\right|_{\theta=\hat{\theta}} - \frac{1}{n}\left.\frac{\partial^2 \pi(\theta)}{\partial\theta\partial\theta^T}\right|_{\theta=\hat{\theta}}$$

$$\simeq J_0$$

if $\ell(\theta; y) = \sum_i \ell(\theta; y_i)$ is $O(n)$ and $\pi(\theta)$ is $O(1)$. This gives

$$\log(p(y)) \simeq \ell(\hat{\theta}; y) + \log(\pi(\hat{\theta})) + \frac{p}{2} \log(2\pi/n) - \frac{1}{2} \log(|J_0|) \tag{6.1}$$

$$\simeq \ell(\hat{\theta}; y) - \frac{p}{2} \log(n) + O(1) \tag{6.2}$$

$$\text{BIC} = -2\ell(\hat{\theta}_{MLE}; y) + p \log(n) \tag{6.3}$$

Equation (6.1) is often reasonably accurate at large $n$. It is implemented in many software packages including the one used in the example below.

The BIC in (6.3) is $-2$ times the $O(1)$ approximation to the log marginal likelihood in (6.2). It is evaluated at the MLE, which is approximately equal to the posterior mode $\hat{\theta}$ at large $n$. The BIC is a zeroth order approximation in $O(1/n)$ as it drops model-dependent constant terms (like $J_0$) in (6.1) not just terms which go to zero! The model with the least BIC score has the largest marginal likelihood, approximately.[12] In "The Bayesian Choice", Christian Robert notes that the BIC isn't Bayesian, as the prior doesn't enter. This may have been a virtue for its inventor[13], as it is widely used in Frequentist inference. Expect it to be useful at large $n$ only.

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

Recall the O-ring data from Section 1.3.2 and 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 model selection 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 $x_i \in R$ is the 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 $\beta_1, \beta_2 \sim N(0, 3^2)$ from Section 1.3.2. The marginal likelihoods are

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

The estimators above are computed from samples generated by MCMC in Estimating-ML-BF.R (code online). The resulting estimates of Bayes Factors: 2.7 (naive - pretty good!) 1.9 (harmonic) 2.8 (bridge $\hat{B}_{1,2}^{(br)}$ from Proposition 6.5) 2.7 (Laplace-approximation estimator from package implementing Equation 6.1). We report $\hat{B}_{1,2} = 2.75$, weak evidence "barely worth mentioning" for Logit over Probit.

---

[12]It isn't part of the course, but under a unit information prior on $\theta$ the terms $O(1)$ also cancel and it becomes $O(1/\sqrt{n})$ approximation.

[13]Schwarz, G.E. (1978), "Estimating the dimension of a model", Annals of Statistics, 6 (2): 461–464

# 7 Likelihood-free methods: Approximate Bayesian Computation

## 7.1 Lecture 9: Motivation and Definitions

### 7.1.1 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 7.1.** 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. $\diamondsuit$

**Definition 7.2.** (*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 *doubly intractable*. $\diamondsuit$

*Remark* 7.3. 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. 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 $\theta$ and some random variables $Z$ 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 $p(y|\theta)$ is and how it depends on $\theta$. This appears in some physics-based models including some climate models. ✠

*Remark* 7.4. 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* 7.5. (*Ising model*) Denote by $\mathcal{Y} = \{0, 1\}^{m^2}$ the set of all binary $m \times m$ "images" $y = (y_1, y_2, ..., y_{m^2})$, $y_i \in \{0, 1\}$, where $i = 1, 2, ..., m^2$ 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 $3 \times 3$ square lattice and (Right) realisation of an Ising model.

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

$$\#y = \frac{1}{2} \sum_{i=1}^{m^2} \sum_{j \sim i} \mathbb{I}_{y_j \neq y_i},$$

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 $y \in \mathcal{Y}$. In the example in Table 1 we have $\#y = 4$.

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

$$p(y|\theta) = \exp(-\theta \# y)/c(\theta). \tag{7.1}$$

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 can't compute for $n$ large. There are $2^{m^2}$ terms in the sum, and no one can solve it (it is an important model in physics, so many have looked at this). [14]

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

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

$$= \min\left\{1, \frac{c(\theta)}{c(\theta')} \ \times \ \exp((\theta - \theta')\#y) \ \times \ \frac{\pi(\theta')}{\pi(\theta)}\right\}$$

and $p(y)$ cancels but $c(\theta)/c(\theta')$ doesn't, and so we can't evaluate the acceptance probability. ♠

### 7.1.2 The ABC posterior

*Exercise* 7.6. (repeat Exercise 1.2) Suppose the data $Y$ and the parameter $\Theta$ 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 $\Theta'$ 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

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

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

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

(events in each loop are independent)

$$= \sum_{n=1}^{\infty} \left[\prod_{i=1}^{n-1} \Pr(y_i \neq y_{obs})\right] \Pr(\theta_n = \theta, y_n = y_{obs})$$

so we have a factor $\Pr(y_i \neq y_{obs}) = 1 - p(y_{obs})$ for each rejection, with $p(y_{obs}) = \sum_\theta \pi(\theta)p(y_{obs}|\theta)$, and we draw $\theta_n = \theta, \ y_n = y_{obs}$ at step $n$ with probability $\pi(\theta)p(y_{obs}|\theta)$,

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

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

---

[14]There is a formula for $c(\theta)$ for Periodic Boundary Conditions, but PBC don't see much use in statistical modelling.

The idea in ABC is to relax the requirement that $y_n = y_{obs}$ and stop if the simulated data set $y_n$ is "close" to $y_{obs}$. We like $\theta$-values with a large likelihood value $p(y_{obs}|\theta)$ so might like $\theta$ 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 7.7.** (*simulation of $(y, \theta)$ 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 $\theta \sim \pi(\theta)$. ♡

What does it mean for $y$ to be close to $y_{obs}$? We need to define the distance between two data sets.

**Definition 7.8.** (*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 $D(s, s')$. ◇

*Remark* 7.9. 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 $\theta$, 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. However, according to the Pitman-Koopman-Darmois theorem, for independent identically distributed data points from a model not belonging to the exponential family, the dimension of any sufficient statistic remains unbounded and keeps increasing with the sample size, so this dimension reduction generally comes at the price of loss of data-information. ✠

*Remark* 7.10. To simulate $(S, \theta)$-pairs, simulate $\theta \sim \pi(\cdot)$, $Y \sim p(\cdot|\theta)$ and evaluate $S(Y)$. ✠

**Definition 7.11.** Let $\Delta_\delta(y_{obs})$ be a "ball" of radius $\delta$ 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 if $p(\Delta_\delta(y_{obs})|\theta) \equiv \Pr(Y \in \Delta_\delta(y_{obs})|\theta)$ we have

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

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 $p(\Delta_\delta(y_{obs})) \equiv \Pr(Y \in \Delta_\delta(y_{obs}))$ is

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

**Definition 7.12.** (*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* 7.13. 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 7.14.** *The ABC posterior can be written the following form,*

$$\pi_{ABC}(\theta|y_{obs}) = \int_{\Delta_\delta(y_{obs})} \pi(\theta|y)p(y|Y \in \Delta_\delta(y_{obs}))dy, \qquad (7.2)$$

*where*

$$p(y|Y \in \Delta_\delta(y_{obs})) = \frac{p(y)\mathbb{I}_{y \in \Delta_\delta(y_{obs})}}{p(\Delta_\delta(y_{obs}))}. \qquad (7.3)$$

*Proof.* Substituting Equation 7.3 in the RHS of Equation 7.2,

$$\int_{\Delta_\delta(y_{obs})} \pi(\theta|y)p(y|Y \in \Delta_\delta(y_{obs}))dy = \int_{\Delta_\delta(y_{obs})} \frac{\pi(\theta|y)p(y)}{p(\Delta_\delta(y_{obs}))}dy$$

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

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

Proposition 7.14 helps us interpret the ABC-posterior. It is what we get when we average the regular posterior $\pi(\theta|y)$ over $y \in \Delta_\delta(y_{obs})$, so averaging over $y$-values around $y_{obs}$, weighted by the prior predictive $p(y)$ in $\Delta_\delta(y_{obs})$. ABC assumes the exact posterior $\pi(\theta|y)$ doesn't change much as $y$ varies over $\Delta_\delta(y_{obs})$, so the average is approximately equal to $\pi(\theta|y_{obs})$.

## 7.2 Computational methods for ABC

### 7.2.1 Simulating the ABC posterior via rejection

We wouldn't make an approximation unless it helped us somehow. The point is that, if $\delta$ 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 7.15.** *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 $\Theta_{ABC} = \theta_n$ and $N = n$ else goto step 2.*

*The ABC-Rejection Algorithm returns samples $(\Theta_{ABC}, N)$ with $\Theta_{ABC} \sim \pi_{ABC}(\cdot|y_{obs})$.*

*Remark* 7.16. We introduced a new random variable $\Theta_{ABC}$. 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 a set $A \subseteq \Omega$, $\Pr(\Theta_{ABC} \in A)$ is the long-run proportion of times the output satisfies $\Theta_{ABC} \in A$. We would like to show that $\Pr(\Theta_{ABC} \in A) = \pi_{ABC}(A|y_{obs})$.

Each loop is an independent trial which succeeds with probability $\Pr(Y \in \Delta_\delta(y_{obs})) = \int_{\Delta_\delta(y_{obs})} p(y)dy$ so $N \sim Geom(p(\Delta_\delta(y_{obs})))$. We get $\Theta_{ABC} \in A$ and $N = n$ iff $y_m \notin \Delta_\delta(y_{obs})$ 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

$$\Pr(\Theta_{ABC} \in A) = \sum_{n=1}^{\infty} \Pr(y_1 \notin \Delta_\delta(y_{obs}), ..., y_{n-1} \notin \Delta_\delta(y_{obs}), (\theta_n, y_n) \in A \times \Delta_\delta(y_{obs}))$$

(events in each loop are independent and $\Pr(y_n \notin \Delta_\delta(y_{obs})) = 1 - p(\Delta_\delta(y_{obs}))$)

$$= \sum_{n=1}^{\infty} (1 - p(\Delta_\delta(y_{obs})))^{n-1} \Pr((\theta_n, y_n) \in A \times \Delta_\delta(y_{obs})).$$

Here $\Pr((\theta_n, y_n) \in A \times \Delta_\delta(y_{obs})) = \int_A \int_{\Delta_\delta(y_{obs})} \pi(\theta) p(y|\theta) dy d\theta$ doesn't depend on $n$ so

$$\Pr(\Theta_{ABC} \in A) = \frac{\int_A \pi(\theta) p(\Delta_\delta(y_{obs})|\theta) d\theta}{p(\Delta_\delta(y_{obs}))}$$

$$= \int_A \pi_{ABC}(\theta|y) d\theta \qquad \square$$

*Example* 7.17. Here is a very simple example in which the data are five samples from a Poisson with mean $\lambda$ and we have a Gamma-prior for $\lambda$.

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

Data space $\mathcal{Y} = \{0, 1, 2, ..., \}^n$ and prior $\lambda \sim \Gamma(\alpha = 1, \beta = 1)$.

Summary statistic: 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 7.15 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 $\lambda$ and stop, otherwise goto (1).

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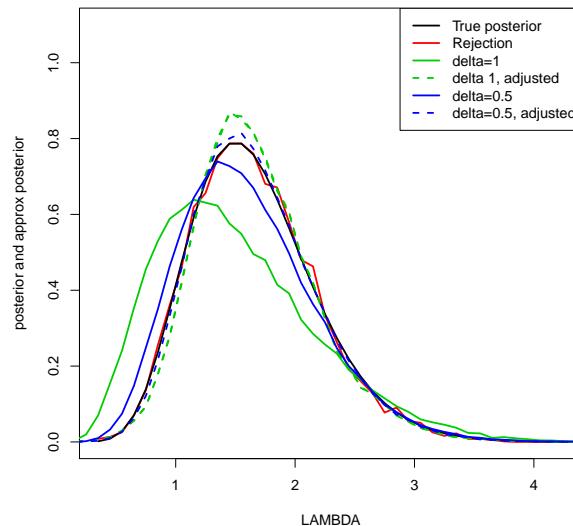


Figure 11: ABC-poisson example: (black curve) exact posterior; (red curve) rejection algorithm in Exercise 7.6; (green curve) samples from $\pi_{ABC}$ with $\delta = 1$; (blue curve) samples from $\pi_{ABC}$ with $\delta = 0.5$; (dashed curves) Regression adjusted samples. For example code see ABC-Lecture.R.

We can estimate expectations, plot histograms and calculate an HPD set using these samples. Figure 11 shows density plots of samples. When we make $\delta$ 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. ♠

*Remark* 7.18. There is a slightly different ABC sampler which is often convenient. Instead of fixing $\delta$ and running ABC-rejection, we simulate $\theta_t \sim \pi(\cdot), y_t \sim p(\cdot|\theta^{(t)}), t = 1, ..., T$, evaluate $d_t = d(S(y_t), s_{obs})$ and sort $d_t$ smallest to largest to get $(\theta_{(t)}, d_{(t)})$. If we want to keep a fraction

$\alpha \in (0,1]$ of the samples we set $\tau = \lfloor \alpha T \rfloor$ and $\delta = d_{(\tau)}$ and keep $\theta_{(t)} : d_{(t)} < \delta$. This still has the right distribution as conditioning a random variable to be less than the value of one of its order statistics is the same as conditioning the variable to be less than that value. ✠

### 7.2.2 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 $\theta^{(adj)} \sim \pi(\cdot|y_{obs})$. This is not achievable in general. We will set out some assumptions 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 $\theta \in \mathbb{R}$ to simplify notation (in regression with a multivariate response, $\beta$ below is a matrix). The idea extends easily to $\theta \in \mathcal{R}^p$.

Let $\theta \sim \pi(\cdot|y)$ and $\theta' \sim \pi(\cdot|y_{obs})$. Let $s = S(y)$ and $s_{obs} = S(y_{obs})$.

**Assumption 7.19.** The posterior mean $\mu(S(y)) = E(\theta|Y = y)$ is a linear function of $S(y)$ alone, that is, for some $\beta \in R^p$,

$$(7.4) \qquad \mu(s) = \mu(s_{obs}) + (s - s_{obs})^T \beta. \qquad \heartsuit$$

**Assumption 7.20.** The random variables $\theta - \mu(s)$ and $\theta' - \mu(s_{obs})$ are identically distributed. $\heartsuit$

These assumptions are obviously very strong. However, for $s$ such that $D(s, s_{obs}) \leq \delta$ and $\mu(s)$ a smooth function, Eqn. 7.4 is a local linear approximation good at small $\delta$. If these conditions do hold then the regression-adjusted parameter

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

is distributed according to the posterior.

**Proposition 7.21.** *If Assumptions 7.19 and 7.20 hold then $\theta_{adj} \sim \pi(\cdot|y_{obs})$.*

*Proof.* By Assumption 7.20 we have $\theta \sim \theta' + \mu(s) - \mu(s_{obs})$, so subtracting $(s - s_{obs})^T \beta$,

$$\theta - (s - s_{obs})^T \beta \sim \theta' + \mu(s) - \mu(s_{obs}) - (s - s_{obs})^T \beta$$

with $\mu(s) - \mu(s_{obs}) - (s - s_{obs})^T \beta = 0$ by Assumption 7.19 and hence $\theta - (s - s_{obs})^T \beta \sim \theta'$. $\square$

We can use Proposition 7.21 to motivate a regression adjustment. The regression correction adjusts the distribution of $\theta$ at $y$ to move it onto the distribution of $\theta$ at $y_{obs}$. Let $\epsilon \sim \theta - \mu(s)$ be the random difference to the "local" mean (ie, at $s$). Then

$$\theta = \mu(s) + \epsilon$$
$$= \mu(s_{obs}) + (s - s_{obs})^T \beta + \epsilon$$

Now suppose we run ABC at fixed $\delta$ and get pairs of samples $(\theta^{(t)}, y^{(t)})_{t=1}^T$ with $y^{(t)} \in \Delta_\delta(y_{obs})$. Let $s^{(t)} = S(y^{(t)})$. If we regress $\theta^{(t)}$ on $(s^{(t)} - s_{obs})$ setting

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha, \beta} \sum_{t=1}^T (\alpha + (s^{(t)} - s_{obs})^T \beta - \theta^{(t)})^2$$

then the intercept $\hat{\alpha}$ estimates $\mu(s_{obs})$ and we can use the estimated effects $\hat{\beta}$ to make an adjustment,

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

This gives $\hat{\theta}_{adj}^{(t)} \sim \pi(\cdot|y_{obs})$ (approximately, as the assumptions will only be good for $s$ close to $s_{obs}$, and anyway, we only have $\hat{\beta}$).
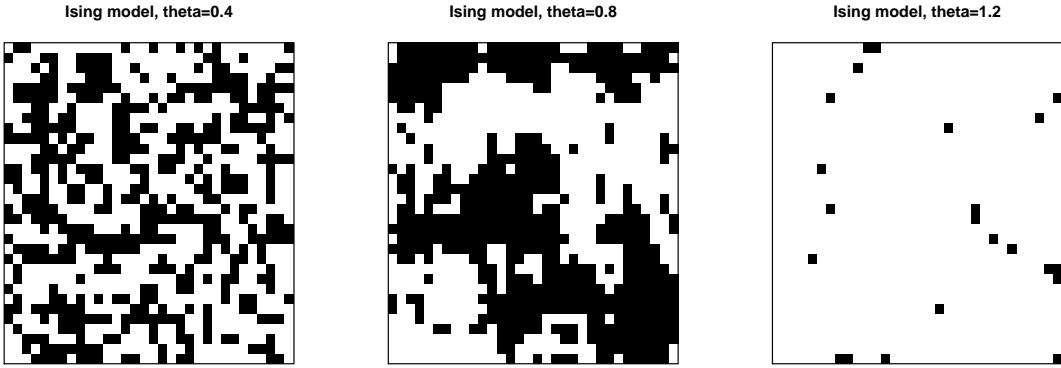
Figure 12: Three samples from the Ising model at different $\theta$-values (see image titles).

*Example* 7.22. We did this for the Poisson-Gamma posterior in Example 7.17. The code for this is available on github. The improvement after regression adjustment is significant. Referring to Figure 11 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* 7.23. One of the main reasons for doing regression adjustment is that it allows us to take $\delta$ quite large and fix the poor approximation using the regression adjustment. Some experimentation is usually necessary (to choose $\delta$ small enough so the approximation doesnt change significantly when we make it smaller). We like to take $\delta$ large as we accumulate samples more rapidly (in Proposition 7.15, we throw out samples if $D(S(y), S(y_{obs})) > \delta$). ✠

## 7.3   ABC example: the Ising Model

The Ising observation model $Y \sim p(\cdot|\theta)$ in Eqn. 7.1 is easy to sample for moderate $m$-values using MCMC. Figure 12 shows samples simulated at three values of $\theta$. These samples are not exactly distributed according to $p(y|\theta)$ as the MCMC won't have converged exactly to the target. However, we can make this error small by taking a long MCMC run and checking convergence.

*Remark* 7.24. Here is an MCMC algorithm simulating a Markov Chain $\{X_k\}_{k=1\ldots K}$ which targets $p(\cdot|\theta)$ in Eqn. 7.1. We will need to run this algorithm to convergence *at each value of $n$* in the ABC algorithm in Proposition 7.15. Our simulated data will be $y_n = X_K$, the last state we simulated, since that gives $X_K \sim p(\cdot|\theta_n)$ approximately at large $K$. Notice that the intractable normalising constant (as a function of $\theta$) isn't a problem as this MCMC samples $y$ not $\theta$.

First we give a proposal distribution. Suppose the current state is $X_k = x$ with $x \in \{0,1\}^{m^2}$. Choose a cell $i \sim U\{1, 2, ..., m^2\}$. Set $x'_i = 1 - x_i$ and $x'_j = x_j$ for $j \neq i$. Notice that $q(x'|x) = q(x|x') = 1/m^2$ for $x', x$ differing at exactly one cell.

Here is the MCMC algorithm itself. If $X_k = x$ then $X_{k+1}$ is determined in the following way.

1. Simulate $x' \sim q(x'|x)$ as above, and $u \sim U(0,1)$.
2. If $u \leq \alpha(x'|x)$ with
$$\alpha(x'|x) = \min\left\{1, \ \exp(-\theta(\#x' - \#x))\right\}$$
   set $X_{k+1} = x'$ and otherwise set $X_{k+1} = x$.

The Markov chain is irreducible ($q$ is irreducible and $\alpha$ is never zero) and aperiodic (rejection is possible), so it is ergodic for $p(y|\theta)$. See ABC-Lecture.R for an implementation. Some samples produced using this code are shown in Figure 12. ✠
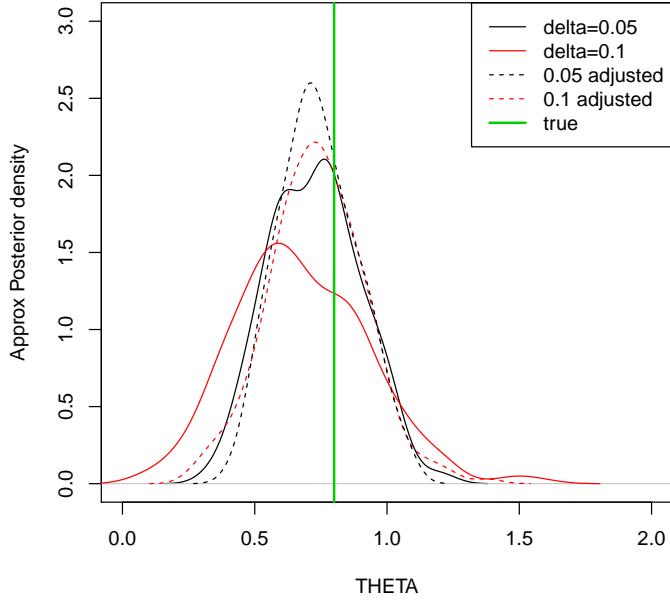
Figure 13: Density plots for samples from $\pi_{ABC}$ varying $\delta$ and applying regression adjustment. The vertical line is the true $\theta$-value in this synthetic-data example.

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 $\delta$ is on a scale of $O(1)$.

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

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

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

Summary statistic: if $y = (y_1, ..., y_{m^2})$ 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 7.15 for this case.

1. Simulate $\theta \sim \text{Exp}(1)$ and $y \sim \text{Ising}(\theta)$ (using MCMC in Remark 7.24).
2. If $|\#y - \#y_{obs}| < M\delta$ return $\theta$ and stop, otherwise goto (1).

Run this algorithm $T$ times returning $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(T)}$. These are samples from $\pi_{ABC}(\theta|y_{obs})$. See ABC-Lecture.R for an implementation. Results are shown in Figure 13.

The distribution converges to the true posterior as $\delta \to 0$ (mathematically, not from the figure). 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 $\delta$-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 $\delta = 0.05$.

# 8 Model averaging

We can allow for uncertainty in which model is the right model when we estimate a function $h(\theta)$ of the parameters $\theta$. 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!

## 8.1 Lecture 10: Model averaging distributions and decisions

### 8.1.1 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 $m \in \mathcal{M}$. Let $\pi(\theta|m)$ be the prior for $\theta$ when the unknown true model $M$ is model $m$, let $p(y|\theta, m)$ be the observation model for $y$ given $\theta$, let

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

be the marginal likelihood under model $M = m$, let

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

be the posterior for $\theta$ given $M = m$ and let

$$\pi(m|y) = \frac{p(y|m)\pi_M(m)}{p(y)} \tag{8.1}$$

be the posterior probability if the true model is $m$, where $\pi_M(m) = \Pr(M = m)$ is the prior probability for $M$ to equal $m$. The model-averaged marginal likelihood is

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

When we speak of the "true model" we mean nature's true generative model for the data.

Suppose we are interested in the expectation $E_{\theta|y}(h(\theta))$. Instead of selecting a model, $M = \hat{m}$ say, and estimating $E_{\theta|y,\hat{m}}(h(\theta))$ in that model, we integrate over the model uncertainty and estimate $E_{\theta,m|y}(h(\theta))$. We now define the joint distribution of $\Theta, M|y$.

**Definition 8.1.** The extended parameter space including the model index is

$$\Omega^* = \bigcup_{m \in \mathcal{M}} \bigcup_{\theta \in \Omega_m} \{(\theta, m)\}$$

The joint posterior distribution for the model and parameter is

$$\tag{8.2} \pi(\theta, m|y) = \pi(\theta|y, m)\pi(m|y), \ (\theta, m) \in \Omega^*. \qquad \diamondsuit$$

**Definition 8.2.** The model-averaged posterior[15] for the parameter is

$$\pi(\theta|y) = \sum_{m \in \mathcal{M}} \pi(\theta, m|y), \ \theta \in \Omega \tag{8.3}$$

where now $\theta$ could come from any one of the model spaces $\Omega_m, \ m \in \mathcal{M}$ so

$$\Omega = \bigcup_{m \in \mathcal{M}} \Omega_m. \qquad \diamondsuit$$

---

[15]here I take $\pi(\theta, m|y) = 0$ if $\theta \notin \Omega_m$

*Example* 8.3. If $\theta = (\theta_1, \dots \theta_m)$ is a realisation of a Poisson process in $[0, T]$ (a "point pattern" say, see Section 1.5.2) with a random number of points $M = m$ then the model index is the parameter dimension, $\mathcal{M} = \{0, 1, 2 \dots\}$ and $\Omega_m = \{\theta \in [0, T]^m : 0 < \theta_1 < \dots < \theta_m < T\}$. The space of all point patterns $\theta$ with any number of points is $\Omega$ and $\Omega^*$ pairs each point pattern in $\Omega$ with the number of points in the pattern, so $\Omega^*$ is all pairs $(\theta, m)$ with $\theta = (\theta_1, \dots, \theta_m)$. ♠

*Remark* 8.4. Write the model averaged posterior in terms of the marginal likelihoods, by substituting Equation 8.2 into Equation 8.3 and using Equation 8.1 to expand $\pi(m|y)$. This gives

$$\pi(\theta|y) \propto \sum_{m \in \mathcal{M}} \pi(\theta|y, m) p(y|m) \pi_M(m). \tag{8.4}$$

Terms in the sum with larger $p(y|m)$-values get a higher weight. The model-averaged posterior puts more weight on models which give a larger prior predictive probability for the data. ✠

*Remark* 8.5. For computation we typically start from the joint distribution in the form

$$\pi(\theta, m|y) \propto p(y|\theta, m) \pi(\theta|m) \pi(m), \qquad (\theta, m) \in \Omega^*, \tag{8.5}$$

since everything here is usually tractable. The normalising constant is $p(y)$ above. ✠

*Example* 8.6. *(averaging over link functions, see Model-Averaging-Lecture.R for an implementation.)* Recall our two models for the Challenger O-ring data,

$$y_i \sim \text{Bernoulli}(\mu_m(\beta_1 + \beta_2 x_i)) \; i = 1, \dots, n,$$

with $x_i$ scaled temperature and a choice $\mu_m(\beta_1 + \beta_2 x_i)$, $m = 1, 2$ of link functions with $\mu_1$ the logit and $\mu_2$ the probit link. We will estimate the model averaged posterior $\pi(\beta|y)$ (and plot it for $\beta_1$). We are 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 8.3

$$\pi(\beta|y) = \pi(\beta|m = 1, y)\,\pi(m = 1|y) + \pi(\beta|m = 2, y)\,\pi(m = 2|y).$$

We sum over $m$ in Equation 8.3 rather than use Equation 8.5 (which is usually easiest) as we already calculated the Bayes Factor $B_{1,2}$ comparing the same models in Section 6.3, and got $B_{1,2} = 2.75$. Solving $B_{1,2} = \pi(m = 1|y)/(1 - \pi(m = 1|y))$ (model priors cancel) for $\pi(m = 1|y)$ we get

$$\pi(\beta|y) = \pi(\beta|m = 1, y)\frac{B_{1,2}}{1 + B_{1,2}} + \pi(\beta|m = 1, y)\frac{1}{1 + B_{1,2}}.$$

Posterior densities for $\beta_1|y$ estimated for each model using MCMC are illustrated in Figure 14. The model-averaged posterior is slightly weighted towards Model 1 (Logistic). ♠

### 8.1.2 Model averaging is preferred to inference after model selection

Suppose we want to estimate the value of some function $h(\theta)$ and our loss is the squared difference to the true value. Here $h : \Omega \to \mathbb{R}$ so $h$ is defined on each of the spaces $\Omega_m$, $m \in \mathcal{M}$. The coherent action under the squared error loss is to report the model averaged posterior mean, $E_{\theta,m|y}(h(\theta))$.

The model-averaged posterior mean is

$$E_{\theta,m|y}(h(\theta)) = \sum_{m \in \mathcal{M}} \int_{\Omega_m} h(\theta)\pi(\theta, m|y)\, d\theta.$$
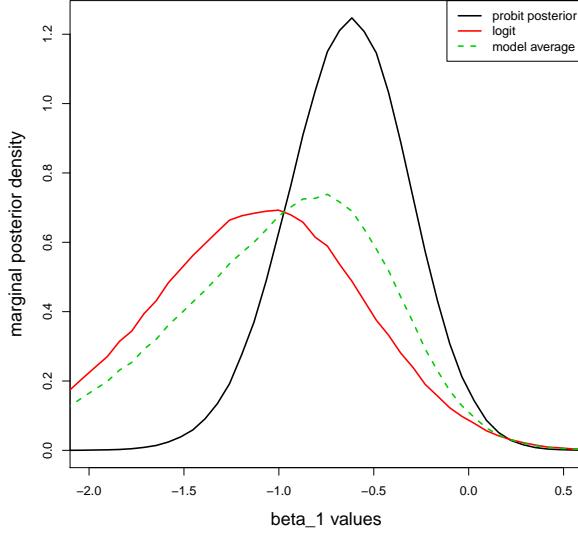
Figure 14: Posterior under logit (red) and probit (black) link functions and model-averaged posterior (green dashed) for Bernoulli-GLM regression of the binary O-ring data.

Compare this with the single-model posterior mean at some selected model $\hat{m}$ say,

$$E_{\theta|y,\hat{m}}(h(\theta)) = \int_{\Omega_{\hat{m}}} h(\theta)\pi(\theta|y,\hat{m})\,d\theta.$$

If we do inference after model selection and report the posterior mean for the model $M = \hat{m}$ we get a higher Bayes Risk.

**Proposition 8.7.** *If the loss for estimation using $\delta(y)$ when the truth is $h$ is $(\delta(y) - h)^2$ then the Bayes risk $\rho(\pi, \delta)$ 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

$$\rho(\pi, \delta|y) = \sum_{m \in \mathcal{M}} \int_{\Omega_m} (\delta(y) - h(\theta))^2 \pi(\theta, m|y)\,d\theta.$$

The calculation is the same as it was for the original case without model averaging, as $m$ is just another parameter,

$$\frac{\partial \rho}{\partial \delta} = \sum_{m \in \mathcal{M}} \int_{\Omega_m} (2\delta - 2h(\theta))\pi(\theta, m|y)\,d\theta$$

and this is zero and when $\delta(y) = E_{\theta,m|y}(h(\theta))$. This is the unique minimum value for $\delta$ at $y$ and so any other estimator cant be better. $\qquad\square$

## 8.2 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. Spike-and-slab priors are 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.

Page 81

### 8.2.1 Spike and slab priors for regression

Consider model averaging in a regression problem with the setup,

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

with $z = (z_1, ..., z_p)$ a vector of binary indicator variables, and $X$ an $n \times p$ matrix of continuous covariates with a first column of ones corresponding to the intercept. 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 of $Y$ on the second covariate, with $E(Y_i) = \theta_1 + \theta_2 X_{i,2}, \ i = 1, \ldots, n$.

There are $2^p$ models and our model index $m \in \mathcal{M}$ is replaced by $z \in \mathcal{Z}$, which takes values in $\mathcal{Z} = \{0, 1\}^p$. The joint posterior in the form given by Equation 8.5 is

$$\pi(\theta, \sigma, z|y) \propto p(y|\theta, \sigma, z)\pi(\theta)\pi_s(\sigma)\pi_Z(z), \tag{8.6}$$

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 $\theta$ 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* 8.8. This $\theta$-prior is called a "spike and slab" prior. Suppose

$$\pi(\theta, \sigma, z) = \pi_\sigma(\sigma) \prod_{i=1}^{p} \pi(\theta_i)\pi_Z(z_i)$$

(all independent a priori). The parameters appearing in the regression are $\theta_z$. 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 $w = 1 - \pi_Z(1)$ give the prior probability $z_i = 0$. We have

$$\Pr(\tilde{\theta}_i \leq c) = \Pr(z_i = 0)\Pr(\theta_i z_i \leq c|z_i = 0) \ + \ \Pr(z_i = 1)\Pr(\theta_i z_i \leq c|z_i = 1)$$

$$= w\mathbb{I}_{c \geq 0} + (1 - w)\int_{-\infty}^{c} \pi(t)dt$$

We get the prior density by differentiating with respect to $c$,

$$\pi_{\tilde{\Theta}}(\tilde{\theta}_i) = w\delta_0(\tilde{\theta}_i) + (1 - w)\pi_i(\tilde{\theta}_i),$$

in terms of the delta-function notation introduced in Chapter 3, Definition 3.13. The prior density for the regression parameters $\tilde{\theta} = \theta_z$ is "spike" plus "slab". ✠

*Remark* 8.9. Normally when we consider different models in regression, with different subsets of effects, the dimension of the parameter space $\Omega_z$ 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 $\Omega_z = \mathbb{R}^p \times \mathbb{R}^+$ and this makes it easy (in principle) to analyse using our standard MCMC tools. If $z_i = 0$ then the posterior distribution for $\theta_i$ is just given by its prior, as the likelihood doesn't depend on $\theta_i$ when $z_i = 0$.

In Section 8.3 we give the equivalent parameterisation and posterior if we choose to just keep the "selected" variables $\theta_i : z_i = 1$ in the probability distribution. That's easy as the "unselected" variables don't enter the likelihood so can be integrated out. ✠

### 8.2.2 Model Averaged Regression of `swiss` data

We illustrate model-averaged regression on the `swiss` dataset in the `MASS` package of R. After some routine transformations of the covariates (recorded as percentages so logit) and response (centered and scaled to unit variance), and removing two outliers, a standard regression

```
> sw.lm<-lm(Fertility~Infant.Mortality+Examination+Education+Catholic+Agriculture,
+           data=sw)
> summary(sw.lm)
...
Coefficients:
                 Estimate Std. Err ... Pr(>|t|)
(Intercept)        0.99    0.99     ... 0.324
Infant.Mortality   2.16    0.66     ... 0.002 **
Examination       -0.54    0.28     ... 0.061 .
Education         -0.48    0.20     ... 0.024 *
Catholic           0.08    0.05     ... 0.095 .
Agriculture       -0.23    0.14     ... 0.116
```

suggests unsurprisingly that `Infant.Mortality` is informative for fertility. Other effects look plausible. The model is

$$Y_i = \sum_{j=1}^{p} z_j \theta_j x_{i,j} \; + \; \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$ and $p = 6$. In this model $z_i \in \{0, 1\}$ controls the effect of covariate $j = 1, \dots, 6$.

*Priors:* For the sake of example $\theta_i \sim t(df = 3)$ and $\sigma \sim 1/\sigma$. 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 $p = 6$ 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.

*Posterior:* The joint distribution of the model index $z$ and parameters $\theta, \sigma$ is, from Equation 8.6,

$$\pi(\theta, z, \sigma | y) \propto N(y; X\theta_z, \sigma^2 I_n) \times t(\theta; df) \times \sigma^{-1} \times \xi^{|z|} (1 - \xi)^{p - |z|}.$$

*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 $\theta$-update, a $z$-update and a $\sigma$-update sequentially.

$\theta$-update: fix $a > 0$. At each step choose $i \sim U\{1, 2, \dots, p\}$ and simulate a proposal

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

The new regression parameters are

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

The acceptance probability is

$$\alpha_{\theta_i}(\theta_z' | \theta_z) = \min \left\{ 1, \frac{N(y; X\theta_z', \sigma^2 I_n) \, t(\theta_i'; df)}{N(y; X\theta_z, \sigma^2 I_n) \, t(\theta_i; df)} \right\}.$$

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

$$\alpha_{z_i}(\theta_{z'}|\theta_z) = \min\left\{1, \frac{N(y; X\theta_{z'}, \sigma^2 I_n)\xi^{z_i'}(1-\xi)^{1-z_i'}}{N(y; X\theta_z, \sigma^2 I_n)\xi^{z_i}(1-\xi)^{1-z_i}}\right\}.$$

$\sigma$-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 $\theta$-update.

The analysis is implemented in Model-Averaging-Lecture.R. We ran it and generated samples

$$(\theta^{(t)}, z^{(t)}, \sigma^{(t)}) \sim \pi(\theta, z, \sigma|y) \quad t = 1, 2, ..., T.$$

The run length was $T = 10^7$ sub-sampled each 5000 steps for plotting. The effective sample sizes (ESS values) for all parameters are all at least one thousand. The MCMC traces for $\theta_z^{(t)}$ are shown in Figure 15 (Left). The traces show the model variation. Notice that when $z_i = 0$, $\theta_{z,i} = 0$. There
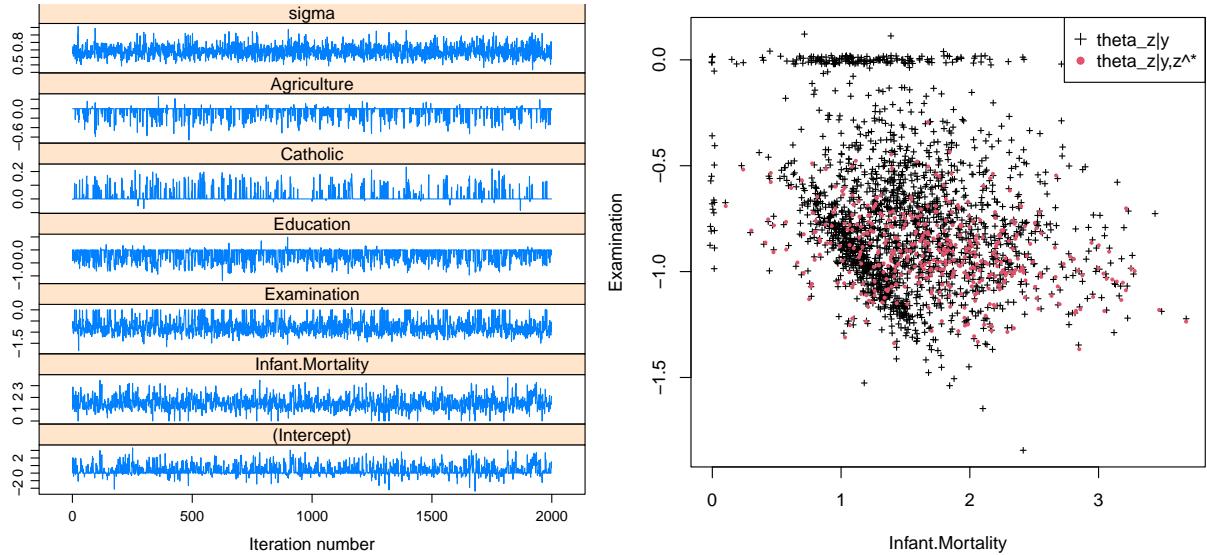


Figure 15: (Left) MCMC traces for each of the parameters (six effects and $\sigma$). (Right) scatter plot of samples from the posterior showing the different distributions of $(\theta_2, \theta_3)$, the effects for `Infant.Mortality` and `Examination`, in the model averaged posterior $\pi(\theta|y)$ ("+" signs) and the posterior $\pi(\theta|y, z^*)$ (red dots) conditioned on the most probable model $z^* = (1, 1, 1, 0, 0, 0)$. This model has an intercept and the two plotted effects.

is little evidence in the data for an effect due to `Catholic`, so $z_5 = 0$ frequently and we see $\theta_{z,5}$ sits on zero in most 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

Page 84

```
111000 011000 111100 011100 011101 111101 011001 111001 111110 011111
  21.8   17.3   13.1   12.9    5.4    4.3    3.8    3.4    1.9    1.8
110110 010110 010111 011110 111111 111010 110111 011010 010100 110100
   1.8    1.7    1.6    1.6    1.5    1.4    1.1    1.0    0.5    0.5
101000 111011 011011 000010 010101 100110 101100 110101 000000 000100
   0.3    0.3    0.2    0.1    0.1    0.1    0.1    0.1    0.0    0.0
```

The top ranked (ie MAP) model includes the intercept, `Infant.Mortality` and `Examination`. However (as the response was centred and scaled) the evidence for the intercept is equivocal. This resembles the evidence from $p$-values in the `lm()` fit. The marginal probabilities $\pi(z_i = 1|y)$ are

```
(Intercept) Infant.Mortality  Examination  Education  Catholic  Agriculture
0.52          0.99              0.92          0.50      0.16      0.24
```

There is clear evidence for an effect on fertility due to infant mortality but having sat a higher exam (not just having basic education) is also informative.

Figure 15 (Right) shows how the posterior changes when we condition on the MAP model. The red points give the posterior conditioned on the MAP model while the black + signs show the more diffuse posterior allowing for uncertainty in the model. The conditioned model misses out the concentrated mass at the top (when $z_3 = 0$, so no `Examination` effect) and lower left running by $(1, -1)$ (when $z_1 = 0$, so no intercept), which is fine, if we trust the MAP as a model estimator. The model averaged posterior gives a better quantification of the uncertainty in downstream inference.

## 8.3   Appendix A: Variable dimension parameterisations for regression

Let's see how it looks if we don't include the unused $\theta$-components in the spike and slab regression model in Section 8.2.1. Let $m = (i \in [p] : z_i = 1)$ be the ordered list of indices for the covariate-effects $\theta_i : z_i = 1$ that we include in $X\theta_z$, so $m = (i_1, ..., i_k)$ when $\sum_i z_i = k$. The space of models is $m \in \mathcal{M}$ where $\mathcal{M}$ is the set of all subsets of $[p] = \{1, ..., p\}$.

Introduce a new parameter vector $\phi$ replacing the selected variables $(\theta_{i_1}, \ldots, \theta_{i_k})$. When there are $|m| = k$ covariates in the model, $\phi = (\phi_{i_1}, \ldots, \phi_{i_k})$ and $\phi \in \mathbb{R}^k$. Let $X_{:,m}$ be the matrix we get by taking the columns $X_i$ of $X$ corresponding to $z_i = 1$, so $X_{:,m} = [X_i]_{i \in m}$. In this notation $X_{:,m}\phi = X\theta_z$. The prior for $m$ is $\pi_M(m) = \pi_Z(z(m))$ where $z_i(m) = \mathbb{I}_{i \in m}$, $i = 1, \ldots, p$. The prior for $\theta$ was $\pi(\theta) = \prod_{i \in [p]} \pi(\theta_i)$ (which didn't depend on $z$) so the prior for $\phi$ is $\pi(\phi|m) = \prod_{i \in m} \pi(\phi_i)$. The dimension of $\phi \in \mathbb{R}^k$ varies depending on the number $k = |m|$ of components in the model.

Some variables $\bar{m} = [p] \setminus m$ were left out of $\phi$. If $|m| = k$ then $|\bar{m}| = p - k$. If $\bar{m} = \{j_1, \ldots, j_{p-k}\}$ then we dropped $(\theta_{j_1}, \ldots, \theta_{j_{p-k}})$. Let $\psi = (\psi_1, \ldots, \psi_{p-k})$ represent these variables in the new parameterisation. Their prior is $\pi(\psi|m) = \prod_{i \in \bar{m}} \pi(\psi_i)$. The re-parameterised posterior is

$$\pi(\phi, \psi, \sigma, m|y) \propto p(y|\phi, \sigma, m)\pi_s(\sigma)\pi_M(m)\pi(\phi|m)\pi(\psi|m)$$

with $p(y|\phi, \sigma, m) = N(y; X_{:,m}\phi, \sigma^2 I_n)$. Integrating $\psi$ to remove unused components we get

$$\pi(\phi, \sigma, m|y) \propto p(y|\phi, \sigma, m)\pi_s(\sigma)\pi_M(m)\pi(\phi|m). \tag{8.7}$$

Written this way the parameter space for $\phi, \sigma$ when the model is $M = m$ is $\Omega_m = \mathbb{R}^{|m|} \times \mathbb{R}^+$ and the joint distribution $\pi(\phi, \sigma, m|y)$ has state space $(\phi, \sigma, m) \in \Omega^*$ where

$$\Omega^* = \bigcup_{m \in \mathcal{M}} \{\{m\} \times \Omega_m\}.$$

# 9 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 presentation of RJ-MCMC (for Oxford students) in Pierre Jacob's old lecture notes for Advanced Simulation where the topic was once taught.

## 9.1 Lecture 11: 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 $\theta$ may vary depending on the model. As an example, consider a simplified version of the setup in Remark 8.9. Suppose we are doing linear regression with $\sigma$ known and we want to sample the posterior we get for a regression with just an intercept and a single covariate $x$, allowing for uncertainty in whether there is an effect due to $x$ at all.

| model index | model | parameter | parameter space |
|---|---|---|---|
| $m = 1$ | $Y = \theta_1 + \epsilon$ | $\theta = (\theta_1)$ | $\Omega_1 = \mathbb{R}$ |
| $m = 2$ | $Y = \theta_1 + \theta_2 x + \epsilon$ | $\theta = (\theta_1, \theta_2)$ | $\Omega_2 = \mathbb{R}^2$ |

In this example $\mathcal{M} = \{1, 2\}$ and $(\theta, m) \in \Omega^*$, where

$$\Omega^* = (\Omega_1 \times \{1\}) \ \cup \ (\Omega_2 \times \{2\}).$$

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 $\theta$ 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.

## 9.2 MCMC with a Jacobian

*We begin by revisiting fixed-dimension MCMC* (so no model index $m$ just yet) *and generalising it so we see the new 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.

### 9.2.1 Proposals from transformations

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 9.1.** (*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 continuous probability 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 $\Omega$ to itself given $u$ so that $\psi_1 : \Omega \times \mathcal{U} \to \Omega$. Given $\theta$, the proposal simulates $u \sim g(u)$ and sets $\theta' = \psi_1(\theta, u)$. Call $u$ the proposal variable and $\psi_1$ the proposal function. $\diamond$

*Remark* 9.2. Here $\psi_1(\theta, u)$ being invertible means the mapping $(\theta, u) \to \theta'$ at fixed $\theta$ is one-to-one between $u$ and $\theta'$. 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* 9.3. 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 $\theta$ since $u = (a + \theta' - \theta)/2a$. ♠

**Proposition 9.4.** *The conditional distribution of $\theta'$ given $\theta$ under the proposal in Definition 9.1 is $q(d\theta'|\theta) = q(\theta'|\theta)d\theta'$ where*

$$q(d\theta'|\theta) = g(u)du \tag{9.1}$$

*and*

$$q(\theta'|\theta) = g(u)\left|\partial\theta'/\partial u\right|^{-1}.$$

*Here $u = u(\theta', \theta)$ on the RHS solves $\theta' = \psi_1(\theta, u)$ at fixed $\theta$ and*

$$\frac{\partial\theta'}{\partial u} = \frac{\partial\psi_1(\theta, u)}{\partial u}.$$

*Proof.* Change variables from $u$ to $\theta'$ with $\theta$ a parameter in the change of variables. Applying the rule for changing variables in a density gives a Jacobian factor $|\partial\theta'/\partial u|^{-1}$. □

*Remark* 9.5. 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* 9.6. Continuing Example 9.3, where $\theta' = \theta + a(2u - 1)$ and $g(u) = \mathbb{I}_{0<u<1}$,

$$\begin{aligned}
q(\theta'|\theta) &= \mathbb{I}_{0<(a+\theta'-\theta)/2a<1}\left|\partial\theta'/\partial u\right|^{-1} \\
&= \frac{1}{2a}\mathbb{I}_{\theta-a<\theta'<\theta+a}
\end{aligned}$$

since $|\partial\theta'/\partial u| = 2a$. That is $\theta' \sim U(\theta - a, \theta + a)$. ♠

**Assumption 9.7.** 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

$$\theta = \psi_1(\psi_1(\theta, u), u'). \tag{9.2}$$

The solution for $u'$ depends on $\theta$ and $u$. We write $u' = \psi_2(\theta, u)$ with $\psi_2(\theta, u)$ a function mapping $(\theta, u)$ 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* 9.8. It is up to us to set things up so that a unique solution $u'$ to Equation 9.2 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* 9.9. Continuing Example 9.3, here $\psi_2(\theta, u) = 1 - u$, since $u' = 1 - u$ will "take us back". To show this starting from Equation 9.2 with $\psi_1(\theta', u') = \theta' + a(2u' - 1)$,

$$\begin{aligned}
\theta &= \psi_1(\psi_1(\theta, u), u') \\
&= \psi_1(\theta + a(2u - 1), u') \\
&= \theta + a(2u - 1) + a(2u' - 1).
\end{aligned}$$

Cancelling $\theta$ both sides gives $a(2u - 1) + a(2u' - 1) = 0$ so $u' = 1 - u$. It follows that Equation 9.2 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 $(\theta', u') = \psi(\theta, u)$ implies $(\theta, u) = \psi(\theta', u')$ (so the inverse of $\psi$ is $\psi$ itself). That means we want $(\theta, u) = (\psi_1(\theta', u'), \psi_2(\theta', u'))$. We already have $\theta = \psi_1(\theta', u')$ as that is how $u'$ was defined in Assumption 9.7, so we need to show that $u = \psi_2(\theta', u')$.

**Proposition 9.10.** *If $\theta' = \psi_1(\theta, u)$ and $u' = \psi_2(\theta, u)$ then $\psi_2(\theta', u') = u$.*

*Proof.* Since $\theta = \psi_2(\theta', u')$ it follows from Assumption 9.7 that there exists a unique $x \in \mathcal{U}$ solving $\theta' = \psi_1(\theta, x)$. But we already have a solution, namely $x = u$, so by the definition of $\psi_2$ in Assumption 9.7 we have $u = \psi_2(\theta', u')$. $\qquad\square$

**Proposition 9.11.** *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 9.2. The second component is

$$\psi_2(\psi_1(\theta, u), \psi_2(\theta, u)) = \psi_2(\theta', u') = u$$

by Proposition 9.10. $\qquad\square$

*Example* 9.12. If $\psi(\theta, u) = (\theta + a(2u - 1),\ 1 - u)$ as in Example 9.3 then

$$\begin{aligned}
\psi(\psi(\theta, u)) &= \psi(\theta + a(2u - 1), (1 - u)) \\
&= (\theta + a(2u - 1) + a(2(1 - u) - 1), 1 - (1 - u)) \\
&= (\theta, u)
\end{aligned}$$

$\spadesuit$

### 9.2.2   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 $\psi_1$ and let $q$ be whatever it is. This is often easier to do well. If you haven't read Section 5.1.7, now would be a good time.

We write down the Metropolis Hastings Algorithm in a slightly different form from Section 5, using our proposal functions and proposal variables, and show the algorithm satisfies detailed balance.

**Theorem 9.13.** *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*

$$\alpha(\theta'|\theta) = \min\left\{1,\ r(\theta', u'|\theta, u)\right\} \tag{9.3}$$

*with $r(\theta', u'|\theta, u)$ given by*

$$r(\theta', u'|\theta, u) = \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)} J_\psi(\theta, u), \tag{9.4}$$

*and $J_\psi$ the Jacobian for the transformation $(\theta', u') = \psi(\theta, u)$,*

$$J_\psi(\theta, u) = \left| \frac{\partial(\theta', u')}{\partial(\theta, u)} \right|, \tag{9.5}$$

*satisfies detailed balance in Equation 5.4 with respect to $\pi(\theta)$.*

*Proof.* we need to verify detailed balance, Eqn 5.4. Here it is again without the integral signs.

$$\pi(d\theta')q(d\theta|\theta')\alpha(\theta|\theta') = \pi(d\theta)q(d\theta'|\theta)\alpha(\theta'|\theta). \tag{9.6}$$

When we write it like this we mean it holds for all measureable sets $A$ and $B$ as in Eqn 5.4. Change variables on the RHS from $\theta'$ to $u$ at fixed $\theta$. Using Eqn 9.1,

$$\pi(d\theta)q(d\theta'|\theta)\alpha(\theta'|\theta) = \pi(\theta)g(u)\alpha(\theta'(\theta, u)|\theta)dud\theta.$$

We wrote $\theta'(\theta, u)$ instead of $\psi_1(\theta, u)$ for ease of reading. Suppose it holds that

$$r(\theta', u'|\theta, u) = 1/r(\theta, u|\theta', u'). \tag{9.7}$$

If that's true then either $r(\theta, u|\theta', u') \geq 1$ or $r(\theta', u'|\theta, u) \geq 1$ so we can take $r(\theta', u'|\theta, u) \leq 1$ WLOG. The RHS of (9.6) is

$$\begin{aligned}
\pi(\theta)g(u)\alpha(\theta'(\theta, u)|\theta)dud\theta &= \pi(\theta)g(u)\frac{\pi(\theta'(\theta, u))g(u'(\theta, u))}{\pi(\theta)g(u)}J_\psi(\theta, u)dud\theta \\
&= \pi(\theta'(\theta, u))g(u'(\theta, u))\left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right|dud\theta \\
&= \pi(\theta')g(u')du'd\theta' \tag{9.8}
\end{aligned}$$

since the Jacobian we have is correct for the change of variables. By our starting assumption $r(\theta, u|\theta', u') \geq 1$ so the LHS of DB in (9.6) is

$$\pi(d\theta')q(d\theta|\theta')\alpha(\theta|\theta') = \pi(\theta')g(u')du'd\theta'.$$

This matches the expression we got for the RHS in (9.8) so detailed balance is satisfied.

We assumed Eqn 9.7 holds. We now verify this. From (9.4) 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 $\theta, u$ 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 use the fact that $\psi$ is an involution so $\psi^{-1}(\theta, u) = \psi(\theta, u)$. In our setting,

$$\begin{aligned}
J_\psi(\theta, u) &= J_{\psi^{-1}}(\theta, u) \\
&= J_\psi(\theta', u')^{-1} \tag{9.9}
\end{aligned}$$

as the Jacobian of the inverse transformation is the inverse of the Jacobian for the transformation. We have shown Eqn 9.7 holds so the proof is almost complete.

When we make a change of variables we should transform the integration domain. Since we have shown that the integrands in Eqn 5.4 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. $\square$

*Remark* 9.14. Let's verify the result in Eqn 9.9 for the Jacobian in a bit more detail. Take

$$(\theta, u) = \psi(\psi(\theta, u))$$

and differentiate the column vector $(\theta, u)$ with respect to the row vector $\partial/\partial(\theta, u)$ on both sides.

$$\frac{\partial(\theta, u)}{\partial(\theta, u)} = \frac{\partial}{\partial(\theta, u)}\psi(\psi(\theta, u))$$

$$I_{\dim(\theta, u)} = \frac{\partial\psi(\theta, u)}{\partial(\theta, u)}\frac{\partial\psi(\psi(\theta, u))}{\partial(\psi(\theta, u)))}$$

$$= \frac{\partial(\theta', u')}{\partial(\theta, u)}\frac{\partial(\theta, u)}{\partial(\theta', u')}\bigg|_{(\theta', u')=\psi(\theta, u)},$$

using the chain rule for the first step and $\psi(\psi(\theta, u)) = (\theta, u)$ for the second. Multiply both sides by the inverse of the second matrix above (which we know is invertible) gives

$$\left[\frac{\partial(\theta, u)}{\partial(\theta', u')}\right]^{-1} = \frac{\partial(\theta', u')}{\partial(\theta, u)},$$

which gives $J_\psi(\theta', u')^{-1} = J_\psi(\theta, u)$ on taking the absolute values of the determinants. We have verified Eqn 9.9 starting from the definition of an involution and using the chain rule. ✠

*Remark* 9.15. 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 9.5 is a square matrix. This is called "dimension matching". ✠

*Example* 9.16. (*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 $g(u) = \mathbb{I}_{0.5 < u < 2}/(2 - 0.5)$ 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

$$\left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right| = \left|\begin{array}{cc} u & \theta \\ 0 & -1/u^2 \end{array}\right| = 1/u,$$

The algorithm (implemented in MCMC-Transformations-Lecture.R) 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 $\alpha$. 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* 9.17. show that the Jacobian for a simple random walk proposal $u \sim U(-a, a)$ with transformation $(\theta', u') = (\theta + u, -u)$ is equal one. ♣

*Exercise* 9.18. In the regression example in Section 8.2.2, the update for $\sigma$ was

$$u \sim U(\delta, 1/\delta), \qquad \sigma' = u\sigma$$

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, pretty much the same as Example 9.16). ♣

### 9.2.3   Matched proposals

We saw in Section 5.1.9 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

$$\sum_i \pi(d\theta)\xi_i K_i(\theta, d\theta') \;=\; \sum_i \pi(d\theta')\xi_i K_i(\theta', d\theta) \tag{9.10}$$

We can now "pair up" the kernels.

**Definition 9.19.** (*matched kernels*) Let $\sigma \in \mathcal{P}_N$ be a permutation of $\{1, ..., N\}$ satisfying $\sigma_{\sigma_i} = i$, so if $\sigma_i = j$ then $\sigma_j = i$ (for example $\sigma = (2, 1, 6, 5, 4, 3)$ has $\sigma_3 = 6$ and $\sigma_6 = 3$, we just swap pairs). 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

$$c_i(\theta) = 1 - \int_\Omega \alpha_i(\theta'|\theta)q_i(d\theta'|\theta)$$

so that

$$K_i(\theta, d\theta') = \alpha_i(\theta'|\theta)q_i(d\theta'|\theta) + c_i(\theta)\delta_\theta(d\theta').$$

$\diamond$

*Remark* 9.20. 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 9.21.** *Detailed balance in Equation 9.10 is satisfied if the following distributions match in $(i, \sigma_i)$ 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 9.10, the terms involving $c_i$ will cancel in detailed balance as they did in Proposition 5.10 (*ie*, between $K_i(\theta, d\theta')$ and $K_i(\theta', d\theta)$), so detailed balance is the condition that the integrands,

$$\sum_i \pi(d\theta)\xi_i\alpha_i(\theta'|\theta)q_i(d\theta'|\theta) \;=\; \sum_i \pi(d\theta')\xi_i\alpha_i(\theta|\theta')q_i(d\theta|\theta'),$$

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 $\alpha_i$ with $i \to \sigma_i$ we get

$$\alpha_{\sigma_i}(\theta|\theta') = \min\left\{1, \frac{\pi(\theta)\xi_{\sigma_{\sigma_i}}q_{\sigma_{\sigma_i}}(\theta'|\theta)}{\pi(\theta')\xi_{\sigma_i}q_{\sigma_i}(\theta|\theta')}\right\}$$
$$= \min\left\{1, \frac{\pi(\theta)\xi_i q_i(\theta'|\theta)}{\pi(\theta')\xi_{\sigma_i}q_{\sigma_i}(\theta|\theta')}\right\} = 1$$

since we required $\sigma$ 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. $\square$

*Remark* 9.22. A consequence is that *the individual kernels $K_i$, $i = 1, ..., N$ do not need to satisfy detailed balance* by themselves. In Example 9.24 below we will have two kernels in one of which $\theta \to \theta'$ is possible and $\theta' \to \theta$ is impossible, and the converse in the other. ✠

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 $\rho$ 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 9.1 is $q_i(d\theta'|\theta) = g_i(u)du$ and we can write

$$q_i(\theta'|\theta) = g_i(u)\left|\partial\theta'/\partial u\right|^{-1}.$$

Detailed balance in Proposition 9.21 becomes

$$\pi(d\theta)\rho q_1(d\theta'|\theta)\alpha_1(\theta'|\theta) = \pi(d\theta')(1 - \rho)q_2(d\theta|\theta')\alpha_2(\theta|\theta') \tag{9.11}$$

**Proposition 9.23.** *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 $\rho$ 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 $\theta'$ (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 $\theta'$ (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 9.21 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 $\psi$ transformation has been set up so that that side of things goes through as before. $\square$

*Example* 9.24. (*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 (implemented in MCMC-Transformations-Lecture.R) 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 $\theta'$ 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 $\theta'$ wp

$$\alpha(\theta'|\theta) = \min\{1, 0.5 \times e^{-\theta'+\theta}u^{-1}\}.$$ ♠

## 9.3 Reversible Jump MCMC

### 9.3.1 A shortcut to RJ-MCMC

Before we do RJ-MCMC properly, let's see how straightforward it can be. Thinking about the setup in model averaging, we need a proposal and an acceptance probability for $(\theta, m) \rightarrow (\theta', m')$.

*Remark* 9.25. 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 $(\theta, m) \rightarrow (\theta', m')$. If the distributions are interpreted correctly then we can actually write

$$\alpha(\theta', m'|\theta, m) = \min\left\{1, \frac{\pi(\theta', m'|y)\,Q(\theta, m|\theta', m')}{\pi(\theta, m|y)\,Q(\theta', m'|\theta, m)}\right\} \tag{9.12}$$

✠

*Example* 9.26. To be concrete, consider the variable dimension setup in Remark 8.9 in Section 8 (but here I use $\theta$ where Remark 8.9 uses $\phi$). 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 $\theta$. The joint model-parameter posterior is given in Equation 8.7,

$$\pi(\theta, \sigma, m|y) \propto p(y|\theta, \sigma, m)\pi(\theta)\pi_s(\sigma)\pi_M(m)$$

where $\pi_M(m)$ is given (we took $\pi_M(m) = \xi^k(1-\xi)^{p-k}$ in Section 8.2.1), and the likelihood is $p(y|\theta, \sigma, m) = N(y; X_{:,m}\theta, \sigma^2 I_n)$. Here $X_{:,m}$ is the matrix with 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

$$\tilde{\Omega}^* = \bigcup_{m\in\mathcal{M}} \{R^{|m|} \times R^+ \times \{m\}\}.$$

Suppose we propose $m \rightarrow m'$ 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 proposal is the prior, say) 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 $(\theta, m)$ to $(\theta', m')$. If $|m| = k$ we have

$$Q(\theta', m'|\theta, m) = \begin{cases} 1/2 \times 1/(p - k) \times \pi(\theta'_i) & \text{if we choose to add,} \\ 1/2 \times 1/k & \text{if we choose to delete,} \end{cases}$$

and "going back" from the new state, we have $|m'| = k + 1$ if we added and $|m'| = k - 1$ if we deleted. In order to reverse the move we have to pick the component we changed so,

$$Q(\theta, m|\theta', m') = \begin{cases} 1/2 \times 1/(p - |m'|) \times \pi(\theta_i) & \text{add back if we chose to delete,} \\ 1/2 \times 1/|m'| & \text{delete if we chose to add.} \end{cases}$$

$$= \begin{cases} 1/2 \times 1/(p - k + 1) \times \pi(\theta_i), \\ 1/2 \times 1/(k + 1). \end{cases}$$

The acceptance probabilities are

$$\alpha(\theta', m'|\theta, m) = 1 \wedge \frac{\pi(\theta', \sigma, m'|y) \times 1/2 \times 1/(k + 1)}{\pi(\theta, \sigma, m|y) \times 1/2 \times 1/(p - k) \times \pi(\theta'_i)}, \tag{9.13}$$

$$= 1 \wedge \frac{\pi(\theta')\pi_M(m')p(y|\theta', \sigma, m')(p - k)}{\pi(\theta)\pi_M(m)p(y|\theta, \sigma, m)(k + 1)\pi(\theta'_i)} \quad \text{if we chose to add,}$$

and

$$\alpha(\theta', m'|\theta, m) = 1 \wedge \frac{\pi(\theta', \sigma, m'|y) \times 1/2 \times 1/(p - k + 1) \times \pi(\theta_i)}{\pi(\theta, \sigma, m|y) \times 1/2 \times 1/k} \tag{9.14}$$

$$= 1 \wedge \frac{\pi(\theta')\pi_M(m')p(y|\theta', \sigma, m')k\pi(\theta_i)}{\pi(\theta)\pi_M(m)p(y|\theta, \sigma, m)(p - k + 1)} \quad \text{if we chose to delete}$$

This is implemented in RJ-MCMC-Swiss-Regress-Spike-Slab.R, using the same $\theta$-updates at fixed $m$ as before and the same $\sigma$ update. This is the same model we analysed in Section 8.2.1 (integrating out $\theta_{z=0}$). I applied it to model averaging on the `swiss` dataset from the `MASS` package in R. I got pretty much the same results as in Section 8.2.2 (up to MC error, and actually these new estimates come with higher ESS values so I expect are more accurate). The distribution over models was

```
111000 011000 011100 111100 011101 111101 011001 111001 011111 110110
 20.7   17.5   13.1   12.6    6.3    4.3    3.6    3.3    2.2    1.9
010110 111110 010111 011110 111111 110111 111010 011010 010100 110100
  1.8    1.7    1.6    1.6    1.5    1.3    1.2    1.1    0.6    0.6
011011 101000 111011 000010 010101 100111 101010 110101 000011 000110
  0.3    0.3    0.3    0.1    0.1    0.1    0.1    0.1    0.0    0.0
```

The dimension of the state varies with the MCMC step. If the simulated MCMC state at step $t = 1, \ldots, T$ is $X_t = (\theta^{(t)}, m^{(t)}, \sigma^{(t)})$ then $|m^{(t)}| = \dim(\theta^{(t)})$ is the state dimension so it is the number of covariates which are explanatory for `Fertility` in the $t$'th sample. This is plotted in Fig. 16. You can check MCMC diagnostic plots and compare then with the (more or less identical) summary plots for Section 8.2.2 using the code for this lecture. ♠

### 9.3.2 Reversible jump proposals

Consider MCMC targeting a general model-averaging posterior

$$\pi(\theta, m|y) \propto p(y|\theta, m)\pi(\theta|m)\pi(m), \ \theta \in \Omega_m, \ m \in \mathcal{M}$$
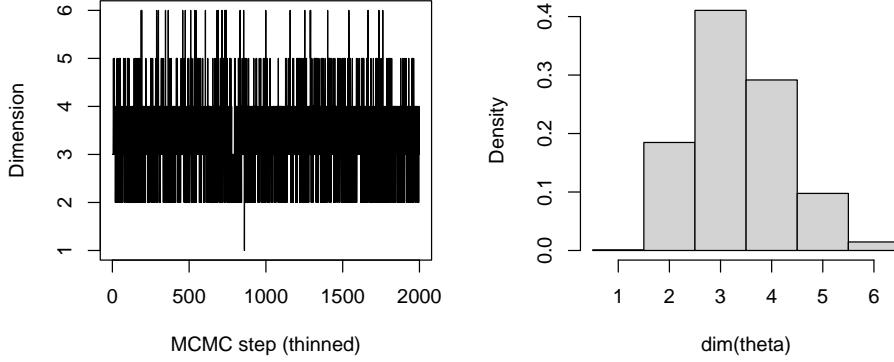
Figure 16: (Left) MCMC trace for the number of components, $|m^{(t)}| = \dim(\theta^{(t)})$ and (Right) a histogram of the sampled dimensions. This give the posterior distribution for the number of explanatory effects.

with $d_m = \dim(\Omega_m)$ so if $\theta \in \Omega_m$ then $d\theta$ is the element of volume in $\mathbb{R}^{d_m}$. When we move between models the dimension of $\theta$ my change. We can use the proposal matching setup of the last section to handle this.

Let $\rho_{m,m'}$ give the probability to propose a move to model $m'$ given the current model is $m$ so $\sum_{m' \in \mathcal{M}} \rho_{m.m'} = 1$ and suppose $\rho_{m,m'} > 0 \Leftrightarrow \rho_{m',m} > 0$. In the following, when we propose a state in $\Omega_{m'}$ the current model $m$ is known and the choice of $m'$ has been made, so $\psi_1$ and $\psi_2$ are specific to a given pair $m$ and $m'$. We don't make this explicit in the notation, but what follows applies for any pair $(m, m')$ such that $\rho_{m,m'} > 0$ including fixed dimension moves with $m = m'$.

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 $\mathcal{U}_{m,m'} = \mathbb{R}^{d_{m,m'}}$. Here $u$ are all the continuous random variables we need to make a state $\theta' \in \Omega_{m'}$ out of a state $\theta \in \Omega_m$. 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. If we have $\theta'$ and $\theta$ then we require $\theta = \psi_1(\theta', u')$ for some unique $u' \in \mathcal{U}_{m',m}$ and since that means $u'$ solves $\theta = \psi_1(\psi_1(\theta, u), u')$ we see that $u'$ is a function of $\theta$ and $u$ and we let $u' = \psi_2(\theta, u)$ be that function.

Assembling these 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. If

$$(\theta', u') = (\psi_1(\theta, u), \psi_2(\theta, u)),$$

then we require

$$(\theta, u) = (\psi_1(\theta', u'), \psi_2(\theta', u'))$$

so we require that $\psi$ is an involution. If we give $\psi$ a state in $\Omega_m \times \mathcal{U}_{m,m'}$ it maps it to a state in $\Omega_{m'} \times \mathcal{U}_{m',m}$ and *vis versa* so $\psi$ is a function defined on the two domains $\Omega_m \times \mathcal{U}_{m,m'}$ and $\Omega_{m'} \times \mathcal{U}_{m',m}$ which maps each to the other invertibly. We have set up $\psi$ 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 $\psi$ 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.

Page 95

### 9.3.3 The RJ-MCMC algorithm

**Proposition 9.27.** *Reversible Jump MCMC algorithm*

*For each $m, m' \in \mathcal{M}$ such that $\rho_{m,m'} > 0$ let*

$$\psi : \Omega_m \times \mathcal{U}_{m,m'} \to \Omega_{m'} \times \mathcal{U}_{m',m}$$

*be an involution which is invertible and differentiable.*

*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 $(\theta', u') = \psi(\theta, u)$ and*

$$\alpha(\theta', m'|\theta, m) = \min\left\{1, \frac{\pi(\theta', m'|y)\rho_{m',m}g_{m',m}(u')}{\pi(\theta, m|y)\rho_{m,m'}g_{m,m'}(u)} J_\psi(\theta, u)\right\} \tag{9.15}$$

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 9.16 below with respect to $\pi(\theta, m|y)$ between the pair of transition kernels associated with the updates $m \to m'$ and $m' \to m$.*

*Remark* 9.28. Here $J_\psi(\theta, u)$ must be non-singular so $(\theta', u') = \psi(\theta, u)$ must be invertible and differentiable. Making $\psi$ a differentiable involution ensures this. A necessary condition is

$$\dim(\Omega_m \times \mathcal{U}_{m,m'}) = \dim(\Omega_{m'} \times \mathcal{U}_{m',m}).$$

This is dimension matching. &#x2720;

*Proof.* Detailed balance (omitting the case where we reject) is

$$\int_B \pi(d\theta', m'|y)\rho_{m',m} \int_A q_{m',m}(d\theta|\theta')\alpha(\theta, m|\theta', m') =$$
$$\int_A \pi(d\theta, m|y)\rho_{m,m'} \int_B q_{m,m'}(d\theta'|\theta)\alpha(\theta', m'|\theta, m) \tag{9.16}$$

where $A$ and $B$ are any two sets for which the integrals are defined (so $A \times B$ is measureable in $\pi(d\theta, m|y)q_{m,m'}(d\theta'|\theta)$ which will imply $B \times A$ is $\pi(d\theta', m'|y)q_{m',m}(d\theta|\theta')$-measurable as we see below). We have

$$q_{m,m'}(d\theta'|\theta) = g_{m,m'}(u)du$$

and for the reverse update it is

$$q_{m',m}(d\theta|\theta') = g_{m',m}(u')du'.$$

Proceeding as in the proof of Theorem 9.13, and assuming $\alpha(\theta', m'|\theta, m) \leq 1$ (WLOG), under the integrals on the RHS of Equation 9.16 we have

$$\text{RHS}(9.16) = \pi(\theta, m|y)\rho_{m,m'}g_{m,m'}(u)\frac{\pi(\theta', m'|y)\rho_{m',m}g_{m',m}(u')}{\pi(\theta, m|y)\rho_{m,m'}g_{m,m'}(u)}\left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right|d\theta du$$
$$= \pi(\theta', m'|y)\rho_{m',m}g_{m',m}(u')d\theta'du'$$
$$= \pi(d\theta', m'|y)\rho_{m',m}q_{m',m}(d\theta|\theta')$$

which is the left hand side of (9.16) as $\alpha(\theta, m|\theta', m') = 1$ by the same reasoning as for Theorem 9.13 (ie exploiting the fact that $J_\psi(\theta, u) = J_\psi^{-1}(\theta', u')$ for a transform $\psi$ which is an involution). Again, we have not tracked the integration domains as we should, but we integrate over a product space $A \times B$ in Eqn 9.16 the transformations are all invertible so when we transform $d\theta d\theta' \to d\theta du \to d\theta'du' \to d\theta'd\theta$ we end up integrating over the same sets we started with in the opposite order. $\square$

*Remark* 9.29. The main new thing here compared to the simplest DB we wrote down for the discrete case in Definition 5.1 is that we need

$$\left| \frac{\partial(\theta', u')}{\partial(\theta, u)} \right| d\theta du = d\theta' du'.$$

Setting up the transformation as a differentiable involution ensures this. ✠

*Example* 9.30. Lets set how this applies for Example 9.26. We used the shortcut rule for constructing RJ acceptance probabilities, just "writing down what we did" to give the numerator and denominator in RJMCMC. Here we check that if we apply Proposition 9.27 we get the same acceptance probabilities.

Suppose the state is $(\theta, m)$ with $m = (i_1, \ldots i_k)$ and $\theta = (\theta_{i_1}, \ldots, \theta_{i_k})$. Here $m$ and $\theta$ are ordered sets with the components of $m$ increasing, because the effects are associated with particular covariates and we want them to line up correctly in the inner product $X_{:,m}\theta$. We choose to add a component with probability $1/2$ and otherwise we choose to delete. If we choose to add then we select $i \in [p] \backslash m$, simulate the new component $\theta'_i \sim \pi(\cdot)$ from the prior and set $m' = m \cup \{i\}$ and $\theta' = \theta \cup \{\theta'_i\}$ (putting them in the right place in order, not just tacking the new elements on the end). This is the same as choosing $m'$ with probability $\rho_{m,m'} = 1/2 \times 1/(p-k)$, drawing $u \sim g_{m,m'}(\cdot)$ with $g_{m,m'}(u) = \pi(u)$ and setting $\theta'_i = u$.

The spaces defined so far are $\Omega_m = \mathbb{R}^k$, $\mathcal{U}_{m,m'} = \mathbb{R}$ and $\Omega_{m'} = \mathbb{R}^{k+1}$. The reverse move will delete the element we just added. In order to match dimensions we need $\dim(\Omega_m \times \mathcal{U}_{m,m'}) = \dim(\Omega_{m'} \times \mathcal{U}_{m',m})$. The LHS has dimension $k+1$ ($k$ for $\theta$ and 1 for $u$) The RHS has dimension $k+1+\dim(\mathcal{U}_{m',m})$ ($k+1$ for $\Omega_{m'}$). We see that we need $\dim(\mathcal{U}_{m',m}) = 0$ in the reverse move.

If we choose to delete then we select $i \in m$, set $m' = m \setminus \{i\}$ and set $\theta' = \theta \setminus \{\theta_i\}$ so $\theta' \in \mathbb{R}^{k-1}$. This is the same as choosing $m'$ with probability $\rho_{m,m'} = 1/2 \times 1/k$. We don't need $u$ as there is no further randomness in the proposal. In order to make sure everything is defined we draw $u \sim g_{m,m'}(\cdot)$ with $g_{m,m'}(u) = \mathbb{I}_{u=\emptyset}$. This gives us $\mathcal{U}_{m,m'} = \{\emptyset\}$ and $\dim(\mathcal{U}_{m,m'}) = 0$.

The mapping for addition is $\psi_1(\theta, u) = (\theta_{i_1}, \ldots, \theta_{i_j}, u, \theta_{i_{j+1}}, \ldots, \theta_{i_k})$ and the mapping for deletion is $\psi_1(\theta, \emptyset) = (\theta_{i_1}, \ldots, \theta_{i_j}, \theta_{i_{j+1}}, \ldots, \theta_{i_k})$ when $i_j < i < i_{j+1}$. If we delete then $u = \emptyset$ and $u' = \theta_j$, the element we deleted, so that $\psi_1(\theta', u') = \theta$ puts back what we took out according to the map defined for addition. We have $\emptyset = \psi_2(\theta, \theta'_i)$ if we are adding and $\theta_i = \psi_2(\theta, \emptyset)$ if we are deleting.

Now calculate the acceptance probabilities using Proposition 9.27. If we choose to add then $m \to m'$ takes us from $k$ to $k+1$ components so $\rho_{m,m'} = 1/2(p-k)$. The reverse move $m' \to m$ is a deletion from $k+1$ to $k$ components so $\rho_{m',m} = 1/2(k+1)$. The acceptance probability is

$$\alpha(\theta', m'|\theta, m) = \min\left\{ 1, \frac{\pi(\theta', m'|y)\rho_{m',m}g_{m',m}(u')}{\pi(\theta, m|y)\rho_{m,m'}g_{m,m'}(u)} J_\psi(\theta, u) \right\} \tag{9.17}$$

$$= \min\left\{ 1, \frac{\pi(\theta', m'|y) \times 1/2(k+1) \times 1}{\pi(\theta, m|y) \times 1/2(p-k) \times \pi(\theta'_i)} \times 1 \right\} \tag{9.18}$$

where the Jacobian is 1 as its rows correspond to elements of $\theta'$ and it columns to $\theta, u$. These are mapped one to one so every row and column has exactly one entry equal 1 and the rest equal 0. This matches Eqn 9.17 for the addition update in Example 9.26. ♠

*Exercise* 9.31. Show the acceptance probability in Eqn. 9.14 equals the acceptance probability in Proposition 9.27. ANS: for deleting $\rho_{m,m'} = 1/2k$, $\rho_{m',m} = 1/2(p-k+1)$, $g_{m,m'}(u) = \mathbb{I}_{u=\emptyset}$ and $g_{m',m}(u') = \pi(\theta_i)$. The Jacobian is one as the elements of $\theta, \emptyset$ are mapped one to one to elements of $\theta', \theta_i$. Plug these into the acceptance probability in Proposition 9.27 to get Eqn. 9.14. ♣

## 9.4 Galaxy radial velocity data: RJ MCMC for mixture models

The Galaxy radial velocity data are shown in Figure 19. It is natural to model this via a mixture of normals in which mixture components might capture different classes of galaxy. This is a 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.

### 9.4.1 Observation model

We model our data $y_i \in \mathbb{R}, i = 1, 2, ..., n$ 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 vector of $\theta$'s is one of $m!$ copies of the same state.

Notice the model index is just a positive integer $m$ now, not an ordered set, as it was in the regression Examples 9.26 and 9.30. This is because we dont need to keep track of the order of the $\theta$'s because any permutation of the indices $(\theta_1, ..., \theta_m) \to (\theta_{\sigma_1}, ..., \theta_{\sigma_m})$ gives the same likelihood. That wasn't true in Example 9.26 as the link between parameters and covariates (columns of $X$) meant we had to maintain the order in $\theta$.

### 9.4.2 Priors

We need a prior $\pi_M$ 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

$$w \sim \text{Dirichlet}(\alpha 1_m) \qquad \text{with } 1_m \text{ a vector of } m \text{ ones}$$

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 $2\sigma$ - 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 $\sigma$; the small rate equal 0.5 gives a relatively heavy tail, and the standard deviation of the prior for $\sigma$ is about 2.5.

### 9.4.3 Mixture model posterior

The posterior for the model and parameters $(\theta, m)$, with $\theta_j = (\mu_j, \sigma_j, w_j)$, $j = 1, ..., m$, $\theta = \{\theta_1, ..., \theta_m\}$ and $m \in \{1, 2, 3, ...\}$ is

$$\pi(\theta, m | y) \propto p(y | \theta, m) \pi(\theta | m) \pi(m)$$

$$\propto p(y | \mu, \sigma, w, m) \times \text{Dirichlet}(w; \alpha 1_m) \times \prod_{j=1}^{m} N(\mu_j; \mu_0, v_0) \times \text{Gamma}(\sigma_j; 1.5, 0.5)$$

$$\times \quad \text{Poisson}(m; \lambda) \times m! \tag{9.19}$$

The extra $m!$ is needed as the RHS is the posterior probability for a particular vector of $\theta$'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 $\lambda$, so the condition that we have at least one cluster can be dropped. We just constrain the space so that $m > 0$.

### 9.4.4 RJ MCMC algorithm targeting a normal mixture posterior

In the following we use the "shortcut" in Remark 9.25 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 $\theta, m$ to $\theta', m'$ and call this $Q(\theta', m' | \theta, m)$ in Equation 9.12.

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 the 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 9.12 for $(\mu', \sigma', w', m')$ given $(\mu, \sigma, w, m)$ 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 $(\mu, \sigma, w, m)$ given $(\mu', \sigma', w', m')$ 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 $(\mu', \sigma', w', m')$ with probability

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

where

$$\alpha^+ = \min \left\{ 1, \; \frac{\pi(\mu', \sigma', w', m' | y) \rho_{m',m} \frac{1}{m(m+1)}}{\pi(\mu, \sigma, w, m | y) \rho_{m,m'} q_{\mu\sigma}(\mu'_{m+1}, \sigma'_{m+1}) \times \frac{1}{m} \times \frac{1}{w_j}} \right\}$$

from Equation 9.12.

*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 $(\mu', \sigma', w', m')$ 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 $\mu$, $\sigma$ and $w$ respectively in fixed dimension moves.

### 9.4.5 RJ-MCMC fitting a normal mixture model for the Galaxy data

We apply the RJ-MCMC algorithm in Section 9.4.4 to target the posterior in Equation 9.19. See RJ-MCMC-Galaxy-Data.R for an implementation. 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 17. Convergence seems reasonable. The posterior distribution of the number of mixture components is shown in Figure 18. Our analysis averages over the number of clusters and allows us to estimate the likely number of distinct clusters. Figure 19 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. The fit seems reasonable.
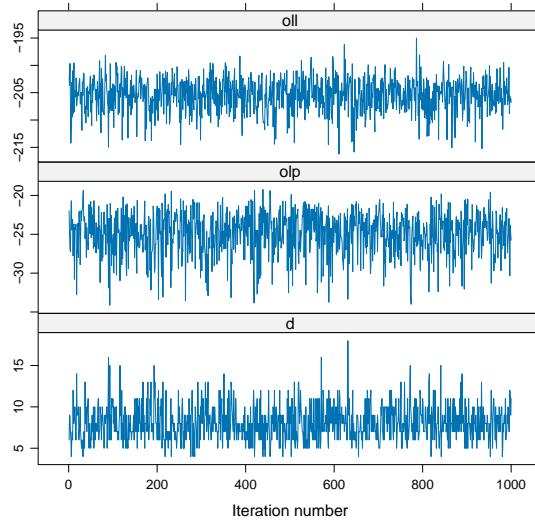
Figure 17: MCMC traces for the log-prior, log-likelihood and number of components, (as the number of parameters vary, the parameters themselves are not easily plotted).
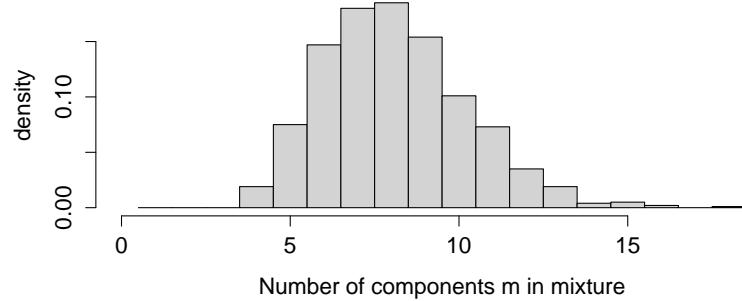


Figure 18: RJ-MCMC for the mixture model for the Galaxy radial velocity data: posterior distribution over the number of components. 6-9 components is the number favored.
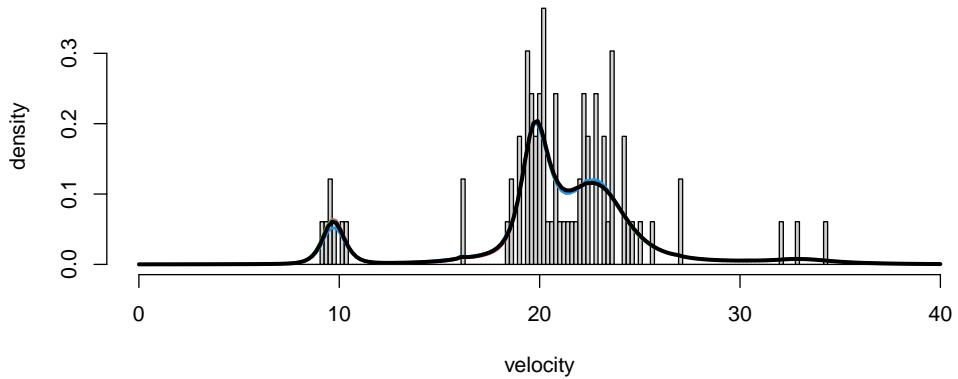


Figure 19: Measurements of radial velocities for $n = 82$ galaxies shown as a frequency histogram. Posterior predictive density (black curve) overlaid on the data. The underlying histogram is a histogram of the data, $y$. Other lines are $p(y'|y, m)$-estimates (m=4 red, m=6 green, m=8 blue).

Notes updated for Michaelmas 2025 to here.

# 10 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 10.41, 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.

## 10.1 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 have parametric elements so Prior elicitation and careful modeling still matter.

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 developed (model averaging, marginal likelihoods...) is the same.

Here is an informal view of where we are going. Recall the discussion in Section 3.3. 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

$$d\pi(\theta | y_1, ..., y_n) \propto p(y_1, ..., y_n | \theta) dG(\theta)$$

all exist. However we don't know $G$ (like we didn't know $\theta$). 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 $\Pi$ 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!

If we *informally* treat $G$ as another parameter then the joint prior distribution of $\theta$ and $G$ is

$$d\Pi(G, \theta) = G(d\theta) d\Pi(G)$$

in which case the posterior is

$$d\Pi(G, \theta | y) \propto p(y | \theta) G(d\theta) d\Pi(G).$$

It would in general be hopeless to express the infinite dimensional $G$ explicitly on a computer, so we will work with the marginal prior

$$\pi(d\theta) \propto \int_{\mathcal{G}} G(d\theta) d\Pi(G)$$

and posterior $\pi(d\theta|y) \propto p(y|\theta)\pi(d\theta)$ averaged over uncertainty in $G$. Prior elicitation moves up a level in the hierarchy to the choice of $\Pi$, a distribution over distributions.

## 10.2 The Dirichlet Process and the Chinese Restaurant Process

### 10.2.1 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

$$\pi(w_1, w_2, ..., w_M) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} w_1^{\alpha_1 - 1} w_2^{\alpha_2 - 1} ... w_M^{\alpha_M - 1}. \tag{10.1}$$

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 $w|n_1, ..., n_M \sim \mathsf{Dirichlet}(\alpha_1 + n_1, ..., \alpha_M + n_M)$.

### 10.2.2 The Multinomial Dirichlet process

We want to define a random probability distribution $G$, that is, we want to distribute probability randomly in parameter space $\Omega$ (which is always $\mathbb{R}^p$ or an open subset here). We will build this up using a simple base distribution $H$. Let $\theta_k^*$, $k = 1, 2, 3, ...$ 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 $\Omega$ with density $h$. We need $H$ to be in some sense "simple", but otherwise generalisation is straightforward.

**Definition 10.1.** 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 $w_1, ..., w_M \sim \mathsf{Dirichlet}(\alpha/M, ..., \alpha/M)$.

3. Set $G_M(d\theta) = \sum_{k=1}^{M} w_k \delta_{\theta_k^*}(d\theta)$. We alternatively write $G_M = \sum_{k=1}^{M} w_k \delta_{\theta_k^*}$.

This drops $M$ random probability masses $w_k$, $k = 1, ..., M$ at a random locations $\theta_k^*$ in $\Omega$. $\qquad \diamond$

*Remark* 10.2. Here $G_M$ gives the distribution of $\theta$ 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 $w_k : \theta_k^* \in A$:

$$
\begin{aligned}
G_M(A) &= \int_\Omega \mathbb{I}_{\theta \in A} G_M(d\theta) \\
&= \int_A \sum_{k=1}^M w_k \delta_{\theta_k^*}(d\theta) \\
&= \sum_{k=1}^M w_k \mathbb{I}_{\theta_k^* \in A}.
\end{aligned} \tag{10.2}
$$

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 $\theta^*$ we determine a prior *on a probability distribution*. The choice $w \sim \mathsf{Dirichlet}(\alpha/M, ..., \alpha/M)$, $\theta_k^* \sim H$, $k = 1, ..., M$ determines a Multinomial Dirichlet Process (MDP) $\Pi_M(\alpha, H)$. Let us look at some of its properties.

**Proposition 10.3.** *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.* The mean of the Dirichlet distribution in Eqn 10.1 is $E(w_k) = \alpha_k / \sum_j \alpha_j$ so from Eqn 10.2,

$$
\begin{aligned}
E(G_M(A)) &= \sum_{k=1}^M E(w_k \mathbb{I}_{\theta_k^* \in A}) \\
&= \sum_{k=1}^M E(w_k) E(\mathbb{I}_{\theta_k^* \in A}) \\
&= \sum_{k=1}^M \frac{1}{M} H(A) \\
&= H(A)
\end{aligned}
$$

as $\alpha_k = \alpha/M$, $k = 1, \ldots, M$ in Definition 10.1 and using the fact that $w$ and $\theta^*$ are independent.  □

*Exercise* 10.4. Use the Dirichlet-variance formula (look this up) to calculate $\mathrm{var}(G_M(A))$.  ♣

*Remark* 10.5. An important property of the MDP and the DP below is that it is "atomic". It puts atoms of probability mass at points in $\Omega$. As a consequence, if $G_M \sim \Pi_M(\alpha, H)$ and $\theta_1, \theta_2 \sim G_M$, then marginally $\Pr(\theta_1 = \theta_2) > 0$ (even though $\theta \sim H$ is a continuous random variable).  ✠

*Exercise* 10.6. Show that $\Pr(\theta_1 = \theta_2 | \theta^*, w) = \sum_k w_k^2$ and hence $\Pr(\theta_1 = \theta_2) > 0$ for $\alpha > 0$.  ♣

### 10.2.3 The Dirichlet *Process*

**Definition 10.7.** (*Dirichlet process*): $G \sim \Pi(\alpha, H)$ is a Dirichlet Process (DP) iff for all partitions $A_1, ..., A_r$ of $\Omega$ (with $A_k \in \mathcal{B}$, $k = 1, ..., r$), we have

$$
G(A_1), ..., G(A_r) \sim \mathsf{Dirichlet}(\alpha H(A_1), ..., \alpha H(A_r)).
$$

Perhaps surprisingly, there exists a unique process satisfying these conditions.  ◇

*Remark* 10.8. 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 $\Omega$. ✠

*Remark* 10.9. 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 (the Blackwell-MacQueen urn scheme) realising an IES $\theta_1, ..., \theta_n$ directly given $\alpha, H$. By de Finetti (in a statement of the theorem more general than the one we wrote down) a random variable $G$ and distribution $d\Pi(G; \alpha, H)$ must exist. ✠

*Remark* 10.10. 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 $\theta_k^*$ having weight $w_k \geq 0$, and $\sum_{k=1}^{\infty} w_k = 1$. ✠

### 10.2.4 Some properties of the Dirichlet Process

A number of properties follow immediately from Definition 10.7.

**Proposition 10.11.** *If $G \sim \Pi(\alpha, H)$ then for any set $A \in \mathcal{B}$ we have $E_{G \sim \Pi}(G(A)) = H(A)$*

*Proof.* Since $G \sim \Pi(\alpha, H)$ and $H(A^c) = 1 - H(A)$, we have from Definition 10.7

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

The two-component Dirichlet is a Beta for $G(A)$, so $G \sim \Pi(\alpha, H) \Rightarrow G(A) \sim \text{Beta}(\tilde{\alpha}, \tilde{\beta})$ with $\tilde{\alpha} = \alpha H(A)$ and $\tilde{\beta} = \alpha(1 - H(A))$. That gives $E_{G \sim \Pi}(G(A)) = E_{G(A) \sim B(\tilde{\alpha}, \tilde{\beta})}(G(A))$ with

$$E_{G(A) \sim B(\tilde{\alpha}, \tilde{\beta})}(G(A)) = \frac{\alpha H(A)}{\alpha H(A) + \alpha(1 - H(A))}$$

and hence $E_{G \sim \Pi}(G(A)) = H(A)$. □

**Proposition 10.12.** *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:

$$\begin{aligned} \Pr(\theta \in A) &= E_{G \sim \Pi}(E_{\theta \sim G}(\mathbb{I}_{\theta \in A} | G)) \\ &= E_{G \sim \Pi}(G(A)), \end{aligned}$$

so $\Pr(\theta \in A) = H(A)$ for all $A \in \mathcal{B}$ by Proposition 10.11. It follows that $\theta \sim H$. □

*Remark* 10.13. If $G \sim \Pi(\alpha, H)$ and $\theta \sim G$ then for $B \subseteq A$ both in $\mathcal{B}$,

$$\Pr(\theta \in B | \theta \in A) = \frac{\Pr(\theta \in B)}{\Pr(\theta \in A)} = \frac{H(B)}{H(A)}$$

If $H(dx) = h(x)dx$, $x \in \Omega$ then marginally $\theta | \theta \in A$ has density $h(\theta | \theta \in A)$. ✠

### 10.2.5 Clustering with the Dirichlet Process

This section gives motivation for what follows. Suppose our data $y \in \mathbb{R}^n$ come in $K$ groups, where for $k \in [K]$ the set $S_k$ gives the labels $i \in [n]$ of the samples in group $k$. That means $S = (S_1, \ldots, S_K)$ is a partition of $[n]$. The samples $y_i$, $i \in S_k$ in group $k$ are iid with observation model $y_i \sim f(\cdot|\theta_k^*)$ so $f(y|\theta^*, S)$ is just a product over data. Now suppose we don't know the number of groups $K$, we don't know how the data are partitioned $S$ and we don't know $\theta^* = (\theta_1^*, \ldots, \theta_K^*)$.

We would like to cluster the data and recover the parameters, so we would like to estimate the number of clusters $K$, the grouping $S$ and the parameters $\theta^*$. We can think of $S$ as a model index. If there are $K$ clusters in $S$ then our prior for the cluster parameters conditioned on the model is $\pi(\theta^*|S)$. We also need a prior $\pi(S)$ for the unknown partition. With all this we have a posterior

$$\pi(\theta^*, S|y) \propto \pi(S)\,\pi(\theta^*|S)\,f(y|\theta^*, S). \tag{10.3}$$

In the model-averaging notation of Chapter 8, Eqn 10.3 is $\pi(\theta, m|y) \propto \pi_M(m)\pi(\theta|m)p(y|\theta, m)$.

We can equivalently supply each observation $y_i$ with its own $\theta_i$ so $y_i \sim f(\cdot|\theta_i)$, $i \in [n]$. What prior should we use for $\theta = (\theta_1, \ldots, \theta_n)$? We want the $\theta$'s to come in groups, with all $\theta$'s in a group equal. This is going to be a complicated prior because the probability for two components of $\theta$ to be equal is not zero, so it won't be a simple continuous density.

However, we have seen that the DP is a discrete distribution, so if $\theta_1, \ldots, \theta_n|G \sim G$ jointly independent given $G$, with $G \sim \Pi(\alpha, H)$, then (1) the marginal prior density for each $\theta_i$ is $h$, which we elicit as in Chapter 1, and (2) there will be ties among the $\theta$'s, as several $\theta$'s may choose the same atom of $G$. If we take $S$ to be the partition defining the groups of tied $\theta$'s and $\theta_k^*$ to give the parameter-value shared by the $\theta$'s in group $k$ then the DP will determine prior distributions for $S$ and $\theta^*$ and we get the setup we wanted in Eqn 10.3.

In Section 10.2.6 we derive the joint distribution $d\pi(\theta)$ for $\theta$. As noted above this is not a simple continuous density. In Section 10.2.7 we reparameterise with $(S, \theta^*)$ and derive its generative model from the DP. This generative model is called the Blackwell-MacQueen urn scheme. In Sections 10.2.8 and 10.2.9 we work out the joint distribution $\pi(\theta^*, S)$ from the Blackwell-MacQueen urn scheme.

### 10.2.6 DP generative model and predictive distributions

Suppose we make a sequence of draws $\theta = (\theta_1, ..., \theta_n)$ from the same "random prior" $G$. Our aim in this and the next section will be to compute $d\pi(\theta)$, the joint marginal distribution of $\theta = (\theta_1, ..., \theta_n)$ which we get by integrating out the common random measure $G$. As in Section 10.2.4, 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 10.14.** The generative model for $\theta = (\theta_1, ..., \theta_n)$ is given by the following process:

1. $G \sim \Pi(\alpha, H)$
2. $\theta_i \sim G$, jointly independent for $i = 1, 2, ..., n$.

Denote by $d\pi(\theta)$ the marginal distribution of $\theta$ generated in this way. $\qquad \diamond$

*Remark* 10.15. Here $G$ is a "parameter" like $\theta$, but we want to work with the marginal $d\pi(\theta)$ not the joint for $G$ and $\theta$. We will use the identity

$$d\pi(\theta) = d\pi(\theta_n|\theta_{1:n-1})d\pi(\theta_{n-1}|\theta_{1:n-2})...d\pi(\theta_1) \tag{10.4}$$

and calculate each term in the product. We know $d\pi(\theta_1) = H(d\theta_1)$ (the marginal given in Proposition 10.12) so for $j = 1, ..., n-1$ the predictive distributions $d\pi(\theta_{j+1}|\theta_{1:j})$ are needed. ✠

**Proposition 10.16.** *(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*

$$G|\theta_1 \sim DP\left(\tilde{\alpha}_1, \tilde{H}_1\right)$$

*where $\tilde{\alpha}_1 = \alpha + 1$ and*

$$\tilde{H}_1 = \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}.$$

*It further holds that if $\theta_2 \sim G|\theta_1$ then marginally $\theta_2|\theta_1 \sim \tilde{H}_1$ or equivalently,*

$$d\pi(\theta_2|\theta_1) = \frac{\alpha H(d\theta_2) + \delta_{\theta_1}(d\theta_2)}{\alpha + 1},$$

*in Equation 10.4.*

*Proof.* By the definition of a DP in Definition 10.7 $G|\theta_1 \sim DP(\tilde{\alpha}_1, \tilde{H}_1)$ if and only if

$$G(A_1), ..., G(A_r)|\theta_1 \sim \mathsf{Dirichlet}(\tilde{\alpha}_1\tilde{H}_1(A_1), ..., \tilde{\alpha}_1\tilde{H}_1(A_r))$$

for all partitions $A_1, ..., A_r$ of $\Omega$, so let us show this holds.

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 $g = (g_1, ..., g_r)$, $H_i \equiv H(A_i)$ and let

$$f(g) = \mathsf{Dirichlet}(g; \alpha H_1, ...\alpha H_r),$$

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

$$f(g|\theta_1) \; \propto \; \pi(\theta_1|g_1, ..., g_r) \; f(g). \tag{10.5}$$

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 10.13. Here $\theta_1$ 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 10.5,

$$\begin{aligned}
f(g|\theta_1) &\propto h(\theta_1|\theta_1 \in A_j) \times g_j \times g_1^{\alpha H_1 - 1} \times ... \times g_r^{\alpha H_r - 1} \\
&\propto g_1^{\alpha H_1 - 1 + \mathbb{I}_{\theta_1 \in A_1}} \times ... \times g_r^{\alpha H_r - 1 + \mathbb{I}_{\theta_1 \in A_r}},
\end{aligned}$$

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 $G|\theta_1 \sim \Pi(\tilde{\alpha}_1, \tilde{H}_1)$ if we choose $\tilde{\alpha}_1$ and $\tilde{H}_1$ so that

$$\tilde{\alpha}_1\tilde{H}_1(A_j) = \alpha H(A_j) + \mathbb{I}_{\theta_1 \in A_j}, \quad \text{for } j = 1, ..., r.$$

As $\tilde{H}_1(\Omega) = H(\Omega) = 1$ we must have $\tilde{\alpha}_1 = \alpha + 1$ so

$$\tilde{H}_1 = \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}.$$

Finally, $G|\theta_1 \sim \Pi(\tilde{\alpha}_1, \tilde{H}_1)$ and $\theta_2 \sim G|\theta_1$ so marginally $\theta_2|\theta_1 \sim \tilde{H}_1$ by Proposition 10.12. □

*Remark* 10.17. 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 $(\theta_1, \theta_2)$ 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 $\theta_1$ so we may have $\theta_2 = \theta_1$. ✠

**Proposition 10.18.** *If $\theta_1, ..., \theta_j \sim G$ with $G \sim \Pi(\alpha, H)$ then*

$$G|\theta_{1:j} \sim \Pi\left(\tilde{\alpha}_j, \tilde{H}_j\right)$$

*with $\tilde{\alpha}_j = \alpha + j$ and*

$$\tilde{H}_j = \frac{\alpha H + \sum_{i=1}^{j}\delta_{\theta_i}}{\alpha+j}. \tag{10.6}$$

*It further holds that if $\theta_{j+1} \sim G|\theta_{1:j}$ then marginally $\theta_{j+1}|\theta_{1:j} \sim \tilde{H}_j$ or equivalently,*

$$d\pi(\theta_{j+1}|\theta_{1:j}) = \frac{\alpha H(d\theta_{j+1}) + \sum_{i=1}^{j}\delta_{\theta_i}(d\theta_{j+1})}{\alpha+n}$$

*Proof.* see Problem Sheet 4 (PS4) using induction with Propostion 10.16 above. The last part follows by Proposition 10.12 because $G|\theta_{1:j}$ is a draw from $\Pi\left(\tilde{\alpha}_j, \tilde{H}_j\right)$. $\square$

*Remark* 10.19. It follows that $d\pi(\theta)$ in Equation 10.4 is a product of the factors $\tilde{H}_j$ in Eqn 10.6,

$$d\pi(\theta) = \prod_{j=0}^{n-1} \frac{\alpha H(d\theta_{j+1}) + \sum_{i=1}^{j}\delta_{\theta_i}(d\theta_{j+1})}{\alpha+j}.$$

This is not very easy to work with. It is a product of singular distributions as some $\theta$-values may be exactly equal. Our new notation in the following section removes replicates. ✠

The generative process we wrote down involving $G$ can now be replaced with the marginal process.

**Proposition 10.20.** *Blackwell-MacQueen urn scheme: The 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 $\theta_{j+1} \sim U\{\theta_1, ..., \theta_j\}$.*

*realises the random variables $\theta = (\theta_1, \ldots, \theta_n)$ defined in Definition 10.14 marginally. No $G$ in sight!*

*Proof.* we have seen that $\theta_{j+1}|\theta_{1:j} \sim \tilde{H}_j$ with

$$\tilde{H}_j = \frac{\alpha}{\alpha+j}H + \frac{1}{\alpha+j}\sum_{i=1}^{j}\delta_{\theta_i}$$

and the algorithm simulates $\theta_{j+1}|\theta_{1:j} \sim \tilde{H}_j$ at each step $j = 1, ..., n-1$ so by Remark 10.19 it simulates $\theta \sim d\pi(\cdot)$ as defined in Definition 10.14. $\square$

### 10.2.7   Sequential simulation and repeated values

The marginal distribution $d\pi(\theta)$ on the LHS of Eqn 10.4 is not very helpful as it is a product of sums of singular distributions (ie, distributions involving $\delta$-functions). However, we can make a change of variables that makes it much easier to work with.

**Definition 10.21.** Let $\theta = (\theta_1, ..., \theta_n)$ be generated as in Proposition 10.20. Denote by $\theta^* = (\theta_1^*, ..., \theta_K^*)$ the $1 \leq K \leq n$ unique $\theta$-values in $(\theta_1, ..., \theta_n)$. Let $n_k$ give the number of times $\theta_k^*$ appears in $(\theta_1, ..., \theta_n)$. For $k = 1, ..., K$ let

$$S_k = \{i \in \{1, \ldots n\} : \ \theta_i = \theta_k^*\},$$

so these are the indices of the $\theta$-values matching $\theta_k^*$. We have a partition $S = \{S_1, ..., S_K\}$ of $1, ..., n$ with $n_k = |S_k|$ and a new set of variables $(\theta^*, S)$ with which to represent the distribution $d\pi(\theta)$. For $k = 1, ..., K$ we can call $S_k$ a "cluster".   $\diamondsuit$

*Remark* 10.22. 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* 10.23. 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* 10.24. Since $\theta_k^*$ appears $n_k$ times in the sum over $i$ in $\tilde{H}_j$ (Equation 10.6), we can write

$$\tilde{H}_j = \frac{\alpha}{\alpha + j}H + \frac{1}{\alpha + j}\sum_{k=1}^{K} n_k \delta_{\theta_k^*}. \tag{10.7}$$

In order to simulate $\theta_{j+1}$, sample a new $\theta_{K+1}^* \sim H$ with probability $\alpha/(\alpha+j)$ and set $\theta_{j+1} = \theta_{K+1}^*$, and otherwise choose a $k \in [K]$ with weights $(n_1, \ldots, n_K)$ and set $\theta_{j+1} = \theta_k^*$.   ✠

*Remark* 10.25. Blackwell-MacQueen Urn Scheme of Proposition 10.20 in the new variables:

1. $\theta_1^* \sim H$; set $K = 1$, $S_1 = \{1\}$ and $S = \{S_1\}$.
   # $\theta_1 = \theta_1^*$ *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 10.14.   ✠

### 10.2.8  The joint distribution of $\theta^*, S$

We observed that $d\pi(\theta)$ was a mess. What does $\pi(\theta^*, S)$ 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 $(\theta^*, S)$, output from the marginal generating process in Remark 10.25, 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 10.26.** Let $[n] = \{1, ..., n\}$ and let $\Xi_{[n]}$ be the set of all partitions of $[n]$. The distribution $\pi(\theta^*, S)$ is defined on the space

$$\Omega^* = \bigcup_{S \in \Xi_{[n]}} \Omega^{K(S)} \times \{S\}. \qquad \diamond$$

**Proposition 10.27.** *Let $(\theta^*, S) \sim \pi(\theta^*, S)$ be determined by the process in Remark 10.25 and suppose $H(dx) = h(x)dx$ is a continuous distribution on $\Omega$ with density $h$. The distribution of $(\theta^*, S)$ is*

$$\pi(\theta^*, S) = \pi_S(S)\pi(\theta^*|S), \quad (\theta^*, S) \in \Omega^*$$

*where $\pi_S(S)$, $S \in \Xi_{[n]}$ is a distribution over partitions and, if $S = (S_1, \dots, S_K)$,*

$$\pi(\theta^*|S) = \prod_{k=1}^{K} h(\theta_k^*),$$

*is the conditional probability density of $\theta^* \in \Omega^K$ given $S$.*

*Proof.* We can rearrange the Blackwell-MacQUeen urn scheme in Remark 10.25 so that we realise $S$ first and then $\theta_k^* \sim H$ independently for $k = 1, \dots, K$. This determines a distribution $\pi_S$ over partitions of $[n]$ and the conditional distribution for $\pi(\theta^*|S)$ given in the proposition. $\qquad \square$

*Remark* 10.28. We work out $\pi_S(S)$, $S \in \Xi_{[n]}$ in the next section, where we call it $P_{\alpha,[n]}(S)$. ✠

*Remark* 10.29. The dimension of $\theta^* \in \Omega^K$ is random, as $S$ is random, so $K = K(S)$ is random. If $\Omega = \mathbb{R}^p$ then $\dim(\Omega^K) = pK$ and $H(dx) = h(x)dx$ with $dx$ the element of volume in $\mathbb{R}^p$. ✠

*Remark* 10.30. If $A \subset \Omega^*$ is a set, chosen so the integrals below exist, and $(\theta^*, S) \sim \pi(\theta^*, S)$, then

$$\Pr\left((\theta^*, S) \in A\right) = \sum_{S \in \Xi_{[n]}} \int_{\Omega^{K(S)}} \mathbb{I}_{(\theta^*, S) \in A} \left[ P_{\alpha,[n]}(S) \prod_{k=1}^{K(S)} h(\theta_k^*) \right] d\theta_1^*, ..., d\theta_{K(S)}^*$$

so this is how we do integration in $\Omega^*$. We use Monte-Carlo to evaluate these integrals! ✠

### 10.2.9  The Chinese Restaurant Process

We now calculate $P_{\alpha,[n]}(S)$ (ie, $\pi_S(S)$). The sequential simulation of parameters in Remark 10.25 is analogous to restaurant seating. The authors of the CRP paper were impressed with the seemingly infinite capacity of Chinese restaurants - a table could always be found!

In this analogy $n$ customers with labels $j \in \{1, \dots, n\}$ are seated one by one at tables in a restaurant. The restaurant has an infinite number of tables with labels $k \in \{1, 2, 3, \dots\}$. After the $j$'th customer has arrived, customers are seated at tables $1, \dots, K_j$. The next customer, with

label $j + 1$, can choose to start a new table (which will be table $K_j + 1$) or they can sit at one of the already-occupied tables. Once all the customers are seated $K = K_n$ tables are occupied. If $S_k$ is the set of customers seated at table $k$ then $S = (S_1, \ldots, S_K)$ is a partition of $[n]$. Now a dish $\theta_k^* \sim H$ is served at each table $k \in [K]$. All the customers at table $k$ get dish $\theta_k^*$ so if customer $j \in [n]$ is seated at table $k$ then the dish $\theta_j$ served to customer $j$ is $\theta_j = \theta_k^*$.

**Definition 10.31.** (*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 $n_k^1 = 1$ people seated at table $k = 1$, and $K_1 = 1$ tables are occupied.
2. for $j = 1, \ldots, n - 1$
   (a) the $j + 1$'st arrival chooses new table $K_j + 1$ with probability $\alpha/(\alpha + j)$ and table $k \in \{1, \ldots, K_j\}$ with probability $n_k^j/(\alpha + j)$.
   (b) after customer $j + 1$ arrives, there are $n_k^{j+1}$ 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 $k = 1, \ldots, K$, set $S_k$ lists customers at table $k$ and $n_k = n_k^j$ 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* 10.32. If at the end we put an independent parameter $\theta_k^* \sim H$ on table $k = 1, \ldots, K$ (a single dish, which is shared!) then this is the same as the $(\theta^*, S)$ algorithm in Remark 10.25 above: the $\theta^*$'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 10.25. ✠

*Exercise* 10.33. (see PS4) Show from the CRP that

$$E(K) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}$$

♣

**Proposition 10.34.** *The CRP in Defintion 10.31 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* 10.35. 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 $T$,

$$P_{\alpha,[n]}(S) = 1 \times \frac{1}{\alpha + 1} \times \frac{\alpha}{\alpha + 2} \times \frac{2}{\alpha + 3} \times \frac{1}{\alpha + 4} \times \frac{\alpha}{\alpha + 5} \times \frac{1}{\alpha + 6} \times \frac{2}{\alpha + 7} \times \frac{3}{\alpha + 8} \times \frac{\alpha}{\alpha + 9}$$

$$= \alpha^3 \, 2! \, 3! \, 1! \, 0! \prod_{i=2}^{10} \frac{1}{\alpha + i - 1}$$

$$= \alpha^4 \, \Gamma(3) \, \Gamma(4) \, \Gamma(2) \, \Gamma(1) \prod_{i=1}^{10} \frac{1}{\alpha + i - 1}$$

$$= \alpha^K \left[ \prod_{k=1}^{K} \Gamma(n_k) \right] \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}$$

with $K = 4$ and $n = 10$ and using $\prod_{i=1}^{n}(\alpha + i - 1) = \Gamma(\alpha + n)/\Gamma(\alpha)$ from Exercise 10.36. ✠

*Exercise* 10.36. Show $\Gamma(\alpha + n) = \Gamma(\alpha) \prod_{i=1}^{n}(\alpha + i - 1)$ for $\alpha > 0$ (use $\Gamma(x+1) = x\,\Gamma(x)$, $x \geq 1$). ♣

*Proof of Proposition 10.34.* 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

$$\frac{j-1}{\alpha + i_j - 1}$$

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

$$P_{\alpha,[n]}(S) = \alpha^{K-1} \prod_{k=1}^{K}(n_k - 1)! \prod_{i=2}^{n} \frac{1}{\alpha + i - 1}$$

$$= \alpha^{K} \prod_{k=1}^{K}(n_k - 1)! \prod_{i=1}^{n} \frac{1}{\alpha + i - 1}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^{K} \prod_{k=1}^{K} \Gamma(n_k)$$

where we extended the product in the denominator down to one in the first step and in the second step we used Exercise 10.36. □

*Remark* 10.37. 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* 10.38. Let $\mathcal{P}_n$ be the set of all permutations of $\{1, ..., n\}$. For $\sigma \in \mathcal{P}_n$ let $P_{\alpha,\sigma}(S)$ be the distribution over partitions we get if the customers arrive in the order $\sigma = (\sigma_1, ..., \sigma_n)$ and let $S(\sigma)$ be the partition obtained by permuting the customer labels in $S$ according to $\sigma$. For example if $S = (\{1,2\}, \{3\})$ and $\sigma = (3, 2, 1)$ then $S(\sigma) = (\{1\}, \{2,3\})$ because the new partition is $\{\{\sigma_1, \sigma_2\}, \{\sigma_3\}\} = \{\{3,2\}, \{1\}\}$ and recall the set-labelling convention in Remark 10.22.

Show that $P_{\alpha,[n]}(S) = P_{\alpha,[n]}(S(\sigma)) = P_{\alpha,\sigma}(S)$ for all $S \in \Xi_{[n]}$, so CRP outcomes don't depend on customer arrival order. *Hint: $n_k$ doesn't change and see PS4 for solution.* ♣

*Exercise* 10.39. Let $(\theta^*, S)$ be a draw from $\pi(\theta^*, S)$ given $\alpha, H$ and $n \geq 1$. Let $\theta = \theta(\theta^*, S)$. Show that $\theta_1, ..., \theta_n, ...$ is an IES. *ANS: for $\sigma \in \mathcal{P}_n$, $\theta_\sigma = \theta(\theta^*, S(\sigma))$ and $\theta(\theta^*, S(\sigma)) \sim \theta(\theta^*, S)$ using Ex 10.38, so $\theta_{\sigma_1}, ..., \theta_{\sigma_n} \sim \theta_1, ..., \theta_n$ for every $n$. As the $\theta$'s are simulated sequentially the random variables are defined for every $n \geq 1$. Note that $\theta$ are now defined by the Blackwell-MacQueen urn scheme, so $G$ and $d\Pi(G; \alpha, H)$ exist per Remark 10.9.* ♣

*Exercise* 10.40. 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. Here $K^{-i} = K - 1$ if we create an empty cluster when we remove $i$ and otherwise $K^{-i} = K$. For example if $S = (\{1,2\}, \{3\})$ then $K = 2$ and $S^{-3} = (\{1,2\})$ so $K^{-3} = 1$.

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 before $S'$ is simulated from the CRP. Show that $S^{-i} \sim P_{\alpha,[n]\setminus\{i\}}(S^{-i})$. *Hint: making $i$ the last customer to arrive doesnt change the probability for $S$ by Exercise 10.38. Stopping the CRP before $i$ arrives or adding $i$ then removing it give the same partition $S^{-i}$. See PS4 for solution.* ♣

## 10.3 Inference for a Dirichlet process mixture

We now return to the clustering problem we set out in Section 10.2.5.

Suppose we have $n$ independent samples $y_i \sim f(y_i|\theta_i), i = 1, ..., n$ each with its own parameter $\theta_i$. The likelihood for $\theta = (\theta_1, ..., \theta_n)$ is

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

There seems to be a parameter for every observation, but the $\theta$'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 $(\theta^*, S)$ 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

$$f(y|\theta^*, S) = \prod_{k=1}^{K} \prod_{i \in S_k} f(y_i|\theta_k^*).$$

**Theorem 10.41.** *The posterior for $\theta$ under a DP prior $\theta \sim G$ with $G \sim \Pi(\alpha, H)$ is given by*

$$\pi(\theta^*, S|y) \propto f(y|\theta^*, S)\, \pi(\theta^*|S)\, P_{\alpha,[n]}(S), \tag{10.8}$$

*with $\theta = \theta(\theta^*, S)$ and $(\theta^*, S) \in \Omega^*$ given in Definition 10.26.*

*Proof.* apply Bayes rule and Proposition 10.27 with $\pi_S = P_{\alpha,[n]}$ by Proposition 10.34. □

*Remark* 10.42. I called this a theorem because it is really the point and outcome of Chapter 10. Equation 10.8 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* 10.43. In detail, from Proposition 10.27 and Proposition 10.34,

$$\pi(\theta^*, S|y) \propto \alpha^K \prod_{k=1}^{K} \Gamma(n_k)\, h(\theta_k^*) \prod_{i \in S_k} f(y_i|\theta_k^*), \tag{10.9}$$

where $S = (S_1, ..., S_K)$ so $K = |S|$ and $n_k = |S_k|$, $k = 1, ..., K$. ✠

*Remark* 10.44. This is a model averaging setup in which $\pi_M$ in Chapter 8 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 $\theta^*$ is a random variable. ✠

### 10.3.1 Normal mixture for the Galaxy data

Recall the Galaxy radial velocity data which we saw in Section 9.4. 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 9.4 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 $\mu_i$ and $\sigma_i$ are the parameters for the cluster to which $i$ belongs.

In terms of our $(\theta^*, S)$ 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

$$h_\sigma(\sigma_k^{*2}) = I\Gamma(\sigma_k^{*2}; \alpha_0, \beta_0).$$

with $\mu_0 = 20, \sigma_0 = 10, \alpha_0 = 2$ and $\beta_0 = 1/9$ fixed hyper-parameters, so from Theorem 10.41,

$$
\begin{aligned}
\pi(S, \mu^*, \sigma^* | y) \quad &\propto \quad f(y | \mu^*, \sigma^*, S) \pi(\mu^*, \sigma^* | S) P_{\alpha, [n]}(S) \\
&\propto \quad \prod_{k=1}^{K} \prod_{i \in S_k} N(y_i; \mu_k^*, \sigma_k^{*2}) \\
&\quad \times \prod_{k=1}^{K} N(\mu_k^*; \mu_0, \sigma_0^2) \, I\Gamma(\sigma_k^{*2}; \alpha_0, \beta_0) \\
&\quad \times \alpha^K \prod_{k=1}^{K} \Gamma(n_k).
\end{aligned}
$$

We dropped the denominator in the expression for $P_{\alpha, [n]}(S)$ as it does not depend on $S$. Comparing with Equation 9.19 for the posterior we wrote down in Section 9.4, there is no $n!$ - or $K!$ - that is because $S = (S_1, ..., S_K)$ is ordered by the labelling convention in Remark 10.22). Any given partition of the data $(y_1, ..., y_n)$ into clusters has a unique $S$. 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* 10.45. We take Normal/inv-Gamma for the $\mu^*/\sigma^*$-prior to keep things simple and conjugate, so that we can can Gibbs-sample $\mu^*, \sigma^*$. If we just did straightforward MH-MCMC on $\mu^*$ and $\sigma^*$ 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 $\mu^*$ and $\sigma^*$ with freedom. ✠

*Remark* 10.46. We took $\alpha = 1$ in this example. This controls the prior distribution on the number of clusters. I used simulation (of the CRP) to check this 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 $\alpha$ 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 $\alpha$ and infer it along with everything else using Metropolis-Hastings. ✠

### 10.3.2 Gibbs sampler for the mixture parameters $\mu^*, \sigma^*$

Iterate through the parameters sampling them conditionally. The conditional posterior distribution for $\mu_k^*$ 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 $\sigma^{*2}|\mu_k^*, y \sim I\Gamma(c, d)$ with

$$c = \alpha_0 + n_k/2, \qquad d = \beta_0 + \frac{1}{2}\sum_{i \in S_k}(y_i - \mu_k^*)^2.$$

### 10.3.3 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 $\Xi_{[n]}$. We 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. We give the MCMC step in Figure 20 and verify it is correct using reversible jump.

The intuition for this update is clear from Eqns. 10.10 and 10.11. When we remove $i$ from its cluster $S_{k_i}$ we create a new partition $S^{-i}$. Now we make $i$ the last arrival in the CRP and drop it back into a randomly chosen cluster as if it was the last arrival in the CRP. The cluster is chosen according to the CRP weights appropriate for $S^{-i}$. The main change is that we now have data, so we have to weight by a factor $f(y_i|\theta_k^*)$ for choosing cluster $k$. The two main branches of the algorithm handle the two cases where $i$ is in a cluster with more than one element (so removing $i$ wont empty the cluster) and the case there $i$ is by itself, so moving $i$ to other cluster will leave an empty cluster.

**Proposition 10.47.** *The acceptance probabilities for the proposals in Figure 20 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 in 2(c)ii when $n_{k_i} > 1$ and we choose $k' = K+1$ using the cluster selection distribution $\vec{p}(S, k_i)$. The reverse move is 3(b)ii and uses $k \sim \bar{p}(S', K+1)$: the current state is $(\theta', S')$ and we must select $i$ to move and $k = k_i$ to move

Let $X_t = (\theta^*, S)$.

1. Choose $i \sim U\{1, ..., n\}$ and suppose $i \in S_{k_i}$.

2. If $n_{k_i} > 1$ - moving $i$ out of $S_{k_i}$ wont empty $S_{k_i}$.
   (a) Simulate $\theta'_{K+1} \sim H$
   (b) let $\vec{q}(S, k_i) = (\vec{q}_1(S, k_i), \ldots, \vec{q}_{K+1}(S, k_i))$ with components

   $$\vec{q}_k(S, k_i) = \begin{cases} (n_k - 1)f(y_i|\theta^*_k) & \text{if } k = k_i, \\ \alpha f(y_i|\theta'_{K+1}) & \text{if } k = K + 1, \\ n_k f(y_i|\theta^*_k) & \text{for } k \in [K] \setminus \{k_i\}. \end{cases}, \qquad (10.10)$$

   (c) choose a new cluster $k' \sim (\vec{p}_k)_{k=1,...,K+1}$ for $i$ where $\vec{p}_k = \vec{q}_k / \sum_{j=1}^{K+1} \vec{q}_j$.
      i. (no change, put $i$ back where it was) If $k' = k_i$ then $(\theta', S') = (\theta^*, S)$.
      ii. (increase dim, put $i$ in a new cluster) 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, put $i$ in a different existing cluster) 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 $n_{k_i} = 1$ - moving $i$ out of $S_{k_i}$ will empty $S_{k_i}$.
   (a) let $\breve{q}(S, k_i) = (\breve{q}_1(S, k_i), \ldots, \breve{q}_K(S, k_i))$ with components

   $$\breve{q}_k(S, k_i) = \begin{cases} \alpha f(y_i|\theta^*_k) & \text{if } k = k_i \\ n_k f(y_i|\theta^*_k) & \text{for } k \in [K] \setminus \{k_i\}. \end{cases}, \qquad (10.11)$$

   (b) choose a new cluster $k' \sim (\breve{p}_k)_{k=1,...,K}$ for $i$ where $\breve{p}_k = \breve{q}_k / \sum_{j=1}^{K} \breve{q}_j$.
      i. (no change, put $i$ back where it was) If $k' = k_i$ then $(\theta', S') = (\theta^*, S)$.
      ii. (decrease dim, put $i$ in a different existing cluster and delete its old cluster) 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 the empty cluster $S'_{k_i}$: set $S' \leftarrow S' \setminus S'_{k_i}$ and $\theta' = \theta^*_{-k_i}$.

4. $X_{t+1} = (\theta', S')$.

We may need to resort the cluster labels to meet our labeling convention (see Remark 10.22) in step 2(c)ii,iii and 3(b)ii.

Figure 20: Gibbs sampler for the random partition $S$. See Algorithm 8 in RM Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models", JCGS 9:249-265 (2000)

it back into $S'_{k_i}$. Since $|S'_{K+1}| = 1$ we delete $S'_{K+1}$ after the update. Using Remark 9.25 we have,

$$\alpha(\theta', S'|\theta^*, S) = \min\left\{1, \frac{\pi(\theta', S'|y)}{\pi(\theta, S|y)} \times \frac{n^{-1}\breve{p}_{k_i}(S', K+1)}{n^{-1}h(\theta'_{K+1})\vec{p}_{K+1}(S, k_i)}\right\}$$

$$= \min\left\{1, \frac{f(y_i|\theta'_{K+1})h(\theta'_{K+1})\alpha^{K+1}\Gamma(n_{K+1})\Gamma(n_{k_i}-1)}{f(y_i|\theta^*_{k_i})\alpha^K\Gamma(n_{k_i})} \times \frac{\breve{q}_{k_i}(S', K+1)}{h(\theta'_{K+1})\vec{q}_{K+1}(S, k_i)}\right\}$$

$$= \min\left\{1, \frac{f(y_i|\theta'_{K+1})h(\theta'_{K+1})\alpha}{f(y_i|\theta^*_{k_i})(n_{k_i}-1)} \times \frac{(n_{k_i}-1)f(y_i|\theta^*_{k_i})}{h(\theta'_{K+1})\alpha f(y_i|\theta'_{K+1})}\right\},$$

which equals one. Most of this is just careful accounting, but the key step is canceling the denom-

inators in $\breve{p}_{k_i}(S', K+1)$ and $\vec{p}_{K+1}(S, k_i)$. These are equal,

$$\sum_{j=1}^{K(S')} \breve{q}_j(S', K+1) = n_1' f(y_i|\theta_1') + ... + n_{k_i}' f(y_i|\theta_{k_i}') + ... + \alpha f(y_i|\theta_{K(S')}')$$

$$= n_1 f(y_i|\theta_1^*) + ... + (n_{k_i}-1)f(y_i|\theta_{k_i}^*) + ... + \alpha f(y_i|\theta_{K(S)+1}')$$

$$= \sum_{j=1}^{K(S)+1} \vec{q}_j(S, k_i).$$

The first line is the denominator for the reverse move (delete). We are emptying $S_{K+1}'$ so $K+1$ contributes a term $\alpha f(y_i|\theta_{K+1}')$ per case 1 of Eqn 10.11. Also, since $i$ was removed from $S_{k_i}$, we have $n_{k_i}' = n_{k_i} - 1$ in $S'$ and so cluster $k_i$ gives a term $(n_{k_i}-1)f(y_i|\theta_{k_i}^*)$ per case 2. The last line is the denominator for the forward move (add). In this move the old cluster $k_i$ gives a term $(n_{k_i}-1)f(y_i|\theta_{k_i}^*)$ per case 1 of Eqn 10.10 and new cluster $K+1$ gives $\alpha f(y_i|\theta_{K+1}')$ per case 2.

The dimension goes down in 3(b)ii when $n_{k_i} = 1$ and we select $k' \neq k_i$, so we delete a cluster. This 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 update 2(c)iii where $n_{k_i} > 1$ and we don't select $k' = k_i$ or $k' = K+1$ is interesting. This update is a fixed dimension update and is matched in detailed balance with itself. Notice that we have to generate $\theta_{K+1}' \sim h$ in 2(a) in both directions. We have

$$\alpha(\theta', S'|\theta^*, S) = \min\left\{1, \frac{\pi(\theta', S'|y)}{\pi(\theta, S|y)} \times \frac{n^{-1}h(\theta_{K+1}')\breve{p}_{k_i}(S', k')}{n^{-1}h(\theta_{K+1}')\breve{p}_{k'}(S, k_i)}\right\}$$

$$= \min\left\{1, \frac{f(y_i|\theta_{k'}^*)\Gamma(n_{k'}+1)\Gamma(n_{k_i}-1)}{f(y_i|\theta_{k_i}^*)\Gamma(n_{k'})\Gamma(n_{k_i})} \times \frac{\vec{q}_{k_i}(S', k')}{\vec{q}_{k'}(S, k_i)}\right\}$$

$$= \min\left\{1, \frac{f(y_i|\theta_{k'}^*)n_{k'}}{f(y_i|\theta_{k_i}^*)(n_{k_i}-1)} \times \frac{(n_{k_i}-1)f(y_i|\theta_{k_i}^*)}{n_{k'}f(y_i|\theta_{k'}^*)}\right\},$$

so again, the acceptance probability is one. Checking the denominators of $\breve{p}_{k_i}(S', k')$ and $\vec{p}_{k'}(S, k_i)$,

$$\sum_{j=1}^{K(S')+1} \breve{q}_j(S', k') = n_1' f(y_i|\theta_1') + ... + n_{k_i}' f(y_i|\theta_{k_i}') + ... + (n_{k'}'-1)f(y_i|\theta_{k'}') + ... + \alpha f(y_i|\theta_{K(S')+1}')$$

$$= n_1 f(y_i|\theta_1^*) + ... + (n_{k_i}-1)f(y_i|\theta_{k_i}') + ... + (n_k+1-1)f(y_i|\theta_{k'}^*) + ... + \alpha f(y_i|\theta_{K(S)+1}')$$

$$= \sum_{j=1}^{K(S)+1} \vec{q}_j(S, k_i)$$

so they cancel again. $\qquad\qquad\square$

*Remark* 10.48. This is implemented for our mixture model application where $\theta_k^* = (\mu_k^*, \sigma_k^{*2})$,

$$h(\theta_k^*) = N(\mu_k^*; \mu_0, \sigma_0^2)\,\mathrm{I\Gamma}(\sigma_k^{*2}; \alpha_0, \beta_0)$$

and $f(y_j|\theta_k^*) = N(y_i; \mu_k^*, \sigma_k^{*2})$, $k = 1, ..., K$ and $i = 1, ..., n$, in the code for this lecture. ✠

### 10.3.4 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 $f(y'|\theta', \theta^*, S) = f(y'|\theta')$ we get,

$$= \int_{\Omega} f(y'|\theta')p(\theta'|\theta^*, S)d\theta'$$
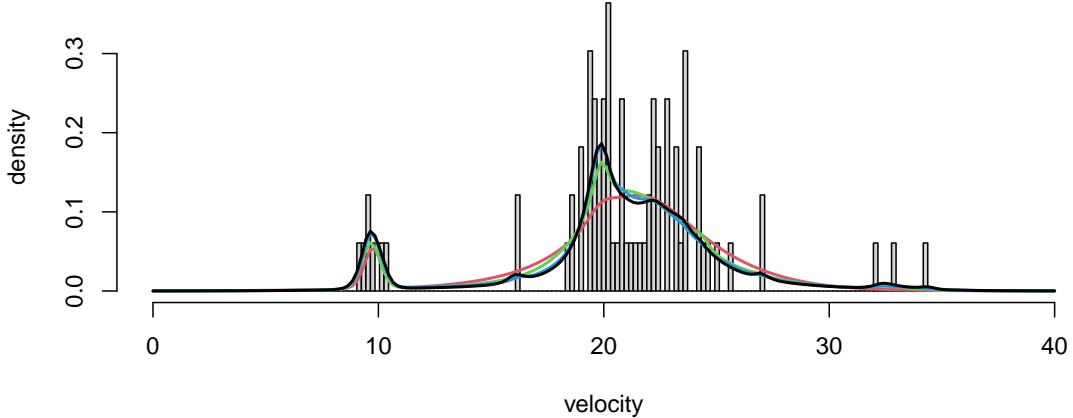
with $\theta'|\theta^*, S \sim \tilde{H}_n$ in Eqn 10.7 so,

$$= \int_{\Omega} f(y'|\theta') \left[ \frac{\alpha h(\theta') + \sum_{k=1}^{K} n_k \delta_{\theta_k^*}(\theta')}{\alpha + n} \right] d\theta'$$

$$= \frac{\alpha}{\alpha + n}p(y') + \sum_{k=1}^{K} f(y'|\theta_k^*)\frac{n_k}{\alpha + n}.$$

An unbiased estimate for $p(y') = E_{\theta' \sim h}(f(y'|\theta'))$ is given by $f(y'|\theta')$ with $\theta' \sim h(\cdot)$. In the conjugate-prior setting we don't 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, $\theta'^{(t)} \sim h(\cdot)$ iid for $t = 1, ..., T$,

$$\widehat{p(y'|y)} = \frac{1}{T}\sum_{t=1}^{T} \left[ \frac{\alpha}{\alpha + n}f(y'|\theta'^{(t)}) + \sum_{k=1}^{K^{(t)}} f(y'|\theta_k^{*,(t)})\frac{|S_k^{(t)}|}{\alpha + n} \right]$$

with $K^{(t)} = K(S^{(t)})$. Note that $T^{-1}\sum_t f(y'|\theta'^{(t)})$ is the naive estimator for $p(y')$ criticised in Section 6.1. However, here it is fine as $y'$ is just one observation, not $n$ observations (like $y$) so $f(y'|\theta')$ is not sharply peaked in $\Omega$ and the usual problems with the naive estimator don't arise.

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 $p(y'|y)$ (plotted solid black line) and the predictive distribution conditioned on $K = 3,4$ and 5 components (red, green and blue) and plot these as functions of $y'$ over the data. The fit seems reasonable for the model averaged predictive distribution in black, and for 4 or more clusters.

The code explores other visualisations of the output, such as the posterior distribution of the dimension-parameter $2K$.

## 10.4 Appendices

### 10.4.1 Appendix for Section 10.2.3: 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 $\Omega$. 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 10.49.** *For each $g \in \Lambda$,*

$$\lim_{M \to \infty} f_M(g) = Dirichlet(g; \alpha H(A_1), ..., \alpha H(A_r))$$

*so $G_M(A_1), ..., G_M(A_r) \xrightarrow{D} Dirichlet(\alpha H(A_1), ..., \alpha H(A_r))$ 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

$$N_1, ..., N_r \sim Multinomial(M, H_1, ..., H_r). \tag{10.12}$$

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 $\Lambda$ 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. 10.12, 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

$$\lim_{M\to\infty} f_M(g) = \text{Dirichlet}(g; \alpha H_1, ..., \alpha H_r).$$

$\square$