Stat 544 Outline
Spring 2008

Steve Vardeman
Iowa State University

August 13, 2009

Abstract

This outline summarizes the main points of the course lectures.

Contents

1 Bayes Statistics: What? Why? A Worry ... ? How?

In this initial qualitative introduction to Bayesian statistics, we'll consider four questions:

1. What is it?
2. Why use it?
3. What about a worry?
4. How does one implement it in practice?

1.1 What is Bayesian Statistics?

Standard probability-based statistical inference begins with (typically vector) **data** $\boldsymbol{Y}$ modeled as an observable random vector/variable. The distribution of $\boldsymbol{Y}$ is presumed to depend upon some (typically vector) **parameter** $\boldsymbol{\theta}$ that is unknown/unobservable, and potentially on some (typically vector) "**covariate**" $\boldsymbol{X}$ that is observed. The object is often to make plausibility statements about $\boldsymbol{\theta}$. Sometimes, one thinks of $\boldsymbol{Y}$ as comprised of two parts, that is

$$\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$$

where one first observes $\boldsymbol{Y}_1$ and also needs to make plausibility statements about $\boldsymbol{Y}_2$.

In any case, one supposes that the $\boldsymbol{\theta}$ distribution of $\boldsymbol{Y}$ is specified by some probability density

$$f(\boldsymbol{y}|\boldsymbol{\theta}; \boldsymbol{X}) \tag{1}$$

and this function then specifies an entire family of probability models for $\boldsymbol{Y}$, one for each different $\boldsymbol{\theta}$. A couple of comments are in order regarding the form (1). In the first place, $f(\boldsymbol{y}|\boldsymbol{\theta}; \boldsymbol{X})$ could be a *probability density* used to get probabilities by doing "ordinary" Riemann integration over some part of $\Re^k$, or it could be a *probability mass function* used to get probabilities by adding over some discrete set of points $\boldsymbol{y}$, or it could be some *combination* of the two, used to get probabilities by doing Riemann integration over some coordinates of $\boldsymbol{y}$ while adding over values of the other coordinates of $\boldsymbol{y}$. Secondly, since one takes the values of the covariates as known/fixed, we will typically not bother to display the dependence of (1) on $\boldsymbol{X}$.

"**Classical**" statistical inference treats the density (with the observed data $\boldsymbol{y}$ plugged in)

$$L(\boldsymbol{\theta}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta})$$

as a (random) function of $\boldsymbol{\theta}$ called the **likelihood function** and uses it alone to guide inference/data analysis. Formally, "**Bayes**" statistical inference adds to the model assumptions embodied in (1) a model assumption on $\boldsymbol{\theta}$, that says

there is a density $g(\boldsymbol{\theta})$ that specifies a "**prior" distribution** for $\boldsymbol{\theta}$. This is intended to describe a "pre-data" view of the parameter. It too can be a probability density, a probability mass function, or a combination of the two. (More generally, it can in some cases simply be a non-negative function of $\boldsymbol{\theta}$, but more on that in a bit.)

In the standard/most easily understood case where $g(\boldsymbol{\theta})$ does specify a probability distribution for $\boldsymbol{\theta}$, the product

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$$

specifies a joint distribution for $(\boldsymbol{Y}, \boldsymbol{\theta})$. This in turn means that the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is specified by the conditional density

$$g(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}, \boldsymbol{\theta})}{\int f(\boldsymbol{y}, \boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{2}$$

(where the "integral" in the denominator of (2) is a Riemann integral, a sum, or some combination of the two). Of course, the denominator of (2) is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) \equiv \int f(\boldsymbol{y}, \boldsymbol{\theta})\, d\boldsymbol{\theta}$$

which is NOT a function of $\boldsymbol{\theta}$. Thus the posterior density $g(\boldsymbol{\theta}|\boldsymbol{y})$ is a function of $\boldsymbol{\theta}$ proportional to

$$f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}) = L(\boldsymbol{\theta})\, g(\boldsymbol{\theta})$$

and Bayes statistical inference is based on the notion that it is this product that should be the basis of plausibility statements about $\boldsymbol{\theta}$ (and in the case that only $\boldsymbol{Y}_1 = \boldsymbol{y}_1$ is observed, that the product $f(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$ should be the basis of all plausibility statements about $\boldsymbol{Y}_2$ and/or $\boldsymbol{\theta}$).

Notice, that it is sufficient but not always necessary that $g(\boldsymbol{\theta})$ be a density for the product $f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$ to be proportional to a density for $\boldsymbol{\theta}$ or for $f(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$ to be proportional to a joint density for $\boldsymbol{Y}_2$ and $\boldsymbol{\theta}$. That is, sometimes $g(\boldsymbol{\theta})$ can fail to be a density because it has an infinite "integral" and yet $f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$ or $f(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$ be perfectly useful (after normalization) as a density for $\boldsymbol{\theta}$ or $(\boldsymbol{Y}_2, \boldsymbol{\theta})$. In this case, it is common to say that $g(\boldsymbol{\theta})$ specifies an "**improper prior**" (a prior "distribution" that has total mass not 1, but rather ∞).

The **Bayes Paradigm** is then:

> All plausibility statements about $\boldsymbol{\theta}$ are based on a product
>
> $$f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}) = L(\boldsymbol{\theta})\, g(\boldsymbol{\theta})$$
>
> –and in the case that only $\boldsymbol{Y}_1 = \boldsymbol{y}_1$ is observed, plausibility statements about $\boldsymbol{Y}_2$ and/or $\boldsymbol{\theta}$ are based on a product
>
> $$f(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$$
>
> – the first of which specifies a "posterior distribution" for $\boldsymbol{\theta}$, the second of which specifies a joint predictive posterior/posterior distribution for $(\boldsymbol{Y}_2, \boldsymbol{\theta})$.

1.2 Why Use Bayesian Statistics?

There are at least 3 kinds of answers to the "why?" question concerning Bayes statistical methods. These are

1. philosophical answers,
2. optimality/decision theoretic answers, and
3. pragmatic answers.

Real "card-carrying" philosophical Bayesians argue that the only rationally coherent way of making statistical inferences is through the use of the Bayes Paradigm. The early parts of most books on Bayes inference provide these kinds of arguments. I'm not terribly interested in them. You should probably have a look at one or two such discussions.

Optimality/decision theory arguments for Bayes methods are based on minimization of expected costs. That is, supposes that

- $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, a **parameter space**,
- there are possible "actions" $a \in \mathcal{A}$ (an **action space**),
- associated with each pair $(\boldsymbol{\theta}, a)$ there is some "**loss**" $L(\boldsymbol{\theta}, a) \geq 0$,
- actions may be chosen on the basis of $\boldsymbol{Y} \sim f(\boldsymbol{y}|\boldsymbol{\theta})$ a distribution over some observation space $\mathcal{Y}$, that is, there are "**decision rules**" $\delta : \mathcal{Y} \to \mathcal{A}$

Then there are theorems that say things roughly like "Essentially only decision rules $\delta_g(\boldsymbol{y})$ that for some prior specified by g have the form

$$\delta_g(\boldsymbol{y}) = \text{an } a \text{ minimizing } \int L(\boldsymbol{\theta}, a) f(\boldsymbol{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{3}$$

can be any good in terms of

$$E_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \delta(\boldsymbol{Y})) = \int L(\boldsymbol{\theta}, \delta(\boldsymbol{y})) f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y} \quad ,$$

the expected loss function." Notice that $\int L(\boldsymbol{\theta}, a) f(\boldsymbol{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is proportional to the "posterior mean loss of action a," that is,

$$\int L(\boldsymbol{\theta}, a) f(\boldsymbol{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} = f_{\boldsymbol{Y}}(\boldsymbol{y}) \int L(\boldsymbol{\theta}, a) g(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

so $\delta_g(\boldsymbol{y})$ of the form (3) is an action that minimizes the (g) posterior expected loss (and is called a "Bayes rule" for prior g).

As a bit of a digression, it's worth noting that most philosophical Bayesians do not like optimality arguments for Bayes procedures (or at least do not find them compelling). This is because an expected loss $\mathrm{E}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \delta(\boldsymbol{Y}))$ involves an integration/averaging over the observation space $\mathcal{Y}$. A philosophical Bayesian

would question the relevance of averaging over outcomes that one knows did not occur ... that is, once $\boldsymbol{Y} = \boldsymbol{y}$ is known, such a person would argue that the only probability structure that is at all relevant is $g(\boldsymbol{\theta}|\boldsymbol{y})$.

The third kind of answer to the "why?" question of Bayesian statistics is purely pragmatic. The Bayesian paradigm provides an almost alarmingly simple and unified framework for statistical analysis. There is the model for the data (the likelihood) and the prior that give a joint distribution for "everything" (data, parameters, and future observations) that in turn gives a conditional (posterior) for everything that is not observed given everything that is observed. End of story. "All" one has to do is describe/understand/summarize the posterior. All of statistical inference has been reduced to probability calculations *within a single probability model*.

In contrast to "classical" statistics with its *family* of probability models indexed by $\boldsymbol{\theta}$, and the seeming necessity of doing "custom development" of methods for each different family and each different inference goal (estimation, testing, prediction, etc.), Bayesian statistics takes essentially the same approach to all problems of inference. (Bayesians might go so far as to say that while there is a well-defined "Bayes approach" to inference, there really is no corresponding classical or non-Bayesian "approach"!) Further, recent advances in Bayesian computation have made it possible to implement sensible Bayes solutions to statistical problems that are highly problematic when attacked from other vantage points. These are problems with particularly complicated data models, and especially ones where $\boldsymbol{\theta}$ is of high dimension. For example, maximum likelihood for a 100-dimensional $\boldsymbol{\theta}$ involves optimization of a function of 100 variables ... something that (lacking some kind of very specific helpful analytic structure) is often numerically difficult-to-impossible. In the same problem, modern Bayesian computation methods can make implementation of the Bayes paradigm almost routine.

1.3 A Worry ...?

Possibly the most worrisome feature of the Bayes paradigm is that the posterior distribution specified by $g(\boldsymbol{\theta}|\boldsymbol{y})$ (or $g(\boldsymbol{\theta}, \boldsymbol{y}_2|\boldsymbol{y}_1)$) of course depends upon the choice of prior distribution, specified by $g(\boldsymbol{\theta})$. Change the form of the prior and the final inferences change. This obvious point has long been a focal point of debate between philosophical Bayesians and anti-Bayesians. Anti-Bayesians have charged that this fact makes Bayes inferences completely "subjective" (a serious charge in scientific contexts). Bayesians have replied that in the first place "objectivity" is largely an illusion, and besides, the choice of prior is a modeling assumption in the same class as the modeling choice of a likelihood, that even anti-Bayesians seem willing to make. Anti-Bayesians reply "No, a likelihood and a prior are very different things. A likelihood is something that describes in probabilistic terms what reality generates for data. In theory at least, its appropriateness could be investigated through repetitions of data collection. Everyone admits that a prior exits only in one's head. Putting these two different kinds of things into a single probability model is not sensi-

ble." Bayesians reply that doing so is the only way to be logically consistent in inferences. And so the debate has gone ...

Two developments have to a large extent made this debate seem largely irrelevant to most onlookers. In the first place, as the probability models that people wish to use in data analysis have grown more and more complicated, the distinction between what is properly thought of as a model parameter and what is simply some part of a data vector that will go unobserved has become less and less clear. To many pragmatists, there are simply big probability models with some things observable and some things unobservable. To the extent that "Bayes" methods provide a way to routinely handle inference in such models, pragmatists are willing to consider them without taking sides in a philosophical debate.

The second development is that Bayesians have put a fair amount of work into the search for "flat" or "diffuse" or "objective" or "non-informative" or "robust" priors (or building blocks for priors) that tend to give posteriors leading to inferences similar to those of "classical"/non-Bayesian methods in simple problems. The idea is then that one could then hope that when these building blocks are used in complicated problems, the result will be inferences that are "like" classical inferences and do not depend heavily on the exact forms used for the priors, i.e. perform reasonably, regardless of what the parameter actually is. (An extreme example of an "informative" prior *lacking* this kind of "robustness" is one that says that with prior probability 1, $\theta = 13$. The posterior distribution of θ given $\boldsymbol{Y} = \boldsymbol{y}$ says that with posterior probability 1, $\theta = 13$. This is fine as long as the truth is that $\theta \approx 13$. But if the prior is badly wrong, the Bayes inference will be badly wrong.)

1.4 How Does One Implement the Bayesian Paradigm?

Conceptually, the Bayes paradigm is completely straightforward. Prior and likelihood are "multiplied" to produce something proportional to the posterior. Nothing could be simpler. But the practical problem is making sense of what one ends with. The questions become "What does a distribution specified by

$$f\left(\boldsymbol{y}|\boldsymbol{\theta}\right) g\left(\boldsymbol{\theta}\right) \quad \text{or} \quad f\left(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{\theta}\right) g\left(\boldsymbol{\theta}\right) \tag{4}$$

look like? What are posterior probabilities that it specifies? What are (posterior) means and standard deviations of the unobserved quantities?"

Except for very special circumstances where ordinary freshman/sophomore pencil-and-paper calculus works, making sense of a posterior specified by (4) is a matter of numerical analysis. But numerical analysis (particularly integration) in any more than a very few (2 or 3) dimensions is problematic. (Asymptotic approximations are sometimes mentioned as a possible "solution" to this computational problem. But that possibility is illusory, as large sample approximations for Bayes methods turn out to be fundamentally non-Bayesian (the prior really washes out of consideration for large samples) and it is, after all, the non-asymptotic behavior that is of real interest.) So it might seem

that the discussion has reached an impasse. While the paradigm is attractive, actually using it to do data analysis seems typically impossible.

But there is another way. The basic insight is that one doesn't have to *compute with* form (4) if one can *simulate from* form (4). Armed with a large number of realizations of simulations from a posterior, one can do simple arithmetic to approximate probabilities, moments, etc. as descriptors of the posterior. The first impulse would be to look for ways of drawing iid observations from the posterior. Sometimes that can be done. But by far the most powerful development in Bayesian statistics has been methods for doing not iid simulation from a posterior, but rather appropriate so-called "Markov Chain Monte Carlo" simulation. This is finding and using a suitable Markov Chain whose state space is the set of θ or $(\boldsymbol{y}_2, \boldsymbol{\theta})$ receiving positive posterior probability and whose empirical distribution of states visited for long runs of the chain approximates the posterior.

People with superior computing skills often program their own MCMC simulations. At the present time, the rest of us typically make use of a free Bayes simulation package called `WinBUGS`. You are welcome to use any means at your disposal to do computing in Stat 544. In practice, that is likely to mean some combination of `WinBUGS` and `R` programming.

2 Some Simulation Methods Useful in Bayesian Computation

There are a number of basic methods of generating realizations from standard simple distributions discussed in Stat 542 that begin from the assumption that one has available a stream of iid $\mathrm{U}(0,1)$ realizations. For example, if $U_1, U_2, \ldots$ are such iid uniform realizations

1. $F^{-1}(U)$ for a univariate cdf F has distribution F,
2. $-\ln(U)$ is exponential with mean 1,
3. $\max\left[\text{integers } j \geq 0 |\ -\sum_{i=1}^{j} \ln(U) < \lambda\right]$ is Poisson(λ),
4. $I[U < p]$ is Bernoulli(p),
5. $\sum_{i=1}^{n} I[U_i < p]$ is Binomial(n, p),
6. $Z_1 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2)$ and $Z_2 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$ are iid $\mathrm{N}(0,1)$,

and so on.

In the following introductory discussion, we consider several much more general simulation methods that are widely useful in Bayesian computation, namely

1. rejection sampling,

2. Gibbs (or more properly, successive substitution) sampling,

3. slice sampling,

4. the Metropolis-Hastings algorithm, and

5. Metropolis-Hastings-in-Gibbs algorithms.

The last four of these are MCMC algorithms. Later in the course we will discuss some theory of Markov Chains and why one might expect the MCMC algorithms to work. This initial introduction will be simply concerned with what these methods are and some aspects of their use.

Throughout this discussion, we will concern ourselves with simulation from some distribution for a vector $\boldsymbol{\eta}$ that is specified by a "density" that is proportional to a function

$$h(\boldsymbol{\eta})$$

We will not need to assume that h has been normalized to produce integral 1 and therefore already be a density. The fact that we don't have to know the integral of $h(\boldsymbol{\eta})$ is an essential point for practical Bayes computation. In Bayesian applications of this material, most often $\boldsymbol{\eta}$ will be either $\boldsymbol{\theta}$ or $(\boldsymbol{\theta}, \mathbf{Y}_2)$, the unknown parameter vector or the parameter vector and some future (possibly vector) observation and $h(\boldsymbol{\eta})$ will be respectively either $f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$ or $f(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$ and computation of the integral may not be feasible.

2.1 The Rejection Algorithm

Suppose that I can identify a *density* $p(\boldsymbol{\eta})$ (of the same type as a normalized version of $h(\boldsymbol{\eta})$) from which I know how to simulate, and such that

1. $p(\boldsymbol{\eta}) = 0$ implies that $h(\boldsymbol{\eta}) = 0$ so that the distribution specified by $p(\boldsymbol{\eta})$ has support at least as large as that of the distribution specified by of $h(\boldsymbol{\eta})$, and

2. one knows a finite upper bound M for the ratio

$$\frac{h(\boldsymbol{\eta})}{p(\boldsymbol{\eta})}$$

(this is essentially a requirement that the $p(\boldsymbol{\eta})$ tails be at least as heavy as those for $h(\boldsymbol{\eta})$ and that one can do the (pencil-and-paper or numerical) calculus necessary to produce a numerical value for M).

Then it is a standard Stat 542 argument to establish that the following works to produce $\boldsymbol{\eta} \sim h(\boldsymbol{\eta})$ (we'll henceforth abuse notation and write "$\sim$" when we mean that the variable on the left has a distribution with density *proportional to* the function on the right):

1. generate $\boldsymbol{\eta}^* \sim p(\boldsymbol{\eta})$,

2. generate $U \sim \mathrm{U}(0,1)$ independent of $\boldsymbol{\eta}^*$, and

3. if
$$\frac{h(\boldsymbol{\eta}^*)}{p(\boldsymbol{\eta}^*)} \geq U \cdot M$$
then set $\boldsymbol{\eta} = \boldsymbol{\eta}^*$, otherwise return to step 1.

Notice that if I can use the rejection algorithm (repeatedly) to create iid realizations $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3, \ldots, \boldsymbol{\eta}_N$ I can use (sample) properties of

$$\text{the empirical distribution of } \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3, \ldots, \boldsymbol{\eta}_N \tag{5}$$

to approximate properties of the distribution specified by $h(\boldsymbol{\eta})$.

An application of this algorithm most naturally relevant to Bayes calculation is that where $p(\boldsymbol{\eta})$ is $g(\boldsymbol{\theta})$ and $h(\boldsymbol{\eta})$ is $L(\boldsymbol{\theta})\, g(\boldsymbol{\theta})$. In this case the ratio $h(\boldsymbol{\eta}^*)/p(\boldsymbol{\eta}^*)$ is simply $L(\boldsymbol{\theta}^*)$, and if one can bound $L(\boldsymbol{\theta})$ by some number M (for example because an MLE of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, can be found and one can take $M = L\left(\hat{\boldsymbol{\theta}}\right)$), the rejection algorithm becomes:

> Generate $\boldsymbol{\theta}^*$ (a "proposal" for $\boldsymbol{\theta}$) from the prior distribution and accept that proposal with probability $L(\boldsymbol{\theta}^*)/M$, otherwise generate another proposal from the prior ...

This may initially seem like a natural and general solution of the Bayes computation problem. But it is not. Both in theoretical and operational terms, there are problems where it is not possible to find a bound for the likelihood. And more importantly (particularly in problems where $\boldsymbol{\theta}$ is high-dimensional) even when a bound for the likelihood can be identified, the part of the parameter space where the likelihood is large can be so "small" (can get such tiny probability from the prior) that the acceptance rate for proposals is so low as to make the algorithm unusable in practical terms. (A huge number of iterations and thus huge computing time would be required in order to generate a large sample of realizations.)

The nature of the typical failure of the rejection algorithm in high-dimensional Bayes computation provides qualitative motivation for the MCMC algorithms that *can* be successful in Bayes computation more generally. Rejection sampling from a posterior would involve iid proposals that take no account of any "success" earlier proposals have had in landing in regions where $h(\boldsymbol{\eta})$ is large. It would seem like one might want to somehow "find a place where $h(\boldsymbol{\eta})$ is large and move around in $\boldsymbol{\eta}$-space typically generating realizations "near" or "like" ones that produce large $h(\boldsymbol{\eta})$". This kind of thinking necessarily involves algorithms that make realized $\boldsymbol{\eta}$'s dependent. It is essential to the success of modern Bayes analysis that there are ways other than iid sampling (like the next four MCMC algorithms) to create (5) with sample properties approximating those of the distribution specified by $h(\boldsymbol{\eta})$.

2.2 Gibbs (or Successive Substitution) Sampling

Suppose now that $\boldsymbol{\eta}$ is explicitly k-dimensional or is divided into k pieces/subvectors (that may or may not each be 1-dimensional), that is, write

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_k)$$

Then with some starting vector

$$\boldsymbol{\eta}^0 = (\eta_1^0, \eta_2^0, \ldots, \eta_k^0)$$

for $j = 1, 2, \ldots$ a Gibbs sampler

1. samples η_1^j from $h\left(\cdot, \eta_2^{j-1}, \eta_3^{j-1}, \ldots, \eta_k^{j-1}\right)$
2. samples η_2^j from $h\left(\eta_1^j, \cdot, \eta_3^{j-1}, \ldots, \eta_k^{j-1}\right)$
3. samples η_3^j from $h\left(\eta_1^j, \eta_2^j, \cdot, \eta_4^{j-1}, \ldots, \eta_k^{j-1}\right)$

$\vdots$

$(k-1)$. samples η_{k-1}^j from $h\left(\eta_1^j, \eta_2^j, \ldots, \eta_{k-2}^j, \cdot, \eta_k^{j-1}\right)$, and

k. samples η_k^j from $h\left(\eta_1^j, \eta_2^j, \ldots, \eta_{k-1}^j, \cdot\right)$

in order to create $\boldsymbol{\eta}^j$ from $\boldsymbol{\eta}^{j-1}$.

Under appropriate circumstances, for large N, at least approximately

$$\boldsymbol{\eta}^N \sim h \tag{6}$$

and theoretical properties of the h distribution can be approximated using sample properties of

$$\left\{\boldsymbol{\eta}^{B+1}, \boldsymbol{\eta}^{B+2}, \ldots, \boldsymbol{\eta}^N\right\} \tag{7}$$

(for B a number "burn-in" iterations disregarded in order to hopefully mitigate the effects of an unfortunate choice of starting vector).

Use of this algorithm requires that one be able to make the random draws indicated in each of the steps 1 through k. This is sometimes possible because the indicated "sections" of the function h (h with all but one η_l held fixed) are recognizable as standard densities. Sometimes more clever methods are needed, like use of rejection algorithm or the "slice sampling" algorithm we will discuss next.

Why one might expect the "Gibbs sampler" to work under fairly general circumstances is something that we will discuss later in the term, as an application of properties of Markov Chains. For the time being, I will present a very small numerical example in class, and then point out what can "go wrong"

in the sense of (6) failing to hold and the empirical properties of (7) failing to approximate properties of h.

The principle failings of the Gibbs sampler occur when there are relatively isolated "islands of probability" in the distribution described by h, leading to "poor mixing" of the record of successive $\boldsymbol{\eta}^j$'s. Tools for detecting the possibility that the output of the Gibbs algorithm can't be trusted to represent h include:

1. making and comparing summaries of the results for several "widely dispersed" starts for the algorithm (different starts producing widely different results is clearly a bad sign!),

2. making and interpreting "history plots" and computing serial correlations for long runs of the algorithm (obvious jumps on the history plots and important high order serial correlations suggest that the Gibbs output may not be useful), and

3. the Brooks-Gelman-Rubin statistic and corresponding plots.

As the term goes along, we will discuss these and some of their applications. At this point we only note that all are available in `WinBUGS`.

2.3 Slice Sampling

The Gibbs sampling idea can be used to sample from a 1-dimensional continuous distribution. In fact, `WinBUGS` seems to use this idea (called "slice sampling") to do its 1-dimensional updates for non-standard distributions of bounded support (i.e. where the density is 0 outside a finite interval). The "trick" is that in order to sample from a 1-dimensional

$$h(\eta)$$

I implicitly invent a convenient 2-dimensional distribution for (η, V) and do what amounts to Gibbs sampling from this distribution to produce

$$\left(\eta^0, V^0\right), \left(\eta^1, V^1\right), \left(\eta^2, V^2\right), \ldots, \left(\eta^N, V^N\right)$$

and then for large N use η^N as a simulated value for η.

The slice sampling algorithm begins with some starting vector

$$\left(\eta^0, V^0\right)$$

and then for $j = 1, 2, \ldots$ one

1. samples η^j from a distribution uniform on $\left\{\eta | h(\eta) \geq V^{j-1}\right\}$, and

2. samples V^j from the Uniform$\left(0, h\left(\eta^j\right)\right)$ distribution

in order to create $\left(\eta^j, V^j\right)$ from $\left(\eta^{j-1}, V^{j-1}\right)$.

Slice sampling is the Gibbs sampler on a distribution that is uniform on

$$\{(\eta, v)\,|v < h(\eta)\} \subset \Re^2$$

The only difficult part of implementing the algorithm is figuring out how to accomplish step 1. Sometimes it's possible to do the algebra necessary to identify the set of η's indicated in step 1. When it is not, but I know that $h(\eta)$ is positive only on a finite interval $[a, b]$, I can instead generate iid $\mathrm{U}(a, b)$ realizations, checking the corresponding values of h until I get one larger than V^{j-1}.

It is worth noting that at least in theory (whether the following is practically efficient is a separate question), the restriction of slice sampling to cases where $h(\eta)$ is known to be positive only on a finite interval $[a, b]$ is not really intrinsic. That is, one may define a smooth strictly monotone transformation $\gamma : \Re \to (0, 1)$, use slice sampling to sample from the distribution $\gamma(\eta)$, and then apply the inverse transform to get realizations of η from $h(\eta)$. Take, for example, the transformation

$$\gamma(\eta) = \frac{1}{1 + \exp(-\eta)}$$

with inverse transformation

$$\gamma^{-1}(t) = -\ln\left(\frac{1}{t} - 1\right) = \ln\left(\frac{t}{1-t}\right)$$

that has derivative

$$\frac{d}{dt}\gamma^{-1}(t) = \frac{1}{t(1-t)}$$

If η has pdf proportional to $h(\eta)$, then $\gamma(\eta)$ has pdf on $(0, 1)$ proportional to the function of t

$$\frac{h\left(\gamma^{-1}(t)\right)}{t(1-t)} \tag{8}$$

and one can do slice sampling for $\gamma(\eta)$ as indicated above based on (8) and apply γ^{-1} to the result to simulate from h.

Together, the rejection algorithm and slice sampling (each with its own limitations) make two ways of implementing one of the k Gibbs updates for the common cases where the indicated density is not one of a standard form (i.e. is not one for which simulation methods are well known).

2.4 The Metropolis-Hastings Algorithm

A second basic MCMC algorithm alternative to or complementary to the Gibbs algorithm is the so-called Metropolis-Hastings algorithm. It begins from some starting vector $\boldsymbol{\eta}^0$. Then for $j = 1, 2, \ldots$

1. let $J_j(\eta'|\eta)$ specify for each η a distribution (for η') over the part of Euclidean space where $h(\eta') > 0$, called the "jumping" or "proposal" distribution for the jth iteration of updating (a distribution that I know how to simulate from), and generate

$$\eta^{j*} \sim J_j\left(\cdot|\eta^{j-1}\right)$$

as a proposal or candidate for η^j,

2. compute

$$r_j = \frac{h\left(\eta^{j*}\right)/J_j\left(\eta^{j*}|\eta^{j-1}\right)}{h\left(\eta^{j-1}\right)/J_j\left(\eta^{j-1}|\eta^{j*}\right)}$$

and generate

$$W_j \sim \text{Bernoulli}\left(\min\left(1, r_j\right)\right)$$

and,

3. take

$$\eta^j = W_j\eta^{j*} + \left(1 - W_j\right)\eta^{j-1}$$

(i.e. one jumps from η^{j-1} to the proposal η^{j*} with probability $\min(1, r_j)$ and otherwise stays put at η^{j-1}).

In contrast to the Gibbs algorithm, this algorithm has the great virtue of requiring only simulation from the proposal distribution (and not from non-standard conditionals of h). These can be chosen to be "standard distributions" with well-known fast simulation methods

The situation where each

$$J_j(\eta'|\eta) = J_j(\eta|\eta')$$

(i.e. the jumping distributions are symmetric) is especially simple and gives the variant of the algorithm known simply as the "Metropolis Algorithm." Note too that the proposal distributions may depend upon the iteration number and the current iterate, η^{j-1}. Strictly speaking, they may not depend upon any more of the history of iterates beyond η^{j-1}. However, it is very common practice to violate this restriction early in a run of an MCMC algorithm, letting the algorithm "adapt" for a while before beginning to save iterates as potentially representing h. The idea of this tuning of the algorithm early in a run is to both "get from the starting vector to the 'important part of the distribution'" and to "tune the parameters of the jumping distributions to make the algorithm efficient" (i.e. make the r_j's tend to be large and create frequent jumps).

2.5 Metropolis-Hastings-in-Gibbs Algorithms

The Gibbs sampler is attractive in that one can use it to break a large simulation problem down into small, manageable chunks, the updating of the k subvectors/pieces of $\boldsymbol{\eta}$. It requires, however, methods of sampling from each of

the (h) conditional distributions of an η_l given the rest of the $\boldsymbol{\eta}$ vector. This requires the recognition of each conditional as of some convenient parametric form, or the use of the rejection or slice sampling algorithm, or yet something else. Sometimes it's not so easy to find a suitable method for sampling from each of these conditionals.

The Metropolis-Hastings algorithm does not require sampling from any distribution defined directly by h, but rather only from proposal distributions that the analyst gets to choose. But, at least as described to this point, it seems that one must deal with the entirety of the vector $\boldsymbol{\eta}$ all at once. But as it turns out, this is not necessary. One may take advantage of the attractive features of *both* the Gibbs and Metropolis-Hastings algorithms in a single MCMC simulation. That is, there are Metropolis-Hastings-in-Gibbs algorithms.

That is, in the Gibbs sampling setup, for the update of any particular subvector η_l, one may substitute a "Metropolis-Hastings step." In place of

$$\text{sampling } \eta_l^j \text{ from } h\left(\eta_1^j, \ldots, \eta_{l-1}^j, \cdot, \eta_{l+1}^{j-1}, \ldots, \eta_k^{j-1}\right)$$

one may

1. let $J_{lj}\left(\eta_l' | \eta_1, \ldots, \eta_{l-1}, \eta_l, \eta_{l+1}, \ldots, \eta_k\right)$ specify for each $\left(\eta_1, \ldots, \eta_{l-1}, \eta_l, \eta_{l+1}, \ldots, \eta_k\right)$ a distribution (for η_l') over the part of Euclidean space where the function of η_l', $h\left(\eta_1, \ldots, \eta_{l-1}, \eta_l', \eta_{l+1}, \ldots, \eta_k\right) > 0$, and generate

   $$\eta_l^{j*} \sim J_{lj}\left(\cdot | \eta_1^j, \ldots, \eta_{l-1}^j, \eta_l^{j-1}, \eta_{l+1}^{j-1}, \ldots, \eta_k^{j-1}\right)$$

   as a proposal or candidate for η_l^j,

2. compute

   $$\begin{aligned} r_{lj} &= \frac{h\left(\eta_1^j, \ldots, \eta_{l-1}^j, \eta_l^{j*}, \eta_{l+1}^{j-1}, \ldots, \eta_k^{j-1}\right)}{h\left(\eta_1^j, \ldots, \eta_{l-1}^j, \eta_l^{j-1}, \eta_{l+1}^{j-1}, \ldots, \eta_k^{j-1}\right)} \\ &\quad \times \frac{J_{lj}\left(\eta_l^{j-1} | \eta_1^j, \ldots, \eta_{l-1}^j, \eta_l^{j*}, \eta_{l+1}^{j-1}, \ldots, \eta_k^{j-1}\right)}{J_{lj}\left(\eta_l^{j*} | \eta_1^j, \ldots, \eta_{l-1}^j, \eta_l^{j-1}, \eta_{l+1}^{j-1}, \ldots, \eta_k^{j-1}\right)} \end{aligned}$$

   and generate

   $$W_{lj} \sim \text{Bernoulli}\left(\min\left(1, r_{lj}\right)\right)$$

   and,

3. take

   $$\eta_l^j = W_{lj}\eta_l^{j*} + \left(1 - W_{lj}\right)\eta_l^{j-1}$$

   (i.e. one jumps from η_l^{j-1} to the proposal η_l^{j*} with probability $\min\left(1, r_{lj}\right)$ and otherwise stays put at η_l^{j-1}).

This kind of algorithm is probably the most commonly used MCMC algorithm in modern Bayesian computation, at least where people do their own programming instead of relying on `WinBUGS`.

3 The Practice of Modern Bayes Inference 1: Some General Issues

We now take for granted computing algorithms for approximating a posterior distribution via MCMC and consider a series of issues in the practical application of the Bayes paradigm.

3.1 MCMC Diagnostics

For purposes of being in a position to detect whether there are potentially problems with "poor mixing"/"islands of probability" in a MCMC simulation from a posterior (or posterior/predictive posterior) distribution, it is standard practice to:

1. pick several widely dispersed and perhaps even "unlikely under the posterior" starting vectors for posterior MCMC iterations,
2. run several (say m) chains in parallel from the starting points in 1.,
3. monitor these several chains until "transient" effects of the starting vectors wash out and they start to have "similar" behaviors, i.e. monitor them until they "burn in," and
4. use for inference purposes only simulated $\boldsymbol{\theta}$ and/or $\boldsymbol{Y}_2$ values coming from iterations after burn-in.

The question is how one is to judge if and when burn-in has taken place.

A fairly qualitative way of trying to assess burn-in is to visually monitor "history plots" (of all parallel chains on a given plot) of individual coordinates of $\boldsymbol{\theta}$ and/or $\boldsymbol{Y}_2$. (These are simply plots of values of the coordinate against iteration number, with consecutive points for a given chain connected by line segments.) `WinBUGS` allows one to run multiple chains and make such plots with each chain getting a different color on the plot. One simply waits until these look "alike" to the statistically practiced eye.

A more or less quantitative tool for judging when burn-in has occurred is the "Gelman-Rubin statistic" and related plots, implemented in `WinBUGS` in a variant form called the "`BGR`" (Brooks-Gelman-Rubin) statistic and plots. The original version of the idea (discussed in the textbook) is the following. Let ψ stand for some coordinate of $\boldsymbol{\theta}$ and/or $\boldsymbol{Y}_2$ (possibly after "transformation to normality"). Beginning after some number of iterations of MCMC simulations, let

$$\psi_i^j = j\text{th saved iterate of } \psi \text{ in chain } i \text{ for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n$$

If burn-in has occurred, I expect that the set of ψ_i^j obtained from each chain i will "look like" the set of ψ_i^j obtained from pooling across all chains. Ways

of measuring the extent to which this is true can be based on within-chain and grand means

$$\overline{\psi}_i = \frac{1}{n}\sum_{j=1}^{n}\psi_i^j \quad \text{and} \quad \overline{\psi}_\cdot = \frac{1}{m}\sum_{i=1}^{m}\overline{\psi}_i$$

and within-chain sample variances and a pooled version of these

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(\psi_i^j - \overline{\psi}_i\right)^2 \quad \text{and} \quad W = \frac{1}{m}\sum_{i=1}^{m}s_i^2$$

and a kind of between-chain variance

$$B = \frac{n}{m-1}\sum_{i=1}^{m}\left(\overline{\psi}_i - \overline{\psi}_\cdot\right)^2$$

W and B are, in fact, respectively the "One-Way ANOVA" error and treatment (within and between) mean squares from a One-Way analysis with "chains" as "treatments." The Gelman-Rubin statistic based on these quantities is

$$\hat{R}_n = \sqrt{\frac{n-1}{n} + \frac{1}{n}\left(\frac{B}{W}\right)}$$

If each chain's record begins to "look like" a random sample from the same distribution as $n \to \infty$, $\hat{R}_n$ should approach 1. If the records of the m chains "look different" one should expect $\hat{R}_n$ to stay larger than 1 with increasing n. (One plots $\hat{R}_n$ against n.) (Note also in passing that the ratio B/W is exactly the one-way ANOVA F statistic for this problem.)

The Brooks-Gelman modification of this idea implemented in `WinBUGS` is as follows. Let

$$\begin{aligned}
L_i^n &= \text{the lower 10\% point of the } n \text{ values } \psi_i^j \text{ (from chain } i\text{)} \\
U_i^n &= \text{the upper 10\% point of the } n \text{ values } \psi_i^j \text{ (from chain } i\text{)} \\
L^n &= \text{the lower 10\% point of the } nm \text{ values } \psi_i^j \text{ (from all chains)} \\
U^n &= \text{the upper 10\% point of the } nm \text{ values } \psi_i^j \text{ (from all chains)}
\end{aligned}$$

Then plotted versus n in `WinBUGS` are 3 quantities:

1. $\left(U^n - L^n\right)/\kappa$ plotted in green,
2. $\left(\frac{1}{m}\sum_{i=1}^{m}\left(U_i^n - L_i^n\right)\right)/\kappa$ plotted in blue, and
3. $\left(U^n - L^n\right) / \left(\frac{1}{m}\sum_{i=1}^{m}\left(U_i^n - L_i^n\right)\right)$ plotted in red.

The idea is that the value in red needs to approach 1 and the values plotted in green and blue need to stabilize. The constant κ used in 1. and 2. is chosen to make the largest plotted green or blue plotted value 1. The `WinBUGS`

manual says that "bins of 50 are used" and I believe that this means that the computation and plotting is done at multiples of 50 iterations.

Ideally, properly burned-in history plots look like patternless "white noise" (iid observations) plots. When instead they show (similar across the chains) behavior that might be characterized as "slow drift" one is faced with a situation where long MCMC runs will be necessary if there is any hope of adequately representing the posterior. In some sense, one has many fewer "observations" from the posterior than one has iterations. A "slowly drifting MCMC record" means that values of a coordinate of $\boldsymbol{\theta}$ and/or $\boldsymbol{Y}_2$ change only slowly. This can be measured in terms of how fast serial correlations in the MCMC records fall off with lag. For example, suppose θ_1 is the first coordinate of the parameter vector $\boldsymbol{\theta}$ and that the jth MCMC iterate of this variable is θ_1^j. One might compute the sample correlation between the first and second coordinates of ordered pairs

$$\left(\theta_1^j, \theta_1^{j+s}\right)$$

for $s = 1, 2, 3, \ldots$ (for j after burn-in) as a measure of "lag-s serial correlation" in the θ_1 record. Nontrivial positive serial correlations for large s are indicative of "slow drift"/"poor mixing" in the simulation and the necessity of long runs for adequate representation of the posterior.

3.2 Considerations in Choosing Priors

How does one choose a prior distribution? The answer to this question is obviously critical to Bayes analysis, and must be faced before one can even "get started" in an application. A couple of points are obvious at the outset. In the first place, a posterior can place probability on only those parts of a parameter space where the prior has placed probability. So unless one is absolutely "sure" that some subset of $\boldsymbol{\theta}$'s simply can not contain the actual parameter vector, it is dangerous to use a prior distribution that ignores that set of parameters. (Unless I am willing to take poison on the proposition that $\theta < 13$, I should not use a prior that places 0 probability on the event that $\theta \geq 13$.) Secondly, all things being equal, if several different choices of prior produce roughly the same posterior results (and in particular, if they produce results consistent with those derivable from non-Bayesian methods) any of those priors might be thought of as attractive from a "robustness" perspective.

3.2.1 "The Prior"

A real philosophical Bayesian would find the previous statement to be heretical-to-irrelevant. That is, for a card-carrying Bayesian, there is only one "true" prior, that reflects his or her carefully considered prior opinions about $\boldsymbol{\theta}$. This probability structure is unashamedly personal and beyond criticism on any other than logical or philosophical grounds. Bayesians have put a fair amount of effort into developing theory and tools for the "elicitation of prior beliefs" and would argue that the way one ought to get a prior is through the careful use

of these. While this logical consistency is in some respects quite admirable, I am unconvinced that it can really be pushed this far in a practical problem. However, you are invited to investigate this line of thought on your own. We will take more eclectic approaches in this set of notes.

3.2.2 Conjugate Priors

Before the advent of MCMC methods, there was a particular premium placed on priors for which one can do posterior calculations with pencil-and-paper calculus, and "conjugate" priors were central to applications of the Bayes paradigm. That is, some simple forms of the likelihood $L(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})$ themselves look like all or parts of a density for $\boldsymbol{\theta}$. In those cases, it is often possible to identify a simple prior $g(\boldsymbol{\theta})$ that when multiplied by $L(\boldsymbol{\theta})$ produces a function that by simple inspection can be seen to be "of the same family or form as $g(\boldsymbol{\theta})$." (For example, a Binomial likelihood multiplied by a Beta prior density produces a product proportional to a different Beta density.) When this is the case and the form of $g(\boldsymbol{\theta})$ is simple, posterior probability calculations can be done without resort to MCMC simulation. The jargon for this kind of nice interaction between the form of a likelihood and a convenient prior form is that the prior is a "conjugate" prior.

These days, conjugate priors are important not so much because they are the only ones for which posterior computations can be done (MCMC methods have removed the necessity of limiting consideration to posteriors that yield to pencil-and-paper calculus), but rather because the explicit formulas that they can provide often enlighten the search for priors that are minimally informative/robust. In fact, many useful "non-informative" prior distributions can be seen to be limits (as one sends parameters of a prior to some extreme) of conjugate priors. (For example, where the elements of $\boldsymbol{Y}$ are iid $\mathrm{N}(\mu, 1)$ variables, one conjugate prior for μ is the $\mathrm{N}\left(0, \sigma^2\right)$ distribution, and in some sense the "limit" of this prior as $\sigma^2 \to \infty$ is the "uniform on $\Re$" improper prior. This is in many respects an attractive non-informative choice for this problem.)

3.2.3 "Flat"/"Diffuse"/"Non-Informative"/"Robust" Priors

The notions of prior "diffuseness," "flatness," and "non-informativeness"/"robustness" are not really terribly concrete concepts. What one hopes to achieve in the search for priors that might be described in these ways is fairly clear: posterior distributions that behave sensibly no matter what be $\boldsymbol{\theta}$. But it is worth saying explicitly here, that whether a prior "looks" flat or diffuse is dependent upon the particular parameterization that one adopts, and thus whether a flat/diffuse choice of prior will function in a robust/non-informative way is *not* obvious from simply examining its shape.

For example, consider a hypothetical inference problem with parameter $p \in (0, 1)$. One "flat"/"diffuse" prior for a Bayes problem involving p would be a $\mathrm{U}(0, 1)$ prior for p. But an alternative parameterization for the problem might

be in terms of the log-odds

$$\gamma = \ln\left(\frac{p}{1-p}\right)$$

and a "flat" improper prior for γ is "uniform on $\Re$." These are not equivalent specifications. For example, the first says that the prior probabilities assigned to the intervals $(.5, .6)$ and $(.6, .7)$ are the same, while the second says that the (improper) prior weights assigned to these sets of p's are in the ratio

$$\frac{\ln\left(\frac{.6}{1-.6}\right) - \ln\left(\frac{.5}{1-.5}\right)}{\ln\left(\frac{.7}{1-.7}\right) - \ln\left(\frac{.6}{1-.6}\right)} = .9177$$

Whether either of these priors will function in a "non-informative" way in a Bayes analysis is not obvious from their qualitative "flatness"/"diffuseness" evident from simple inspection.

3.2.4 Jeffreys Priors

In the case that θ is 1-dimensional, there is a standard method due to H. Jeffreys for identifying a prior (or improper prior) that often turns out to be operationally "non-informative." That is this. Associated with a likelihood $f(\boldsymbol{y}|\theta)$ (differentiable in θ) is the Fisher Information (a function of θ)

$$\mathcal{I}_{\boldsymbol{Y}}(\theta) = \mathrm{E}_\theta\left(\frac{d}{d\theta}\ln f(\boldsymbol{Y}|\theta)\right)^2$$

It is well known that sometimes (but not always) the Fisher Information may also be computed as

$$-\mathrm{E}_\theta\frac{d^2}{d\theta^2}\ln f(\boldsymbol{Y}|\theta)$$

In any case, the Jeffreys prior for a Bayes analysis involving this likelihood is specified by

$$g(\theta) \propto \sqrt{\mathcal{I}_{\boldsymbol{Y}}(\theta)} \tag{9}$$

An especially attractive feature of this prescription is that it is invariant to monotone reparameterization. So one may speak of "the" Jeffreys prior for the problem without ambiguity. That is, for a monotone function $u(\theta)$, consider a second parameterization of this problem with parameter

$$\gamma = u(\theta)$$

With prior (say) pdf (9), Stat 542 transformation theorem material shows that γ has pdf proportional to

$$\sqrt{\mathcal{I}_{\boldsymbol{Y}}(u^{-1}(\gamma))}\left|\frac{1}{u'(u^{-1}(\gamma))}\right| \tag{10}$$

But the information in $\boldsymbol{y}$ about γ is (for $\dot{f}(\boldsymbol{Y}|\theta)$ the partial derivative of $f(\boldsymbol{Y}|\theta)$ with respect to θ)

$$\begin{aligned} \mathrm{E}_{u^{-1}(\gamma)}\left(\frac{d}{d\gamma}\ln f\left(\boldsymbol{Y}|u^{-1}(\gamma)\right)\right)^2 &= \mathrm{E}_{u^{-1}(\gamma)}\left(\frac{\dot{f}\left(\boldsymbol{Y}|u^{-1}(\gamma)\right)\frac{\partial}{\partial\gamma}u^{-1}(\gamma)}{f\left(\boldsymbol{Y}|u^{-1}(\gamma)\right)}\right)^2 \\ &= \mathcal{I}_{\boldsymbol{Y}}\left(u^{-1}(\gamma)\right)\frac{1}{\left(u'\left(u^{-1}(\gamma)\right)\right)^2} \end{aligned} \tag{11}$$

Clearly, the square root of rhs(11) is the pdf (10) that γ inherits from the assumption (9) that θ has a Jeffreys prior.

3.3 Considerations in Choice of Parametrization

The necessity of specifying a prior distribution for $\boldsymbol{\theta}$ and then sampling from a posterior for it probably causes one to think harder about the most convenient way to parameterize a model for $\boldsymbol{Y}$ than might otherwise be necessary. We proceed to make several observations about the issue of parameterization.

3.3.1 Identifiability

A basic requirement for sensible inference, Bayesian or non-Bayesian, is that any two different parameter vectors θ and θ' correspond to genuinely different distributions for $\boldsymbol{Y}$. But it is not impossible to fail to recognize that one has violated this basic sanity requirement. When this happens, MCMC simulations can behave in seemingly inexplicable ways.

For example, consider a mixture problem, where one is presented with iid observations, which each are $\mathrm{N}(\mu_1, 1)$ with probability α and $\mathrm{N}(\mu_2, 1)$ with probability $(1-\alpha)$. As just stated (with the implicit choice of parameter space $\Re\times\Re\times(0,1)$ for (μ_1, μ_2, α)) this model is not identifiable. The parameter vectors $(0, 1, .7)$ and $(1, 0, .3)$ produce the same distribution for the data. MCMC for "obvious" choices of prior in this problem will behave in what seems to be "odd" ways. One needs to somehow either reduce the parameter space to something like

$$\{(\mu_1, \mu_2, \alpha)\,|\mu_1 < \mu_2 \text{ and } 0 < \alpha < 1\} \tag{12}$$

and place a prior on that subset of $\Re^3$ or find an alternative parameterization. For example, one might think of

$$\mu_1 = \text{the smaller of the two means}$$

and set

$$\delta = \mu_2 - \mu_1$$

(so that $\mu_2 = \mu_1 + \delta$) and do the inference in terms of (μ_1, δ, α) rather than (μ_1, μ_2, α) directly. Note then that the parameter space becomes $\Re\times(0,\infty)\times(0,1)$ and that choosing a prior over this space seems less complicated than making a choice of one over (12).

3.3.2 Gibbs and Posterior Independence

In terms of efficiency/properly representing a posterior in as few iterations as possible, Gibbs-like algorithms work best when the subvectors of $\boldsymbol{\theta}$ being updated in turn are (according to the posterior) roughly independent. When the posterior portrays strong dependencies, the range of each update is in effect limited substantially by the dependence, and Gibbs algorithms tend to take very small steps through the parameter space and thus take a large number of iterations to "cover the parameter space" adequately.

This means that all other things being equal, for purposes of efficient computation, one prefers parameterizations with product parameter spaces, and that tend to produce likelihoods that as functions of $\boldsymbol{\theta}$ do not contribute to posterior dependencies. (To the extent that large sample loglikelihoods tend to be approximately quadratic with character determined by the corresponding Fisher information matrix, one prefers parameterizations with essentially diagonal Fisher information matrices.) And again for purposes of computational efficiency (at least if a prior is not going to be effectively "overwhelmed" by a "large sample" likelihood) priors of independence for such parameterizations seem most attractive.

This discussion suggests that at least from a computational standpoint, the parameter space (12) discussed above is less attractive than the $\Re \times (0,\infty) \times (0,1)$ product space associated with the second $((\mu_1, \delta, \alpha))$ parameterization. A second, very familiar, example relevant to this discussion is that of simple linear regression. The simple linear regression Fisher information matrix typically fails to be diagonal in the usual parameterization where the regression coefficients are the slope and y intercept (the mean value of y when $x = 0$). However, if instead of using raw values of covariates one centers them so that regression coefficients become the slope and the mean value of the response when the covariate is at its sample mean value ($\bar{x}$), this potential computational complication disappears.

3.3.3 Honoring Restrictions Without Restricting Parameters

The most convenient/straightforward way of specifying a high-dimensional prior distribution is by making an independence assumption and specifying only marginal distributions for coordinates of $\boldsymbol{\theta}$ on some product space. That makes parameter spaces like (12) that involve some restrictions in a product space problematic. There are at least 2 ways of getting around this unpleasantness. First, one might look for alternate parameterizations that simply avoid the difficulty altogether. (In the mixture example, this is the approach of using (μ_1, δ, α) instead of the original (μ_1, μ_2, α) parameterization.) A second possibility (that might not work in the mixture problem, but will work in other contexts) is to ignore the restrictions and use a prior of independence on a product space for purposes of running an MCMC algorithm, but to "post-process" the MCMC output, deleting from consideration vectors from any iteration whose vector violates the restrictions. For example, in a problem where a parameter vector $(p_1, p_2, p_3) \in (0,1)^3$ must satisfy the order restriction $p_1 \leq p_2 \leq p_3$, one

might adopt and use in an MCMC algorithm independent Beta priors for each p_i. After-the-fact using only those simulated values whose vectors (p_1, p_2, p_3) satisfy the order restrictions essentially then employs a prior with density proportional to the product of Beta densities *but restricted to the part of* $(0,1)^3$ *where the order restriction holds.* (Using `WinBUGS` and `R` this can be accomplished by saving the `WinBUGS` results using the `Coda` option, turning them into a text file, and loading the text file into `R` using the `Coda` package for post-processing.)

3.4 Posterior (Credible) Intervals

A posterior distribution for $\boldsymbol{\theta}$ (or for $(\boldsymbol{Y}_2, \boldsymbol{\theta})$) is often summarized by making representations of the corresponding marginal (posterior) distributions. For sake of discussion here, let η stand for some 1-dimensional element of $\boldsymbol{\theta}$ (or $(\boldsymbol{Y}_2, \boldsymbol{\theta})$). Upon finishing an MCMC simulation from a posterior one has a large number of realizations of η, say $\eta^1, \eta^2, \ldots, \eta^N$. These can be summarized in terms of a histogram, or in the case that η is a continuous variable, with some kind of estimated probability density (`WinBUGS` provides such density estimates). It is also common to compute and report standard summaries of these values, the sample mean, sample median, sample standard deviation, and so on.

Probably the most effective way of conveying where most of the posterior probability is located is through the making and reporting of posterior probability intervals, or so-called Bayesian "credible intervals." The simplest of these are based on (approximate) quantiles of the marginal posterior. That is, if

$$\eta_{.025} = \text{the .025 quantile of } \left\{\eta^1, \eta^2, \ldots, \eta^N\right\}$$

and

$$\eta_{.975} = \text{the .975 quantile of } \left\{\eta^1, \eta^2, \ldots, \eta^N\right\}$$

then the interval

$$[\eta_{.025}, \eta_{.975}]$$

encloses posterior probability .95 (at least approximately) and can be termed a 95% credible interval for η. It might be thought of as a Bayes alternative to a "classical" 95% confidence interval (though there is no guarantee at all that the method that produced it is anything like a 95% confidence procedure).

A theoretically better/smaller construction of credible sets is the "highest posterior density" (hpd) construction. That is, rather than using quantiles to identify a credible interval, one might look for a number c so that with $g(\eta|\boldsymbol{y})$ the posterior marginal density of η, the set

$$\{\eta | g(\eta|\boldsymbol{y}) > c\} \tag{13}$$

has posterior probability .95. That set is then the smallest one that has posterior probability content .95, and can be called the "95% highest posterior density credible set for η."

Unless $g(\eta|\boldsymbol{y})$ is unimodal, there is no guarantee that the hpd construction will produce an interval. And unless $g(\eta|\boldsymbol{y})$ has a simple analytic form, it may not be easy to identify the set (13). Further, while the use of quantiles to make credible intervals is invariant under monotone transformations of the parameter, the result of using the hpd construction is not. (This is really a manifestation of the same phenomenon that makes apparent "flatness" of a prior dependent upon the particular parameterization one adopts.) For these reasons, the quantile method of producing intervals is more common in practice than the hpd construction.

3.5 Bayes Model Diagnostics and Bayes Factors for Model Choice

Since a Bayes statistical model is simply an "ordinary" statistical model with the addition of a prior $g(\boldsymbol{\theta})$, any kind of "model checking" appropriate in a non-Bayesian context (e.g. residual plotting, etc.) is equally appropriate in the Bayes context. The new feature present in the Bayes context is what the prior does. Specifically Bayesian model checking is usually approached from the point of view of "posterior predictive distribution checking." That is, if the density of the observable $\boldsymbol{Y}$ is

$$f(\boldsymbol{y}|\boldsymbol{\theta})$$

let $\boldsymbol{Y}_{\text{new}}$ also have this density and (conditional on $\boldsymbol{\theta}$) be independent of the observable $\boldsymbol{Y}$. So the joint density of all of $(\boldsymbol{Y}, \boldsymbol{Y}_{\text{new}}, \boldsymbol{\theta})$ is proportional to

$$f(\boldsymbol{y}|\boldsymbol{\theta})\, f(\boldsymbol{y}_{\text{new}}|\boldsymbol{\theta})\, g(\boldsymbol{\theta})$$

and one can make posterior (to $\boldsymbol{Y} = \boldsymbol{y}$) simulations of $\boldsymbol{Y}_{\text{new}}$. One can then ask whether the data in hand, $\boldsymbol{y}$, look anything like the simulated values. Chapter 6 of the text discusses some ways of assessing this numerically and graphically. For my money, this strikes me as "stacked in favor of concluding that the Bayes analysis is OK." Roughly speaking, the posterior uncertainty in $\boldsymbol{\theta}$ will have the effect of making the posterior predictive distribution of $\boldsymbol{Y}_{\text{new}}$ more spread out than any single $f(\boldsymbol{y}|\boldsymbol{\theta})$ for a fixed $\boldsymbol{\theta}$. So it seems rare that one will get posterior predictive simulations that fail to "cover" the observed data, unless there is some huge blunder in the modeling or simulation.

A somewhat different question is how one might compare the appropriateness of several (either nested or un-nested) Bayes models for an observable $\boldsymbol{Y}$. So called "Bayes factors" have been offered as one means of doing this. Suppose m different models have densities $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)$ of the same type, where the parameters $\boldsymbol{\theta}_i$ take values (in possibly different) k_i-dimensional Euclidean spaces, $\Re^{k_i}$, and priors are specified by densities (improper densities are not allowed in this development) $g_i(\boldsymbol{\theta}_i)$. Each of these models produces a (marginal) density for $\boldsymbol{Y}$,

$$f_i(\boldsymbol{y}) = \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\, g_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i$$

(where, as usual, the indicated integral is a Riemann integral, a sum, or some combination of the two). One might then look at

$$BF_{i'i} = \frac{f_{i'}(\boldsymbol{y})}{f_i(\boldsymbol{y})} \tag{14}$$

as an appropriate statistic for comparing models i and i'.

In Neyman-Pearson testing of

$$\mathrm{H}_0\text{:the correct model is model } i \ (\boldsymbol{Y} \sim f_i)$$

versus

$$\mathrm{H}_\mathrm{a}\text{:the correct model is model } i' \ (\boldsymbol{Y} \sim f_{i'})$$

$BF_{i'i}$ is the optimal test statistic (is the "likelihood ratio"). Further, if one sets prior probabilities on *models* 1 through m, say $p_1, p_2, \ldots, p_m$ the posterior probability for model i is

$$\frac{p_i f_i(\boldsymbol{y})}{\sum_{l=1}^{m} p_l f_l(\boldsymbol{y})}$$

so that the posterior "odds ratio" for models i' and i is

$$\frac{p_{i'} f_{i'}(\boldsymbol{y})}{p_i f_i(\boldsymbol{y})} = \left(\frac{p_{i'}}{p_i}\right) BF_{i'i}$$

which is the prior odds ratio times the Bayes factor.

Notice that one is typically not going to be able to do the calculus necessary to compute the $f_i(\boldsymbol{y})$ needed to find Bayes factors. But often (especially because it is common to make independence assumptions between coordinates of $\boldsymbol{\theta}$ in specifying priors) it's easy to generate

$$\boldsymbol{\theta}_i^1, \boldsymbol{\theta}_i^2, \ldots, \boldsymbol{\theta}_i^n \text{ that are iid } g_i(\boldsymbol{\theta}_i)$$

Then the law of large numbers implies that

$$\frac{1}{n}\sum_{l=1}^{n} f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i^l\right) \xrightarrow{P} f_i(\boldsymbol{y})$$

from which one can get approximate values for Bayes factors.

How to interpret Bayes factors has been a matter of some dispute. One set of qualitative interpretations suggested by Jeffreys for a Bayes factor BF_{21} is

- $BF_{21} > 1$ favors model 2,
- $0 > \log_{10} BF_{21} > -\frac{1}{2}$ provides *minimal* evidence against model 2,
- $-\frac{1}{2} > \log_{10} BF_{21} > -1$ provides *substantial* evidence against model 2,
- $-1 > \log_{10} BF_{21} > -2$ provides *strong* evidence against model 2, and

- $-2 > \log_{10} BF_{21}$ provides *decisive* evidence against model 2.

One variant on this Bayes factor idea is that for comparing a Bayes model for observable $\boldsymbol{Y}$ (say model 1) to a Bayes model for observable $\boldsymbol{S} = \boldsymbol{s}(\boldsymbol{Y})$ where $\boldsymbol{s}(\cdot)$ is a 1-1 function (say model 2). That is, suppose that what is to be compared are models specified by

$$f_1(\boldsymbol{y}|\boldsymbol{\theta}_1) \quad \text{and} \quad g_1(\boldsymbol{\theta}_1)$$

and by

$$h(\boldsymbol{s}|\boldsymbol{\theta}_2) \quad \text{and} \quad g_2(\boldsymbol{\theta}_2)$$

Now the ratio that is a Bayes factor involves two marginal densities for the same observable. So in this case we must express both models in terms of the same observable. That requires remembering what was learned in Stat 542 about distributions of transformations of random variables. In the case that $\boldsymbol{Y}$ is discrete, it is easy enough to see that

$$f_2(\boldsymbol{y}|\boldsymbol{\theta}_2) = h(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{\theta}_2)$$

so that

$$BF_{21} = \frac{\int h(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{\theta}_2)\, g_2(\boldsymbol{\theta}_2)\, d\boldsymbol{\theta}_2}{\int f_1(\boldsymbol{y}|\boldsymbol{\theta}_1)\, g_1(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}$$

And in the case that $\boldsymbol{Y}$ is continuous, for $J_{\boldsymbol{s}}(\boldsymbol{y})$ the Jacobian of the transformation $\boldsymbol{s}$, the probability density for $\boldsymbol{y}$ under model 2 is

$$f_2(\boldsymbol{y}|\boldsymbol{\theta}_2) = |J_{\boldsymbol{s}}(\boldsymbol{y})|\, h(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{\theta}_2)$$

so that

$$BF_{21} = |J_{\boldsymbol{s}}(\boldsymbol{y})| \frac{\int h(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{\theta}_2)\, g_2(\boldsymbol{\theta}_2)\, d\boldsymbol{\theta}_2}{\int f_1(\boldsymbol{y}|\boldsymbol{\theta}_1)\, g_1(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}$$

3.6 `WinBUGS`, Numerical Problems, Restarts, and "Tighter Priors"

In complicated problems it is not uncommon for `WinBUGS` to stop in the middle of a simulation and report having numerical problems. It is rarely clear from the diagnostics the program provides exactly what has gone wrong. One can usually restart the simulation from the previous iterate (often after several attempts) and continue on in the simulation. The `WinBUGS` documentation suggests "tighter"/more informative priors as a general "fix" for this kind of problem. It is worth thinking about (even in the face of complete ignorance of numerical details) what could be the implications of this kind of difficulty, what might happen if one "ignores it" and routinely restarts the simulation, and the implication of following the manual's advice.

Getting the `WinBUGS` error warning is indication that there is some part of the $\boldsymbol{\theta}$ or $(\boldsymbol{y}_2, \boldsymbol{\theta})$ space that gets nontrivial posterior probability and where the current implementation of some evaluation of some function or some updating

algorithm breaks down. One could hope that in the most benign possible situation, this part of the space is some "relatively small/unimportant isolated corner of the space" and that a strategy of just blindly restarting the simulation will effectively replace the real posterior with a posterior that is the posterior conditioned on being in the "large/important part" of the space. (That counts on restarts from "just inside the 'good' part of the space and restricted to landing in the 'good' part" being equivalent to steps into the good part from inside the 'bad' part.)

Of course there are also less benign possibilities. Consider, for example, the possibility that the region where there are numerical problems serves as a boundary between two large and equally important parts of the $\boldsymbol{\theta}$ or $(\boldsymbol{y}_2, \boldsymbol{\theta})$ space. It's possible that one would then only see realizations from the part in which the chain is started, and thus end up with a completely erroneous view of the nature of the posterior. And it's not clear that there is really any way to tell whether the difficulty that one faces is benign or malignant.

The `WinBUGS` "fix" for this problem is a "fix" only in that it restricts the part of the $\boldsymbol{\theta}$ or $(\boldsymbol{y}_2, \boldsymbol{\theta})$ space that gets nontrivial posterior probability, and thereby keeps the sampler from getting into trouble. That is helpful only if one decides that really a less diffuse prior is adequate/appropriate in the context of the application. At the end of the day, the "real" fix for this kind of problem is doing one's own MCMC coding so that there is a chance of understanding exactly what has happened when something does go wrong.

3.7 Auxiliary Variables

Return to the notation of the exposition of the Gibbs, Metropolis-Hastings, and Metropolis-Hastings-in-Gibbs algorithms. It can sometimes be advantageous to simulate not only realizations of $\boldsymbol{\eta}$, but realizations of $(\boldsymbol{\eta}, \boldsymbol{\gamma})$ for some additional (typically vector) unobserved variable $\boldsymbol{\gamma}$. That is, suppose that $r(\boldsymbol{\gamma}|\boldsymbol{\eta})$ is a conditional density for $\boldsymbol{\gamma}$. Rather than doing MCMC from

$$h(\boldsymbol{\eta})$$

it can be more effective to do MCMC from

$$r(\boldsymbol{\gamma}|\boldsymbol{\eta})\, h(\boldsymbol{\eta})$$

and then simply ignore the values of $\boldsymbol{\gamma}$ so generated, using the $\boldsymbol{\eta}$'s to approximate properties of the $(h(\boldsymbol{\eta}))$ marginal of the joint distribution of $(\boldsymbol{\eta}, \boldsymbol{\gamma})$. As a matter of fact, slice sampling is an example of this idea. But it is also more generally helpful, and is related to the idea of data-augmentation used in the famous EM algorithm for maximization of a likelihood.

One nice application of the idea is in the analysis of interval-censored data from a continuous distribution belonging to some parametric family. In this context, $\boldsymbol{\eta}$ consists of the parameter vector and the likelihood is a product of $\boldsymbol{\eta}$ probabilities of intervals in which a sample of observations are known to lie (one term for each observation). But the Bayes analysis would typically be simpler if

one had instead the exact values of the observations, and could use a likelihood that is the product of the density values for the observations. The application of the auxiliary variables idea is then to let $\boldsymbol{\gamma}$ consist of the unobserved sample of exact values. In a Gibbs algorithm, when one is updating the parameter vector, one gets to use the posterior based on the exact values (instead of the typically more complicated posterior based on the identities of the intervals corresponding to the observations). The updates on the exact values are made using the (fixed parameters) conditional distributions over the intervals in which they are known to lie.

Another helpful application of the auxiliary variables idea is in the analysis of mixture data. That is a context where one has several parametric forms and assumes that data in hand are iid from a weighted (with positive weights summing to 1) average of these. The objects of interest are usually the weights and the parameters of the constituent distributions. A way of using auxiliary variables is to conceive of the individual observations as produced by a two-stage process, where first one of the constituents is chosen at random according the weights, and then the observation is generated from the individual constituent distribution. The "constituent identities" of all observations then become helpful auxiliary variables.

Finally, any "missing data" problem where one would naturally model an entire vector of observations but actually gets to observe only part of the vector is a candidate for use of the auxiliary variables idea. The missing or unobserved values are the obvious auxiliary variables.

3.8 Handling Interval Censoring and Truncation in `WinBUGS`

`WinBUGS` provides an "automatic" implementation of the auxiliary variables idea for interval censoring. Suppose that a part of the data vector, y, amounts to provision of the information that an incompletely observed variable from a parametric probability density $f(\cdot|\theta)$ (θ is part of the parameter vector $\boldsymbol{\theta}$) is somewhere in the interval (a, b). Let y_{aux} be this uncensored observation and $F(\cdot|\theta)$ the cdf corresponding to $f(\cdot|\theta)$. y's contribution to the likelihood is

$$F(b|\theta) - F(a|\theta) \tag{15}$$

and conditioned on y, y_{aux} has pdf

$$\frac{f(y_{\text{aux}}|\theta)}{F(b|\theta) - F(a|\theta)} I[a < y_{\text{aux}} < b] \tag{16}$$

So (multiplying (15) and (16)) we see that the net effect of including the auxiliary variable in MCMC is to replace (15) with

$$f(y_{\text{aux}}|\theta)\, I[a < y_{\text{aux}} < b] \tag{17}$$

in $h(\boldsymbol{\eta})$ from which one must simulate. The `WinBUGS` method for doing this is that instead of even trying to enter something like (15) into consideration, one

specifies that an unobserved variable y_{aux} contributes a term like (17) to $h(\boldsymbol{\eta})$. For the specific correct syntax, see the `Censoring and truncation` subsection of the `Model Specification` section of the `WinBUGS User Manual.`

The pdf (16) is of independent interest. It is the pdf on (a, b) that has the same shape as the density $f(\cdot|\theta)$ (a pdf typically on all of $\Re$ or on $\Re^+$) on that interval. It is usually known as a *truncated* version of $f(\cdot|\theta)$. One might imagine generating observations according to $f(\cdot|\theta)$, but that somehow all escape detection except those falling in (a, b). Density (16) is the pdf of any observation that is detected.

There is no easy/automatic way to use a truncated distribution as a model in `WinBUGS`. In particular one CAN NOT simply somehow make use of the censoring idea, somehow declaring that an *observed* variable has the distribution $f(\cdot|\theta)$ but is censored to (a, b). In the first place, (16) and (17) *are not the same functions*. Besides, if one uses the `WinBUGS` code for censoring and essentially includes terms like (17) in $h(\boldsymbol{\eta})$ but then turns around and provides observed values, one might as well have simply specified that the observation was from $f(\cdot|\theta)$ alone (the indicator takes the value 1). And the density (16) is NOT equivalent to $f(\cdot|\theta)$ as a contributor to a likelihood function.

The only way to make use of a truncated distribution as part of a `WinBUGS` model specification is to essentially program one's own version of the truncated pdf and use the `WinBUGS "zeros trick"` to get it included as a factor in the $h(\boldsymbol{\eta})$ from which `WinBUGS` samples. (See the `"Tricks: Advanced Use of the BUGS Language"` section of the `WinBUGS User Manual.`)

4 The Practice of Bayes Inference 2: Simple One-Sample Models

Both because one-sample statistical problems are of interest in their own right and because what we will find to be true for one-sample models becomes raw material for building and using more complicated models, we now consider the application of the Bayes paradigm to single samples from some common parametric models.

4.1 Binomial Observations

Suppose that observable $Y \sim$Binomial(n, p) for the unknown parameter $p \in (0, 1)$. Then when $Y = y \in \{0, 1, 2, \ldots, n\}$ the likelihood function becomes

$$L(p) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y}$$

From this it is immediate that a convenient form for a prior will be Beta(α, β), that is, a continuous distribution on $(0, 1)$ with pdf

$$g(p) = \frac{1}{\mathrm{B}(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1} \tag{18}$$

for some values $\alpha > 0$ and $\beta > 0$. (α and β thus become parameters of the prior distribution and are thus often termed "hyperparameters.") It is clear that the product $L(p)\, g(p)$ is proportional to a Beta$(\alpha + y, \beta + (n - y))$ density, i.e. with prior specified by (18)

$$g(p|y) \quad \text{is} \quad \mathrm{B}(\alpha + y, \beta + (n - y))$$

The Beta(α, β) distributions are conjugate priors for the simple binomial model.

Notice that the Fisher information in Y about p is

$$\mathcal{I}_Y(p) = -\mathrm{E}_p \frac{d^2}{dp^2} \ln f(Y|p) = \frac{n}{p(1-p)}$$

So the Jeffreys prior for the binomial model is specified by

$$g(p) = \sqrt{\mathcal{I}_Y(p)} = p^{-1/2}(1-p)^{-1/2}$$

That is, in this case the Jeffreys prior is a member of the conjugate Beta family, the Beta$(1/2, 1/2)$ distribution.

The Beta(α, β) mean is $\alpha/(\alpha + \beta)$ so the posterior mean of p with a Beta prior is

$$\begin{aligned} \mathrm{E}[p|Y = y] &= \frac{\alpha + y}{\alpha + \beta + n} \\ &= \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\left(\frac{\alpha}{\alpha + \beta}\right) + \left(\frac{n}{\alpha + \beta + n}\right)\left(\frac{y}{n}\right) \end{aligned}$$

and this motivates thinking about the hyperparameters of a Beta prior in terms of $(\alpha + \beta)$ being a kind of "prior sample size" and α being a corresponding "prior number of successes." (The posterior mean is a weighted average of the prior mean and sample mean with respective weights in proportion to $(\alpha + \beta)$ and n.)

Notice that if one defines

$$\theta = \mathrm{logit}(p) \equiv \ln\left(\frac{p}{1-p}\right)$$

and chooses an improper prior for θ that is "uniform on $\Re$," then

$$p = \mathrm{logit}^{-1}(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)}$$

has an improper prior with "density" (proportional to)

$$g(p) = p^{-1}(1-p)^{-1} \tag{19}$$

The meaning here is that for $0 < a < b < 1$

$$\int_a^b p^{-1}(1-p)^{-1}\, dp \propto \mathrm{logit}(b) - \mathrm{logit}(a) = \int_{\mathrm{logit}(a)}^{\mathrm{logit}(b)} 1 d\theta$$

Now the improper prior specified by (19) is in some sense the $\alpha = 0$ and $\beta = 0$ limit of (proper) Beta(α, β) priors. As long as $0 < y < n$ this improper prior for p and the likelihood combine to give proper Beta posterior for p.

4.2 Poisson Observations

Suppose that observable $Y \sim$Poisson(λ) for the unknown parameter $\lambda \in (0, \infty)$. Then when $Y = y \in \{0, 1, 2, \ldots, \}$ the likelihood function becomes

$$L(\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}$$

A conjugate form for a prior is $\Gamma(\alpha, \beta)$, that is, a distribution on $(0, \infty)$ with pdf

$$g(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda) \tag{20}$$

It is then clear that the product $L(\lambda)\, g(\lambda)$ is proportional to the $\Gamma(\alpha + y, \beta + 1)$ density, that is with prior specified by (20)

$$g(\lambda|y) \quad \text{is} \quad \Gamma(\alpha + y, \beta + 1)$$

The $\Gamma(\alpha, \beta)$ mean is α/β and the variance of the distribution is α/β^2. So for $\alpha = \beta =$ *some small number*, the prior (20) has mean 1 and a large variance. The corresponding posterior mean is $(\alpha + y)/(\beta + 1) \approx y$ and the posterior standard deviation is $\sqrt{(\alpha + y)/(\beta + 1)^2} \approx \sqrt{y}$.

Notice that the Fisher information in Y about λ is

$$\mathcal{I}_Y(\lambda) = -\mathrm{E}_\lambda \frac{d^2}{d\lambda^2} \ln f(Y|\lambda) = \frac{1}{\lambda}$$

So the (improper) Jeffreys prior for λ is specified by

$$g(\lambda) = \frac{1}{\sqrt{\lambda}}$$

and for $\alpha = 1/2$ and $\beta =$ *some small number*, the $\Gamma(\alpha, \beta)$ prior is approximately the Jeffreys prior.

Finally, note that for an improper prior for λ that is "uniform on $(0, \infty)$," i.e. $g(\lambda) = 1$ on that interval, the posterior density is

$$g(\lambda|y) \propto \exp(-\lambda)\lambda^y$$

i.e. the posterior is $\Gamma(y + 1, 1)$.

4.3 Univariate Normal Observations

One- and two-parameter versions of models involving $\mathrm{N}(\mu, \sigma^2)$ observations can be considered. We start with the one-parameter versions.

4.3.1 σ^2 Fixed/Known

Suppose first that $Y \sim \mathrm{N}\left(\mu, \sigma^2\right)$ where σ^2 is a known constant (and thus is not an object of inference). Notice that here Y could be a sample mean of iid normal observations, in which case σ^2 would be a population variance over sample size. (Note too that in such a case, sufficiency considerations promise that inference based on the sample mean is equivalent to inference based on the original set of individual observations.)

The likelihood function here is

$$L(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{21}$$

Then consider a (conjugate) $\mathrm{N}\left(m, \gamma^2\right)$ prior for μ with density

$$g(\mu) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{(\mu-m)^2}{2\gamma^2}\right) \tag{22}$$

Then

$$L(\mu)\, g(\mu) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\gamma^2}\right)\mu^2 + \left(\frac{y}{\sigma^2}+\frac{m}{\gamma^2}\right)\mu\right)$$

So the posterior pdf $g(\mu|y)$ is again normal with

$$variance = \left(\frac{1}{\sigma^2}+\frac{1}{\gamma^2}\right)^{-1} = \frac{\gamma^2\sigma^2}{\sigma^2+\gamma^2} \tag{23}$$

and

$$mean = \left(\frac{y}{\sigma^2}+\frac{m}{\gamma^2}\right) \cdot variance = \frac{\frac{y}{\sigma^2}+\frac{m}{\gamma^2}}{\frac{1}{\sigma^2}+\frac{1}{\gamma^2}} \tag{24}$$

For purposes of Bayes analysis, it is often convenient to think in terms of a distribution's

$$precision = \frac{1}{variance}$$

In these terms, equation (23) says that in this model

$$posterior\ precision = prior\ precision + precision\ of\ likelihood \tag{25}$$

and equation (24) says that in this model the posterior mean is a precision-weighted average of the prior and sample means.

As a bit of an aside, notice that while (25) and (24) describe the posterior distribution of $\mu|Y=y$, one might also be interested in the marginal distribution of Y. This distribution is also normal, with

$$\mathrm{E}Y = m \quad \text{and} \quad \mathrm{Var}Y = \sigma^2 + \gamma^2$$

So in the case of the marginal distribution, it is variances (not precisions) that add.

The Fisher information in Y about μ is

$$\mathcal{I}_Y(\mu) = -\mathrm{E}_\mu \frac{d^2}{d\mu^2} \ln L(\mu) = \frac{1}{\sigma^2}$$

This is constant in μ. So the (improper) Jeffreys prior for μ is "uniform on $\Re$," $g(\mu) = 1$. With this improper prior, the posterior is proportional to $L(\mu)$. Looking again at (21) we see that the posterior density $g(\mu|y)$ is then $\mathrm{N}\left(y, \sigma^2\right)$. Notice that in some sense this improper prior is the $\gamma^2 = \infty$ limit of a conjugate prior (22) and the corresponding $\mathrm{N}\left(y, \sigma^2\right)$ posterior is the $\gamma^2 = \infty$ limit of the posterior for the conjugate prior.

Consider the proper $\mathrm{U}(a, b)$ prior with density

$$g(\mu) \propto I[a < \mu < b]$$

With this prior the posterior has density

$$g(\mu|y) \propto I[a < \mu < b] \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

That is, the posterior is the $\mathrm{N}\left(y, \sigma^2\right)$ distribution truncated to the interval (a, b). Then, as long as $a \ll y \ll b$ (relative to the size of σ) the posterior is essentially $\mathrm{N}\left(y, \sigma^2\right)$. That is, this structure will allow one to approximate a Jeffreys analysis using a proper prior.

4.3.2 μ Fixed/ Known

Suppose now that $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ has components that are iid $\mathrm{N}\left(\mu, \sigma^2\right)$ where μ is a known constant (and thus is not an object of inference and can be used in formulas for statistics to be calculated from the data). The likelihood function here is

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

Let

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

and with this notation note that

$$L\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{n}{2\sigma^2} w\right)$$

A conjugate prior here is the so-called Inv-$\Gamma(\alpha, \beta)$ distribution on $(0, \infty)$ with pdf

$$g\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \tag{26}$$

It is then obvious (upon inspection of the product $L\left(\sigma^2\right) g\left(\sigma^2\right)$) that using prior (26), the posterior is

$$\text{Inv-}\Gamma\left(\alpha+\frac{n}{2}, \beta+\frac{nw}{2}\right) \tag{27}$$

A useful/standard re-expression of this development is in terms of the so-called "scaled inverse χ^2 distributions." That is, one could start by using a prior for σ^2 that is the distribution of

$$\frac{\phi^2\nu}{X} \quad \text{for} \quad X \sim \chi^2_\nu \tag{28}$$

From (28) it it clear that ϕ^2 is a scale parameter for this distribution and that ν governs the shape of the distribution. The textbook uses the notation

$$\sigma^2 \sim \text{Inv-}\chi^2\left(\nu, \phi^2\right) \quad \text{or} \quad \sigma^2 \sim \text{Inv-}\Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\cdot\phi^2\right)$$

for the assumption that σ^2 has the distribution of (28). With this notation, the posterior is

$$\text{Inv-}\Gamma\left(\frac{1}{2}\left(\nu+n\right), \frac{\nu}{2}\phi^2+\frac{nw}{2}\right) \quad \text{or} \quad \text{Inv-}\chi^2\left(\nu+n, \frac{\nu\phi^2+nw}{\nu+n}\right) \tag{29}$$

This second form in (29) provides a very nice interpretation of what happens when the prior and likelihood are combined. The degrees of freedom add, with the prior essentially having the same influence on the posterior as would a legitimate sample of size ν. The posterior scale parameter is an appropriately weighted average of the prior scale parameter and w (the known-mean-n-denominator sample variance).

It's fairly easy to determine that the Fisher information in $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ about σ^2 is

$$\mathcal{I}_Y\left(\sigma^2\right) = \frac{n}{2\sigma^4}$$

so that a Jeffreys (improper) prior for σ^2 is specified by

$$g\left(\sigma^2\right) \propto \frac{1}{\sigma^2} \tag{30}$$

Notice that since for $0 < a < b < \infty$

$$\int_a^b \frac{1}{x}dx = \ln b - \ln a = \int_{\ln a}^{\ln b} 1dx$$

the improper prior for σ^2 specified by (30) is equivalent to an (improper) prior for $\ln\sigma^2$ (or $\ln\sigma$) that is uniform on $\Re$.

Notice also that the Jeffreys improper prior (30) is in some sense the $\alpha = 0$ and $\beta = 0$ limit of the Inv-$\Gamma\left(\alpha, \beta\right)$ prior, or equivalently the fixed ϕ^2 and $\nu = 0$ limit of the Inv-$\chi^2\left(\nu, \phi^2\right)$ prior. The posterior for this improper prior is specified by

$$g\left(\sigma^2|w\right) \propto \left(\sigma^2\right)^{-1}\left(\sigma^2\right)^{-n/2}\exp\left(-\frac{n}{2\sigma^2}w\right)$$

that is, the (proper) posterior is

$$\text{Inv-}\Gamma\left(\frac{n}{2}, \frac{nw}{2}\right) \quad \text{or} \quad \text{Inv-}\chi^2\left(n, w\right)$$

4.3.3 Both μ and σ^2 Unknown

Suppose finally that $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ has components that are iid $\mathrm{N}\left(\mu, \sigma^2\right)$, where neither of the parameters is known. The likelihood function is

$$L\left(\mu, \sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n \left(y_i - \mu\right)^2}{2\sigma^2}\right)$$

There are several obvious choices for a (joint) prior distribution for $\left(\mu, \sigma^2\right)$.

First, one might put together the two improper Jeffreys priors for μ and σ^2 individually. That is one might try using an improper prior on $\Re \times (0, \infty)$ specified by

$$g\left(\mu, \sigma^2\right) \propto 1 \cdot \frac{1}{\sigma^2} \tag{31}$$

Since this a product of a function of μ and a function of σ^2, this is a prior of "independence." As it turns out, provided that $n \geq 2$ the prior (31) has a corresponding proper posterior, that is of course specified by a joint density for μ and σ^2 of the form

$$g\left(\mu, \sigma^2 | \boldsymbol{y}\right) \propto L\left(\mu, \sigma^2\right) g\left(\mu, \sigma^2\right) \tag{32}$$

The posterior density (32) is NOT the product of a function of μ and a function of σ^2, and thus does not specify a posterior of independence. This is not a bad feature of (31). We should, for example, want cases where the usual sample variance s^2 is small to be ones that produce posteriors that indicate 1) that σ^2 is likely small, *and* that therefore 2) μ has been fairly precisely determined ... one does not want the posterior to have the same conditional variance for μ for all σ^2 values.

Pages 73-77 of the text show that posterior (32) has attractive marginals. First, it turns out that

$$g\left(\sigma^2 | \boldsymbol{y}\right) \quad \text{is} \quad \text{Inv-}\chi^2\left(n-1, s^2\right)$$

that is, conditioned on $\boldsymbol{Y} = \boldsymbol{y}$ (and therefore the value of the usual sample variance, s^2) σ^2 has the distribution of

$$\frac{(n-1)\, s^2}{X} \quad \text{for} \quad X \sim \chi^2_{n-1}$$

Further, conditioned on $\boldsymbol{Y} = \boldsymbol{y}$ (and therefore the values of $\bar{y}$ and s^2) μ has the distribution of

$$\bar{y} + T\frac{s}{\sqrt{n}} \quad \text{for} \quad T \sim t_{n-1}$$

These two facts imply that Bayes posterior (credible) intervals for μ and σ^2 will agree exactly with standard Stat 500 confidence intervals (at the same level) for the parameters.

Of course it is possible to approximate the improper prior (31) with proper joint distributions and get posterior inferences that are essentially the same as for this improper prior. For example, as an approximation to (31), one might specify that *a priori* μ and $\ln\sigma^2$ are independent with

$$\mu \sim \mathrm{U}\left(small_1, large_1\right) \quad \text{and} \quad \ln\sigma^2 \sim \mathrm{U}\left(small_2, large_2\right)$$

and expect to get essentially frequentist posterior inferences.

Another possibility is to use a product of two proper conjugate marginal priors for a joint prior. That is, one might specify that *a priori* μ and σ^2 are independent with

$$\mu \sim \mathrm{N}\left(m, \gamma^2\right) \quad \text{and} \quad \sigma^2 \sim \text{Inv-}\chi^2\left(\nu, \phi^2\right)$$

As it turns out, nothing works out very cleanly with this choice of prior. See pages 80-82 of the textbook. Analysis of the posterior here is really a job for simulation. Obviously, one expects that for large γ^2 and small ν, inferences based on this structure should look much like those made using the form (31), and therefore a lot like Stat 500 inferences.

Finally, on pages 78-80 the textbook discusses what seems to me to be a very unattractive but conjugate prior for $\left(\mu, \sigma^2\right)$. I find the assumed *prior* dependence between the two parameters and the specification of the constant κ_0 to be quite unnatural.

4.4 Multivariate Normal Observations

As is completely standard, for a nonsingular covariance matrix $\boldsymbol{\Sigma}$, we will say that a k-dimensional random vector $\boldsymbol{Y} \sim \mathrm{MVN}_k\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ provided it has a pdf on $\Re^k$

$$f\left(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \left(\det \boldsymbol{\Sigma}\right)^{-1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right)$$

Then if $\boldsymbol{Y} = \left(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n\right)$ where the components $\boldsymbol{Y}_i$ are iid $\mathrm{MVN}_k\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, the joint pdf is

$$\begin{aligned} f\left(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) &= \left(\det \boldsymbol{\Sigma}\right)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_i-\boldsymbol{\mu}\right)\right) \\ &= \left(\det \boldsymbol{\Sigma}\right)^{-n/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0\right)\right) \end{aligned}$$

where

$$\boldsymbol{S}_0 = \sum_{i=1}^{n}\left(\boldsymbol{y}_i-\boldsymbol{\mu}\right)\left(\boldsymbol{y}_i-\boldsymbol{\mu}\right)' \tag{33}$$

We proceed to consider models involving multivariate normal observations.

4.4.1 $\boldsymbol{\Sigma}$ Fixed/Known

Suppose that $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ where the components $\boldsymbol{Y}_i$ are iid $\text{MVN}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma}$ is known, the likelihood function is

$$L(\boldsymbol{\mu}) = (\det \boldsymbol{\Sigma})^{-n/2} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0\right)\right)$$

for $\boldsymbol{S}_0$ the function of $\boldsymbol{\mu}$ defined in (33). Then consider a conjugate $\text{MVN}_k(\boldsymbol{m}, \boldsymbol{\Gamma}_0)$ prior for $\boldsymbol{\mu}$ here. As it turns out, in direct generalization of the univariate normal case with known variance and (24) and (23), the posterior pdf $g(\boldsymbol{\mu}|\boldsymbol{y})$ is MVN_k with mean vector

$$\boldsymbol{\mu}_n = \left(\boldsymbol{\Gamma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Gamma}_0^{-1}\boldsymbol{m} + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{y}}\right) \tag{34}$$

and covariance matrix

$$\boldsymbol{\Gamma}_n = \left(\boldsymbol{\Gamma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \tag{35}$$

Thinking of a covariance matrix as an "inverse precision matrix," the sampling precision of $\bar{\boldsymbol{Y}}$ is $n\boldsymbol{\Sigma}^{-1}$, and (35) says that the posterior precision is the sum of the prior precision and the precision of the likelihood, while (34) says the posterior mean is a precision-weighted average of the prior mean and the sample mean. If the matrix $\boldsymbol{\Gamma}_0$ is "big" (the prior precision matrix $\boldsymbol{\Gamma}_0^{-1}$ is "small") then the posterior for the conjugate prior is approximately $\text{MVN}_k\left(\bar{\boldsymbol{y}}, \frac{1}{n}\boldsymbol{\Sigma}\right)$. More directly, if one uses an improper prior for $\boldsymbol{\mu}$ that is uniform on $\Re^k$ one gets this $\text{MVN}_k\left(\bar{\boldsymbol{y}}, \frac{1}{n}\boldsymbol{\Sigma}\right)$ posterior exactly.

4.4.2 $\boldsymbol{\mu}$ Fixed/Known

Suppose that $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ where the components $\boldsymbol{Y}_i$ are iid $\text{MVN}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\mu}$ is known, the likelihood function is

$$\begin{aligned} L(\boldsymbol{\Sigma}) &= (\det \boldsymbol{\Sigma})^{-n/2} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0\right)\right) \\ &= (\det \boldsymbol{\Sigma})^{-n/2} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{S}_0\boldsymbol{\Sigma}^{-1}\right)\right) \end{aligned} \tag{36}$$

for $\boldsymbol{S}_0$ defined in (33). In order to do Bayes inference, we then need to place prior distributions on *covariance matrices*. This requires doing some post-Stat 542 probability (that essentially generalizes the chi-squared distributions to multivariate cases) and introducing the Wishart (and inverse Wishart) distributions.

Wishart Distributions

Let $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_\nu$ be iid $\text{MVN}_k(\boldsymbol{0}, \boldsymbol{\Delta})$ for a non-singular covariance matrix $\boldsymbol{\Delta}$. Then for $\nu \geq k$ consider the "sum of squares and cross-products matrix"

(for these mean $\mathbf{0}$ random vectors)

$$\boldsymbol{W} = \sum_{i=1}^{\nu} \boldsymbol{\alpha}_i' \boldsymbol{\alpha}_i = \left(\sum_{i=1}^{\nu} \alpha_{il} \alpha_{im} \right) \begin{array}{l} l = 1, 2, \ldots, k \\ m = 1, 2, \ldots, k \end{array} \tag{37}$$

This random $k \times k$ matrix has the so-called Wishart$(\nu, \boldsymbol{\Delta})$ distribution. Now $\boldsymbol{W}$ has only $k + \frac{1}{2}\left(k^2 - k\right)$ different entries, as those below the diagonal are the same as their opposite numbers above the diagonal. So if one wishes to write a pdf to describe the distribution of $\boldsymbol{W}$ it will really be a function of those fewer than k^2 distinct elements. It turns out that (thought of on the right as a function of $k \times k$ matrix $\boldsymbol{w}$ and on the left as a function of the elements on and above the diagonal of $\boldsymbol{w}$) $\boldsymbol{W}$ has a pdf

$$f\left(\boldsymbol{w}|\nu, \boldsymbol{\Delta}\right) \propto \left(\det \boldsymbol{w}\right)^{(\nu-k-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Delta}^{-1}\boldsymbol{w}\right)\right) \tag{38}$$

It follows from either representation (37) or density (38) that if $\boldsymbol{W} \sim$Wishart$(\nu, \boldsymbol{\Delta})$

$$\mathrm{E}\boldsymbol{W} = \nu\boldsymbol{\Delta}$$

and in fact the diagonal elements of $\boldsymbol{W}$ are scaled χ^2 variables. That is

$$W_{ii} \sim \Gamma\left(\frac{\nu}{2}, \frac{\delta_{ii}}{2}\right)$$

where δ_{ii} is the ith diagonal element of $\boldsymbol{\Delta}$. That is, W_{ii} has the same distribution as $\delta_{ii}X$ for $X \sim \chi^2_\nu$.

For users of `WinBUGS` a serious **caution** needs to be interjected at this point.

$$\boldsymbol{V} \sim \text{WinBUGS-Wishart}\left(\nu, \boldsymbol{\Gamma}\right)$$

means that

$$\boldsymbol{V} \sim \text{Wishart}\left(\nu, \boldsymbol{\Gamma}^{-1}\right)$$

in the present notation/language. That is, `WinBUGS` parameterizes with precision matrices, *not* covariance matrices.

Inverse Wishart Distributions

Next, for $\boldsymbol{W} \sim$Wishart$(\nu, \boldsymbol{\Delta})$, consider

$$\mathbf{U} = \mathbf{W}^{-1}$$

This $k \times k$ inverse sum of squares and cross-products random matrix can be shown to have probability density

$$f\left(\boldsymbol{u}|\nu, \boldsymbol{\Delta}\right) \propto \left(\det \boldsymbol{u}\right)^{-(\nu+k+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Delta}^{-1}\boldsymbol{u}^{-1}\right)\right) \tag{39}$$

We will call the distribution of $\mathbf{U} = \mathbf{W}^{-1}$ (as specified by the pdf in (39)) Inv-Wishart$(\nu, \mathbf{\Delta})$. That is

$$\boldsymbol{W} \sim \text{Wishart}\,(\nu, \mathbf{\Delta}) \quad \Rightarrow \quad \mathbf{U} = \boldsymbol{W}^{-1} \sim \text{Inv-Wishart}\,(\nu, \mathbf{\Delta})$$

As it turns out,

$$\text{E}\mathbf{U} = \text{E}\mathbf{W}^{-1} = \frac{1}{\nu - k - 1}\mathbf{\Delta}^{-1} \tag{40}$$

and the diagonal entries of $\boldsymbol{U}$ are scaled inverse χ^2 variables. That is, for γ_{ii} the ith diagonal entry of $\mathbf{\Delta}^{-1}$, the ith diagonal entry of $\boldsymbol{U}$ is

$$U_{ii} \sim \text{Inv-}\Gamma\left(\frac{\nu - k + 1}{2}, \frac{\gamma_{ii}}{2}\right) \quad \text{or} \quad \text{Inv-}\chi^2\left(\nu - k + 1, \frac{\gamma_{ii}}{\nu - k + 1}\right)$$

(recall the definitions in Section 4.3.2), i.e. U_{ii} has the same distribution as γ_{ii}/X for $X \sim \chi^2_{\nu-k+1}$. Notice also that with the conventions of this discussion, $\boldsymbol{W} \sim$Wishart$(\nu, \mathbf{\Delta})$ implies that $\mathbf{\Delta}$ is a scaling matrix for $\boldsymbol{W}$ and $\mathbf{\Delta}^{-1}$ is a scaling matrix for $\boldsymbol{U} = \boldsymbol{W}^{-1}$.

Application of Inverse Wishart Priors

So, now comparing the form of the "known $\boldsymbol{\mu}$ multivariate normal likelihood" in (36) and the Inv-Wishart pdf in (39) it is clear that the inverse Wishart distributions provide conjugate priors for this situation. If for $\nu \geq k$ and a given non-singular covariance matrix $\mathbf{\Delta}$, one makes the prior assumption that $\mathbf{\Sigma} \sim$Inv-Wishart$(\nu, \mathbf{\Delta})$, i.e. assumes that

$$g\left(\mathbf{\Sigma}\right) \propto \left(\det \mathbf{\Sigma}\right)^{-(\nu+k+1)/2} \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{\Delta}^{-1}\mathbf{\Sigma}^{-1}\right)\right) \tag{41}$$

multiplying (36) and (41) shows that the posterior is

$$\text{Inv-Wishart}\left(n + \nu, \left(\boldsymbol{S}_0 + \mathbf{\Delta}^{-1}\right)^{-1}\right) \tag{42}$$

So, for example, the posterior mean for $\mathbf{\Sigma}$ is

$$\frac{1}{n + \nu - k - 1}\left(\boldsymbol{S}_0 + \mathbf{\Delta}^{-1}\right)$$

and, for example, Σ_{ii} has the posterior distribution of s_{ii}/X for $X \sim \chi^2_{n+\nu-k+1}$ and s_{ii} the ith diagonal entry of $\boldsymbol{S}_0 + \mathbf{\Delta}^{-1}$, i.e. the posterior distribution of the ith diagonal entry of $\mathbf{\Sigma}$ is Inv-$\Gamma\left(\frac{n+\nu-k+1}{2}, \frac{s_{ii}}{2}\right)$.

It's fairly obvious that the smaller one makes ν the less influential is the prior on the form of the posterior (42). ν is sometimes thought of as a fictitious "prior sample size" in comparison to n.

Using `WinBUGS` With Inverse Wishart Priors

Consider what is required in order to set an Inv-Wishart prior for $\mathbf{\Sigma}$ with a desired/target prior mean. The form (40) implies that if one has in mind some target prior mean for $\mathbf{\Sigma}$, say $\mathbf{\Lambda}$, one wants prior parameters ν and $\mathbf{\Delta}$ such that

$$\mathbf{\Lambda} = \frac{1}{\nu - k - 1}\mathbf{\Delta}^{-1}$$

that is

$$\mathbf{\Delta} = \frac{1}{\nu - k - 1}\mathbf{\Lambda}^{-1} \tag{43}$$

One may set a prior $\mathbf{\Sigma} \sim$Inv-Wishart$(\nu, \mathbf{\Delta})$ by setting $\mathbf{\Sigma}^{-1} \sim$Wishart$(\nu, \mathbf{\Delta})$ and have a desired prior mean for $\mathbf{\Sigma}$ by using (43). If one is then using `WinBUGS` this is done by setting $\mathbf{\Sigma}^{-1} \sim$WinBUGS-Wishart$\left(\nu, \mathbf{\Delta}^{-1}\right)$, and to get a target prior mean of $\mathbf{\Lambda}$ requires that one use $\mathbf{\Sigma}^{-1} \sim$WinBUGS-Wishart$(\nu, (\nu - k - 1)\,\mathbf{\Lambda})$.

An Improper Limit of Inverse Wishart Priors

A candidate for a "non-informative" improper prior for $\mathbf{\Sigma}$ is

$$g(\mathbf{\Sigma}) \propto (\det \mathbf{\Sigma})^{-(k+1)/2} \tag{44}$$

which is in some sense the $\nu = 0$ and "$\mathbf{\Delta} = \infty$" formal limit of the form (41). The product of forms (36) and (44) produces an Inv-Wishart$\left(n, \boldsymbol{S}_0^{-1}\right)$ posterior. So, for example, under (44) the posterior mean is

$$\frac{1}{n - k - 1}\left(\boldsymbol{S}_0^{-1}\right)^{-1} = \frac{1}{n - k - 1}\boldsymbol{S}_0$$

4.4.3 Both $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ Unknown

If one now treats both $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ as unknown, for data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ where the components $\boldsymbol{Y}_i$ are iid MVN$_k(\boldsymbol{\mu}, \mathbf{\Sigma})$ the likelihood function is

$$L(\boldsymbol{\mu}, \mathbf{\Sigma}) = (\det \mathbf{\Sigma})^{-n/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\boldsymbol{S}_0\right)\right)$$

where $\boldsymbol{S}_0$ is still (the function of $\boldsymbol{\mu}$) defined in (33).

Probably the most appealing story to be told concerning Bayes inference in this context concerns what happens when one uses an improper prior for the elements of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ that is put together by taking the product of the two non-informative priors from the "known $\boldsymbol{\mu}$" and "known $\mathbf{\Sigma}$" cases. That is, one might consider

$$g(\boldsymbol{\mu}, \mathbf{\Sigma}) \propto 1 \cdot (\det \mathbf{\Sigma})^{-(k+1)/2} \tag{45}$$

With improper prior (45) as a direct generalization of what one gets in the univariate normal problem with unknown mean and variance, the posterior distribution of $\mathbf{\Sigma}$ is Inv-Wishart, i.e.

$$\mathbf{\Sigma} | \ \boldsymbol{y} \sim \text{Inv-Wishart}\left(n - 1, \boldsymbol{S}^{-1}\right) \tag{46}$$

where

$$\boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}) (\boldsymbol{y}_i - \bar{\boldsymbol{y}})'$$

is the sum of squares and cross-products around the sample means matrix (i.e. is $n-1$ times the sample covariance matrix). By the way, the textbook has this wrong on its page 88 (wrongly substituting $\boldsymbol{S}$ for $\boldsymbol{S}^{-1}$ in the Inv-Wishart form for the posterior). (46) and (40) then imply that the posterior mean for $\boldsymbol{\Sigma}$ using (45) is

$$\frac{1}{n-k-2}\boldsymbol{S}$$

Further, again as a direct generalization of what one gets in the univariate normal problem with unknown mean and variance, the posterior distribution of $\boldsymbol{\mu}$ is multivariate t. That is,

$$(\boldsymbol{\mu} - \bar{\boldsymbol{y}}) \mid \boldsymbol{y} \sim \text{Multivariate } t \left(n-k, \frac{1}{n}\left(\frac{1}{n-k}\boldsymbol{S}\right)\right)$$

meaning that $\boldsymbol{\mu}$ has the (posterior) distribution of

$$\bar{\boldsymbol{y}} + \frac{1}{\sqrt{n}}\left(\frac{1}{n-k}\boldsymbol{S}\right)^{1/2}\sqrt{\frac{n-k}{W}}\boldsymbol{Z}$$

where $\boldsymbol{Z}$ is a $k \times 1$ vector of independent $\text{N}(0,1)$ random variables, independent of $W \sim \chi^2_{n-k}$, and $\left(\frac{1}{n-k}\boldsymbol{S}\right)^{1/2}$ is a matrix square root of $\left(\frac{1}{n-k}\boldsymbol{S}\right)$. (Notice that this fact allows one to easily (at least by simulation) find the posterior distribution of (and thus credible sets for) any parametric function $h(\boldsymbol{\mu})$.)

An alternative to the improper prior (45) is a product of two proper priors for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The obvious choices for the two marginal priors are a $\text{MVN}_k(\boldsymbol{m}, \boldsymbol{\Gamma}_0)$ prior for $\boldsymbol{\mu}$ and an Inv-Wishart$(\nu, \boldsymbol{\Delta})$ prior for $\boldsymbol{\Sigma}$. Nothing works out very cleanly (in terms of analytical formulas) under such assumptions, but one should expect that for $\boldsymbol{\Gamma}_0$ "big," ν small, and $\boldsymbol{\Delta}$ "big," inferences for the proper prior should approximate those for the improper prior (45).

4.5 Multinomial Observations

Consider now n independent identical trials where each of these has k possible outcomes with respective probabilities $p_1, p_2, \ldots, p_k$ (where each $p_i \in (0,1)$ and $\sum p_i = 1$). If

$$Y_i = \text{the number of outcomes that are of type } i$$

then $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)$ is $\text{Multinomial}_k(n, \boldsymbol{p})$ (for $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$) and has (joint) probability mass function

$$f(\boldsymbol{y}|\boldsymbol{p}) = \binom{n}{y_1, y_2, \ldots, y_k} \prod_{i=1}^{k} p_i^{y_i}$$

(for vectors $\boldsymbol{y}$ of non-negative integers y_i with sum n). The coordinate variables Y_i are, of course, Binomial(n, p_i).

Consider inference for $\boldsymbol{p}$ based on $\boldsymbol{Y} \sim$Multinomial$_k(n, \boldsymbol{p})$. The likelihood function is

$$L(\boldsymbol{p}) = \left(\begin{array}{c} n \\ y_1, y_2, \ldots, y_k \end{array} \right) \prod_{i=1}^{k} p_i^{y_i} \tag{47}$$

and in order to do Bayes inference, one must find a way to put a prior distribution on the set of (k-vectors) $\boldsymbol{p}$ that have each $p_i \in (0,1)$ and $\sum p_i = 1$.

The most convenient (and conjugate) form for a distribution on the set of $\boldsymbol{p}$'s that have each $p_i \in (0,1)$ and $\sum p_i = 1$ is the Dirichlet form. If $X_1, X_2, \ldots, X_k$ are independent random variables with $X_i \sim \Gamma(\alpha_i, 1)$ for positive constants $\alpha_1, \alpha_2, \ldots, \alpha_k$ and one defines

$$W_i = \frac{X_i}{\sum_{i=1}^{k} X_i}$$

then

$$\boldsymbol{W} = \left(\begin{array}{c} W_1 \\ W_2 \\ \vdots \\ W_k \end{array} \right) \sim \text{Dirichlet}_k(\boldsymbol{\alpha})$$

Using this characterization it is easy to argue that the ith marginal of a Dirichlet vector is Beta$\left(\alpha_i, \sum_{j \neq i} \alpha_j\right)$ and that conditional distributions of some coordinates given the values of the others are the distributions of multiples of Dirichlet vectors. The pdf for $k-1$ coordinates of $\boldsymbol{W} \sim$Dirichlet$_k(\boldsymbol{\alpha})$ (written in terms of all k coordinates) is

$$f(\boldsymbol{w}|\boldsymbol{\alpha}) \propto \prod_{i=1}^{k} w_i^{\alpha_i} \tag{48}$$

Using (47) and (48) it is clear that using a Dirichlet$_k(\boldsymbol{\alpha})$ prior for $\boldsymbol{p}$, the posterior is

$$\boldsymbol{p}|\boldsymbol{y} \text{ is } \text{Dirichlet}_k(\boldsymbol{\alpha} + \boldsymbol{y})$$

So, for example, the Beta posterior marginal of p_i has mean

$$\frac{\alpha_i + y_i}{\sum_{i=1}^{k} \alpha_i + n} = \frac{\left(\sum_{i=1}^{k} \alpha_i\right)\left(\dfrac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}\right) + n\left(\dfrac{y_i}{n}\right)}{\sum_{i=1}^{k} \alpha_i + n} \tag{49}$$

The form of the posterior mean (49) suggests the common interpretation that $\sum_{i=1}^{k} \alpha_i$ functions as a kind of "prior sample size" in comparison to n for weighting the prior against the sample information (encoded in the relative frequencies y_i/n). If the former is small in comparison to n, the posterior means (49) are nearly the sample relative frequencies. Otherwise, the posterior means are the

sample relative frequencies shrunken towards the prior means $\alpha_i / \left(\sum_{i=1}^{k} \alpha_i\right)$. Of course, the larger is $\sum_{i=1}^{k} \alpha_i$ the more concentrated/less dispersed is the prior and the larger is $\sum_{i=1}^{k} \alpha_i + n$ the more concentrated/less dispersed is the posterior.

5 Graphical Representation of Some Aspects of Large Joint Distributions

This section of the outline covers some material taken from Chapters 17 and 18 of *All of Statistics* by Wasserman. (Wasserman's book refers its readers to *Introduction to Graphical Modeling* by Edwards for a complete treatment of this subject. Lauritzen's *Graphical Models* is another standard reference.) It concerns using graphs (both directed and undirected) as aids to understanding simple (independence) structure in high-dimensional distributions of random variables $(X, Y, Z, \ldots)$ and in relating that structure to functional forms for corresponding densities.

The developers of `WinBUGS` recommend making a "directed graphical" version of every model one uses in the software, and the logic of how one naturally builds Bayes models is most easily related to directed graphs. So although concepts for undirected graphs (and how they represent independence relationships) are simpler than those for directed graphs, we will discuss the more complicated case (of directed graphs) first. But before doing even this, we make some observations about conditional independence.

5.1 Conditional Independence

Random variables X and Y are **conditionally independent given** Z written

$$X \underline{\perp\!\!\!\perp} Y \mid Z$$

provided

$$f_{X,Y|Z}(x, y|z) = f_{X|Z}(x|z)\, f_{Y|Z}(y|z)$$

A basic result about conditional independence is that

$$X \underline{\perp\!\!\!\perp} Y \mid Z \iff f_{X|Y,Z}(x|y, z) = f_{X|Z}(x|z)$$

Conditional independence (like ordinary independence) has some important/useful properties/implications. Among these are

1. $X \underline{\perp\!\!\!\perp} Y \mid Z \Rightarrow Y \underline{\perp\!\!\!\perp} X \mid Z$
2. $X \underline{\perp\!\!\!\perp} Y \mid Z$ and $U = h(X) \Rightarrow U \underline{\perp\!\!\!\perp} Y \mid Z$
3. $X \underline{\perp\!\!\!\perp} Y \mid Z$ and $U = h(X) \Rightarrow X \underline{\perp\!\!\!\perp} Y \mid Z, U$
4. $X \underline{\perp\!\!\!\perp} Y \mid Z$ and $X \underline{\perp\!\!\!\perp} W \mid (Y, Z) \Rightarrow X \underline{\perp\!\!\!\perp} (W, Y) \mid Z$

5. $X \underline{\perp\!\!\!\perp} Y \mid Z$ and $X \underline{\perp\!\!\!\perp} Z \mid Y \Rightarrow X \underline{\perp\!\!\!\perp} (Y, Z)$

A possibly more natural (but equivalent) version of property 3. is

$$X \underline{\perp\!\!\!\perp} Y \mid Z \text{ and } U = h(X) \Rightarrow Y \underline{\perp\!\!\!\perp} (X, U) \mid Z$$

A main goal of this material is representing large joint distributions in graphical ways that allow one to "see" conditional independence relationships in the graphs.

5.2 Directed Graphs and Joint Probability Distributions

A **directed graph** (that might potentially represent some aspects of the joint distribution of $(X, Y, Z, \ldots)$) consists of **nodes** (or vertices) $X, Y, Z, \ldots$ and **arrows** (or edges) pointing between some of them.

5.2.1 Some Graph-Theoretic Concepts

For a graph with nodes/vertices $X, Y, Z, \ldots$

1. if an arrow points from X to Y we will say that X is a **parent** of Y and that Y is a **child** of X
2. a sequence of arrows beginning at X and ending at Y will be called a **directed path** from X to Y
3. if $X = Y$ or there is a directed path from X to Y, we will say that X is an **ancestor** of Y and Y is a **descendent** of X
4. if an arrow pointing in either direction connects X and Y they will be said to be **adjacent**
5. a sequence of adjacent vertices starting at X and ending at Y without reference to direction of any of the arrows will be called an **undirected path** from X to Y
6. an undirected path from X to Y has a **collider** at Z if there are two arrows in the path pointing to Z
7. a directed path that starts and ends at the same vertex is called a **cycle**
8. a directed graph is **acyclic** if it has no cycles

As a matter of notation/shorthand an acyclic directed graph is usually called a DAG (a directed acyclic graph) although the corresponding word order is not really as good as that corresponding to the unpronounceable acronym "ADG."

Example 1 ***A first DAG***

In Figure 1, X and Y are adjacent. X and Z are not adjacent. X is a parent of Y and an ancestor of W. There is a directed path from X to W and an undirected path from X to Z. Y is a collider on the path XYZ and is not a collider on the path XYW.

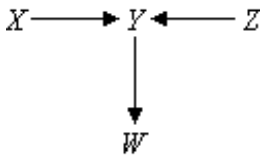

Figure 1: A First DAG

5.2.2 First Probabilistic Concepts and DAG's

For a vector of random variables and vertices $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ and a distribution F for $\boldsymbol{X}$, it is said that a DAG $\mathcal{G}$ **represents** F (or F is **Markov to** $\mathcal{G}$) if and only if

$$f_{\boldsymbol{X}}(x) = \prod_{i=1}^{k} f_{X_i|parents_i}(x_i|parents_i)$$

where

$$parents_i = \{\text{parents of } X_i \text{ in the DAG } \mathcal{G}\}$$

Example 2 ***More on the first DAG***
A joint distribution F for (X, Y, Z, W) is represented by the DAG pictured in Figure 1 if and only if

$$f_{X,Y,Z,W}(x, y, z, w) = f_X(x) f_Y(y) f_{Y|X,Z}(y|x, z) f_{W|Y}(w|y) \tag{50}$$

In `WinBUGS` there is the "Doodles" facility that allows one to input a model in terms of an associated DAG (augmented with information about specific forms of the conditional distributions). The joint distribution that is built by the software is then one represented by the Doodle DAG. Notice, for example, what a DAG tells one about how Gibbs sampling can be done. The DAG pictured in Figure 1 with guaranteed corresponding form (50) implies that when updating X one samples from a distribution specified by

$$f_X(\cdot) f_{Y|X,Z}(y_{\text{current}}|\cdot, z_{\text{current}})$$

updating of Z is similar, updating of Y is done sampling from a distribution specified by

$$f_{Y|X,Z}(\cdot|x_{\text{current}}, z_{\text{current}}) f_{W|Y}(w_{\text{current}}|\cdot)$$

and updating of W is done by sampling from a distribution specified by

$$f_{W|Y}(\cdot|y_{\text{current}})$$

A condition equivalent to the Markov condition can be stated in terms of conditional independence relationships. That is, let $\widetilde{X}_i$ stand for the set of all

vertices $X_1, X_2, \ldots, X_k$ in a DAG $\mathcal{G}$ except for the parents and descendents of X_i. Then

$$F \text{ is represented by } \mathcal{G} \Leftrightarrow \text{ for every vertex } X_i,\ X_i \perp\!\!\!\perp \widetilde{X}_i \,|\, parents_i \qquad (51)$$

Example 3 ***Yet more on the first DAG***

If a joint distribution F for (X, Y, Z, W) is represented by the DAG pictured in Figure 1, it follows that

$$X \perp\!\!\!\perp Z \quad and \quad W \perp\!\!\!\perp (X, Z) \,|\, Y$$

5.2.3 Some Additional Graph-Theoretic Concepts and More on Conditional Independence

Relationship (51) provides some conditional independence relationships implied by a DAG representation of a joint distribution F. Upon introducing some more machinery, other conditional independence relationships that will always hold for such F can sometimes be identified. These can be helpful for thinking about the nature of a large joint distribution.

Example 4 ***A second, more complicated DAG***

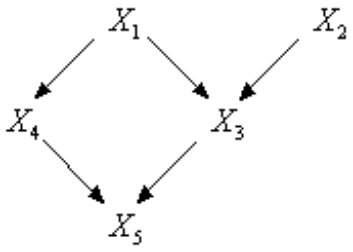

Figure 2: A Second DAG

Figure 2 provides a second example of a DAG. It follows from (51) that for F represented by the DAG in Figure 2, all of the following conditional independence relationships hold:

1. $X_1 \perp\!\!\!\perp X_2$
2. $X_2 \perp\!\!\!\perp (X_1, X_4)$
3. $X_3 \perp\!\!\!\perp X_4 \,|\, (X_1, X_2)$
4. $X_4 \perp\!\!\!\perp (X_2, X_3) \,|\, X_1$
5. $X_5 \perp\!\!\!\perp (X_1, X_2) \,|\, (X_3, X_4)$

But it is also true that

$$(X_4, X_5) \perp\!\!\!\perp X_2 \mid (X_1, X_3)$$

and that with proper additional machinery, this relationship can be read from the DAG.

The basic new graph-theoretic concepts needed concern connectedness and separatedness of vertices on a DAG. For a particular DAG, $\mathcal{G}$,

1. if X and Y are distinct vertices and Q a set of vertices not containing either X or Y, then we will say that X **and** Y **are d-connected given** Q if there is an undirected path P between X and Y such that

   (a) every collider on P has a descendent in Q, and

   (b) no other vertex (besides possibly those mentioned in (a)) on P is in Q.

2. if X and Y are not d-connected given Q, **they are d-separated given** Q.

3. if A, B, and Q are non-overlapping sets of vertices, $A \neq \emptyset$ and $B \neq \emptyset$, then we will say that A **and** B **are d-separated given** Q if every $X \in A$ and $Y \in B$ are d-separated given Q.

4. if A, B, and Q are as in 3. and A and B are not d-separated given Q then we will say that A **and** B **are d-connected given** Q.

Example 5 ***A third DAG (Example 17.9 of Wasserman)***

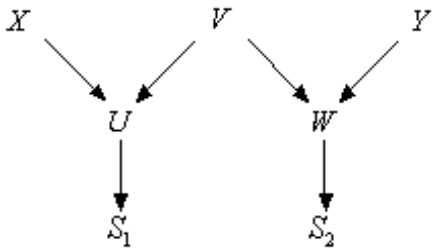

Figure 3: A Third DAG

In the DAG shown in Figure 3

1. X *and* Y *are d-separated given* $\emptyset$.

2. X *and* Y *are d-connected given* $\{S_1, S_2\}$.

3. X *and* Y *are d-connected given* $\{U, W\}$.

4. X and Y are d-separated given $\{S_1, S_2, V\}$.

The relationship between these graph-theoretic concepts and conditional independence for vertices of a DAG is then as follows. For disjoint sets of vertices A, B, and C of a DAG, $\mathcal{G}$, that represents a joint distribution F

$$A \underline{\perp\!\!\!\perp} B \,|\, C \Leftrightarrow A \text{ and } B \text{ are d-separated by } C \tag{52}$$

Example 6 ***More on the second DAG***

Consider a joint distribution F for X_1, X_2, X_3, X_4, and X_5 represented by the DAG shown in Figure 2. Take

$$A = \{X_4, X_5\}, B = \{X_2\}, \text{ and } C = \{X_1, X_3\}$$

Then

1. *X_4 and X_2 are d-separated given C,*
2. *X_5 and X_2 are d-separated given C, so*
3. *A and B are d-separated given C.*

Thus by (52) one may conclude that

$$(X_4, X_5) \underline{\perp\!\!\!\perp} X_2 \,|\, (X_1, X_3)$$

as suggested earlier.

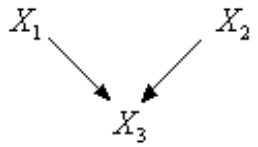

Figure 4: A Fourth DAG

Notice that in Figure 4, X_3 is a collider on the undirected path from X_1 to X_2, and X_1 and X_2 are d-connected given X_3. So in general, X_1 and X_2 will not be conditionally independent given X_3 for F represented by the DAG. This should not surprise us, given our experience with Bayes analysis. For example we know from Section 4.3.3 that in a Bayes model where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ has components that are iid $\mathrm{N}(\mu, \sigma^2)$ (conditioned on (μ, σ^2)), even where *a priori* μ and σ^2 are assumed to be independent, the posterior $g(\mu, \sigma^2 | \boldsymbol{y})$ will typically NOT be one of (conditional) independence (given $\boldsymbol{Y} = \boldsymbol{y}$). The DAG for this model is, of course, the version of Figure 4 shown in Figure 5.

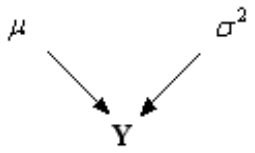

Figure 5: DAG for the 2 Parameter Single Sample Normal Bayes Model

5.3 Undirected Graphs and Joint Probability Distributions

An **undirected graph** (that might potentially represent some aspects of the joint distribution of $(X, Y, Z, \ldots)$) consists of **nodes** (or vertices) $X, Y, Z, \ldots$ and **edges** between some of the possible pairs of vertices. (Formally, one might think of edges as vertex pairs.)

5.3.1 Some Graph-Theoretic Concepts

Some of the terminology introduced above for directed graphs carries over to undirected graphs. And there are also some important additional concepts. For a graph with nodes/vertices $X, Y, Z, \ldots$

1. two vertices X and Y are said to be **adjacent** if there is an edge between them, and this will here be symbolized as $X \sim Y$
2. a sequence of vertices $\{X_1, X_2, \ldots, X_n\}$ is a **path** if $X_i \sim X_{i+1}$ for each i
3. if A, B, and C are disjoint sets of vertices, we will say that C **separates** A **and** B provided every path from a vertex $X \in A$ to a vertex $Y \in B$ contains an element of C
4. a **clique** is a set of vertices of a graph that are all adjacent to each other
5. a clique is **maximal** if it is not possible to add another vertex to it and still have a clique

Example 7 ***A first undirected graph***

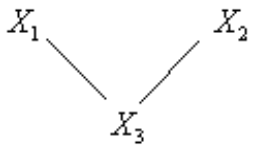

Figure 6: A First Undirected Graph

In Figure 6, X_1, X_2, and X_3 are vertices and there is one edge connecting X_1 and X_3 and another connecting X_2 and X_3.

Example 8 ***A second undirected graph***

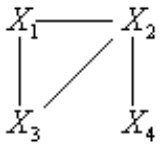

Figure 7: A Second Undirected Graph

In Figure 7

1. $\{X_1, X_3\}$ *and* $\{X_4\}$ *are separated by* $\{X_2\}$
2. $\{X_3\}$ *and* $\{X_4\}$ *are separated by* $\{X_2\}$
3. $\{X_1, X_2\}$ *is a clique*
4. $\{X_1, X_2, X_3\}$ *is a maximal clique*

5.3.2 Some Probabilistic Concepts and Undirected Graphs

Suppose that F is a joint distribution for $X_1, X_2, \ldots, X_k$. For each i and j let $\check{\boldsymbol{X}}_{ij}$ stand for all elements of $\{X_1, X_2, \ldots, X_k\}$ except elements i and j. We may associate with F a **pairwise Markov graph** $\mathcal{G}$ by

$$\text{failing to connect } X_i \text{ and } X_j \text{ with an edge if and only if } X_i \underline{\perp\!\!\!\perp} X_j \mid \check{\boldsymbol{X}}_{ij}$$

A pairwise Markov graph for F can in theory be made by considering only $\binom{k}{2}$ pairwise conditional independence questions. But as it turns out, many other conditional independence relationships can be read from it. That is, it turns out that if $\mathcal{G}$ is a pairwise Markov graph for F, then for non-overlapping sets of vertices A, B, and C,

$$C \text{ separates } A \text{ and } B \Rightarrow A \underline{\perp\!\!\!\perp} B \mid C \tag{53}$$

Example 9 ***A third undirected graph and conditional independence***

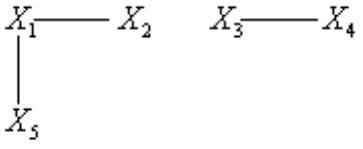

Figure 8: A Pairwise Markov (Undirected) Graph for F

If Figure 8 is a pairwise Markov graph for a distribution F for $X_1, X_2, \ldots, X_5$, we may conclude from (53) that

$$(X_1, X_2, X_5) \perp\!\!\!\perp (X_3, X_4) \quad \textit{and} \quad X_2 \perp\!\!\!\perp X_5 \,|\, X_1$$

Property (53) says that for a pairwise Markov (undirected) graph for F, separation implies conditional independence. Condition (52) says that for a DAG representing F, d-separation is equivalent to conditional independence. A natural question is whether the forward implication in (53) might be strengthened to equivalence. As it turns out, this is possible as follows. For F a joint distribution for $X_1, X_2, \ldots, X_k$ and $\mathcal{G}$ an undirected graph, we will say that F **is globally $\mathcal{G}$ Markov** provided for non-overlapping sets of vertices A, B, and C

$$C \text{ separates } A \text{ and } B \Leftrightarrow A \perp\!\!\!\perp B \,|\, C$$

Then as it turns out,

F is globally $\mathcal{G}$ Markov $\Leftrightarrow$ $\mathcal{G}$ is a pairwise Markov graph associated with F

so that separation on a pairwise Markov graph is equivalent to conditional independence.

Example 10 ***A fourth undirected graph and conditional independence***

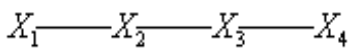

Figure 9: A Second (Undirected) Pairwise Markov Graph

Whether one thinks of Figure 9 as a pairwise Markov graph $\mathcal{G}$ associated with F or thinks of F as globally $\mathcal{G}$ Markov, it follows (for example) that

$$X_1 \perp\!\!\!\perp X_3 \,|\, X_2 \quad \textit{and} \quad X_1 \perp\!\!\!\perp X_4 \,|\, X_2$$

There remains to consider what connections there might be between an undirected graph related to F and a functional form for F. It turns out that subject to some other (here unspecified) technical conditions, a distribution F for $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ is globally $\mathcal{G}$ Markov if and only if there are positive functions ψ_C such that

$$f_{\boldsymbol{X}}(\boldsymbol{x}) \propto \prod_{C \in \mathcal{C}} \psi_C(C)$$

where $\mathcal{C}$ is the set of maximal cliques associated with $\mathcal{G}$. (Any vertices that share no edges get their own individual factors in this kind of product.)

Example 11 ***(Example 18.7 of Wasserman) Another undirected graph and the form of $f_{\boldsymbol{X}}$***

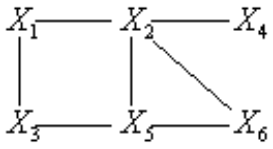

Figure 10: Another (Undirected) Pairwise Markov Graph

The set of maximal cliques associated with the undirected graph $\mathcal{G}$ in Figure 10 is

$$\mathcal{C} = \{\{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_5, X_6\}, \{X_2, X_4\}, \{X_3, X_5\}\}$$

So (subject to some technical conditions) F is globally $\mathcal{G}$ Markov if and only if

$$f_{\boldsymbol{X}}(\boldsymbol{x}) \propto \psi_{12}(x_1, x_2)\, \psi_{13}(x_1, x_3)\, \psi_{24}(x_2, x_4)\, \psi_{35}(x_3, x_5)\, \psi_{256}(x_2, x_5, x_6)$$

for some positive functions $\psi_{12}, \psi_{13}, , \psi_{24}, \psi_{35}$, and ψ_{256}.

6 The Practice of Bayes Inference 3: (Mostly) Multi-Sample Models

Essentially everything that is done in M.S. level statistical methods courses like Stat 500 and Stat 511 (and more besides) can be recast in a Bayes framework and addressed using the kind of methods discussed thus far in this outline. We proceed to indicate how some of these analyses can be built.

6.1 Two-Sample Normal Models (and Some Comments on "Nested" Models)

Several versions of two-sample univariate normal models are possible. That is, suppose that observable $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ consists of iid univariate $\mathrm{N}(\mu_1, \sigma_1^2)$ variables $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ independent of iid univariate $\mathrm{N}(\mu_2, \sigma_2^2)$ variables $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$. The joint pdf of $\boldsymbol{Y}$ is then

$$\begin{aligned} f\left(\boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) &= \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n_1} \exp\left(-\frac{\sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2}{2\sigma_1^2}\right) \\ &\quad \times \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{n_2} \exp\left(-\frac{\sum_{j=1}^{n_2}(y_{2j}-\mu_2)^2}{2\sigma_2^2}\right) \end{aligned} \quad (54)$$

Depending then upon what one wishes to assume about the 4 parameters μ_1, μ_2, σ_1^2, and σ_2^2 there are submodels of the full model specified by this joint

density (54) that might be considered. There is the full 4-parameter model that we will here term model $\mathcal{M}_1$ with likelihood

$$L\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) = f\left(\boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$$

It is fairly common to make the model assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, thereby producing a 3-parameter model that we will call model $\mathcal{M}_2$ with likelihood

$$L\left(\mu_1, \mu_2, \sigma^2\right) = f\left(\boldsymbol{y}|\mu_1, \mu_2, \sigma^2, \sigma^2\right)$$

In both models $\mathcal{M}_1$ and $\mathcal{M}_2$, primary interest usually centers on how μ_1 and μ_2 compare. The assumption $\mu_1 = \mu_2$ imposed on μ_1 and μ_2 in model $\mathcal{M}_2$ produces the one sample univariate normal model of Section 4.3 that we might call model $\mathcal{M}_3$.

"Obvious" priors for $\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$ in model $\mathcal{M}_1$ or $\left(\mu_1, \mu_2, \sigma^2\right)$ in model $\mathcal{M}_2$ can be built using the pieces introduced in Section 4.3. In particular, priors (improper or proper) of "independence" (of product form) for the parameters seem attractive/simple, where

1. means are *a priori* either "iid" uniform on $\Re$ (or some large interval) or are iid normal (typically with large variance), and
2. logvariance(s) is (are) *a priori* either uniform on $\Re$ (or some large interval) or variance(s) is (are) inverse gamma, i.e. scaled inverse χ^2 (typically with small degrees of freedom).

Models $\mathcal{M}_1, \mathcal{M}_2$, and $\mathcal{M}_3$ are nested. But under priors like those just suggested for $\mathcal{M}_1$, $\mathcal{M}_2$ (and therefore $\mathcal{M}_3$) has prior and posterior probability 0. So a Bayes "test of $\mathcal{M}_2$ in model $\mathcal{M}_1$" can never decide in favor of "exactly $\mathcal{M}_2$." Similarly, under priors like those just suggested for $\mathcal{M}_2$, $\mathcal{M}_3$ has prior and posterior probability 0. So a Bayes "test of $\mu_1 = \mu_2$ in model $\mathcal{M}_2$" can never decide in favor of "exactly $\mu_1 = \mu_2$."

This situation is a simple illustration of the fact that in a Bayesian context, rational consideration of whether a lower dimensional submodel of the working model is plausible must be typically be done by either explicitly placing positive probability on the submodel or by taking some other approach. The Bayes factors of Section 3.5 can be employed. Or sticking strictly to calculations with the working model, one can assess the posterior probability "near" the submodel.

Take for explicit example the case of working model $\mathcal{M}_2$ and submodel $\mathcal{M}_3$. If one wants to allow for positive posterior probability to be assigned to the submodel, one will need to do something like assign prior probability p to the working model and then a prior distribution for μ_1, μ_2, and σ^2 in the working model, together with prior probability $1 - p$ to the submodel and then a prior distribution for $\mu = \mu_1 = \mu_2$, and σ^2 in the submodel. Lacking this kind of explicit weighting of $\mathcal{M}_2$ and $\mathcal{M}_3$, one might find a Bayes factor for comparing Bayes models for $\mathcal{M}_2$ and $\mathcal{M}_3$. Or, working entirely within model $\mathcal{M}_2$, one might simply find a posterior distribution of $\mu_1 - \mu_2$ and investigate how much posterior probability for this variable there is near the value $\mu_1 - \mu_2 = 0$.

6.2 r-Sample Normal Models

This is the natural generalization of the two sample normal model just discussed. $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_r)$ is assumed to consist of r independent vectors, where $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ has iid $\mathrm{N}\left(\mu_i, \sigma_i^2\right)$ components. The joint pdf for $\boldsymbol{Y}$ is then

$$f\left(\boldsymbol{y}|\mu_1, \ldots, \mu_r, \sigma_1^2, \ldots, \sigma_r^2\right) = \prod_{i=1}^{r} \left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right)^{n_i} \exp\left(-\frac{\sum_{j=1}^{n_i} \left(y_{ij} - \mu_i\right)^2}{2\sigma_i^2}\right)$$

and the most commonly used version of this model is one where one assumes that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$ and thus has a model with $r + 1$ parameters and likelihood

$$L\left(\mu_1, \ldots, \mu_r, \sigma^2\right) = f\left(\boldsymbol{y}|\mu_1, \ldots, \mu_r, \sigma^2, \ldots, \sigma^2\right)$$

Exactly as in the two-sample case, "obvious" priors for $\left(\mu_1, \ldots, \mu_r, \sigma^2\right)$ can be built using the pieces introduced in Section 4.3. In particular, priors (improper or proper) of "independence" (of product form) for the parameters seem attractive/simple, where

1. means are *a priori* either "iid" uniform on $\Re$ (or some large interval) or are iid normal (typically with large variance), and
2. $\ln \sigma^2$ is *a priori* either uniform on $\Re$ (or some large interval) or σ^2 is inverse gamma, i.e. scaled inverse χ^2 (typically with small degrees of freedom).

(Of course, if one doesn't wish to make the constant variance assumption, it is possible to make use independent priors of the type in 2. above for r different variances.)

6.3 Normal Linear Models (Regression Models)

The first half of Stat 511 concerns statistical analysis based on the linear model

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times k}{\boldsymbol{X}}\underset{k\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

for $\boldsymbol{\epsilon} \sim \mathrm{MVN}_n\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right)$ and known matrix $\boldsymbol{X}$ (that for present purposes we will assume has full rank). This implies that $\boldsymbol{Y} \sim \mathrm{MVN}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right)$ so that this model has parameters $\boldsymbol{\beta}$ and σ^2 and likelihood

$$L\left(\boldsymbol{\beta}, \sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right)$$

(With proper choice of $\boldsymbol{X}$ this is well known to include the constant variance cases of the two- and r-sample normal models just discussed.)

The most obvious priors for $\left(\boldsymbol{\beta}, \sigma^2\right)$ are of a product/independence form, where

1. $\boldsymbol{\beta}$ is either uniform on $\Re^k$ (or some large k-dimensional rectangle) or is MVN_k (typically with large covariance matrix), and
2. $\ln \sigma^2$ is *a priori* either uniform on $\Re$ (or some large interval) or σ^2 is inverse gamma, i.e. scaled inverse χ^2 (typically with small degrees of freedom).

The textbook in Section 14.2 considers the case of the improper prior

$$g\left(\boldsymbol{\beta}, \sigma^2\right) \propto \frac{1}{\sigma^2} \tag{55}$$

and argues that with this choice, the conditional distribution of $\sigma^2 | \boldsymbol{Y} = \boldsymbol{y}$ is

$$\text{Inv-}\chi^2\left(n-k, s^2\right)$$

where

$$s^2 = \frac{1}{n-k}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$$

for $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$ the least squares estimate of $\boldsymbol{\beta}$. (s^2 is the usual linear model estimate of σ^2.) Further, the conditional distribution of $\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) | \boldsymbol{Y} = \boldsymbol{y}$ is multivariate t. That is, the posterior distribution of $\boldsymbol{\beta}$ is that of

$$\hat{\boldsymbol{\beta}} + s\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)^{1/2}\sqrt{\frac{n-k}{W}}\boldsymbol{Z}$$

where $\boldsymbol{Z}$ is a k-vector of iid $\text{N}(0,1)$ random variables, independent of $W \sim \chi^2_{n-k}$ and $\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)^{1/2}$ is a matrix square root of $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$.

Further, if $\boldsymbol{x}_{\text{new}}$ is $k \times 1$ and one has not yet observed

$$y_{\text{new}} = \boldsymbol{x}'_{\text{new}}\boldsymbol{\beta} + \epsilon_{\text{new}}$$

for ϵ_{new} independent of $\underset{n\times 1}{\boldsymbol{\epsilon}}$ with mean 0 and variance $\gamma\sigma^2$, one might consider the posterior predictive distribution of y_{new} based on the improper prior (55). As it turns out, the posterior distribution is that of

$$\boldsymbol{x}'_{\text{new}}\hat{\boldsymbol{\beta}} + s\sqrt{\boldsymbol{x}'_{\text{new}}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_{\text{new}} + \gamma}\; T$$

for $T \sim t_{n-k}$.

The upshot of all this is that Bayes credible intervals for model parameters (and linear combinations of elements of the vector $\boldsymbol{\beta}$) and future observations based on the improper prior (55) for *any full rank linear model* are the same as confidence intervals based on the usual Stat 511 linear model theory. (This, of course, is true for the (constant variance) one-, two-, and r-sample normal models, as they are instances of this model.) A clear advantage of taking this Bayes point of view is that beyond the "ordinary" inference formulas of Stat 511, one can easily simulate the posterior distribution of any parametric function $h\left(\boldsymbol{\beta}, \sigma^2\right)$ and provide credible sets for this.

6.4 One-Way Random Effects Models

The standard r-sample normal model of Section 6.2 is often written in the form

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{56}$$

where the ϵ_{ij} are iid $\mathrm{N}\left(0, \sigma^2\right)$ (and as before, $\mu_1, \ldots, \mu_r, \sigma^2$ are unknown parameters). The one-way random effects model treats $\mu_1, \ldots, \mu_r$ as unobservable iid random draws from a $\mathrm{N}\left(\mu, \sigma_\tau^2\right)$ distribution, producing a model with parameters

$$\mu, \sigma_\tau^2, \text{ and } \sigma^2 \tag{57}$$

Considering only the observables Y_{ij}, the joint distribution of these is multivariate normal where the mean vector is $\mu\mathbf{1}$ and the covariance matrix has diagonal entries

$$\mathrm{Var} Y_{ij} = \sigma_\tau^2 + \sigma^2$$

and off-diagonal elements

$$\mathrm{Cov}\left(Y_{ij}, Y_{ij'}\right) = \sigma_\tau^2 \ \ (\text{if } j \neq j') \ \text{ and } \ \mathrm{Cov}\left(Y_{ij}, Y_{i'j'}\right) = 0 \ \ (\text{if } i \neq i')$$

However, particularly for purposes of setting up a `WinBUGS` simulation from a posterior distribution, it is often very convenient to use the unobservable auxiliary variables $\mu_1, \ldots, \mu_r$. (See Section 3.7 regarding the use of auxiliary variables.) Further, just as in classical treatments of this model that often include prediction of the random effects, there may be independent interest in the realized but unobservable values $\mu_1, \ldots, \mu_r$.

Whether one models only in terms of the observables or includes the unobservable $\mu_1, \ldots, \mu_r$, ultimately the one-way random effects model has only the 3 parameters (57). Once again, choices of joint priors (improper or proper) of "independence" (of product form) for the parameters seem attractive/simple, where

1. μ is *a priori* either uniform on $\Re$ (or some large interval) or is normal (typically with large variance), and
2. $\ln \sigma_\tau^2$ and $\ln \sigma^2$ are *a priori* either uniform on $\Re$ (or some large interval) or variances σ_τ^2 and σ^2 are inverse gamma, i.e. scaled inverse χ^2 (typically with small degrees of freedom).

The formal connection (56) between the one-way random effects model here and the r-sample normal model in Section 6.2 invites consideration of how the present modeling might be appropriate in the earlier r-sample (fixed effects) context. Formally, a Bayes version of the present one-way random effects model with unobservable auxiliary variables $\mu_1, \ldots, \mu_r$ might be thought of as an alternative Bayes model for the r-sample normal situation, where instead of the prior of independence for the r means, one uses a prior of conditional independence given parameters μ and σ_τ^2, and puts priors on these. This kind of modeling might be termed use of a **two stage prior** or **hyper-prior** in

the fixed effects model. (The values μ and σ_τ^2, as parameters of the first-level prior on the r means that themselves get prior distributions, are termed **hyperparameters** in this kind of language.) The ultimate effect of such modeling is that instead of making the r means *a priori* independent, they are dependent. Posterior means for $\mu_1, \ldots, \mu_r$ tend to be shrunken from the r sample means toward a common compromise value representing one's perception of μ (whereas if a proper normal prior for them is used in the style of the discussion of Section 6.2, the shrinking is towards the known prior mean).

This discussion should not be allowed to obscure the basic fact that the "data models," corresponding parameters, and likelihoods here and in Section 6.2 are fundamentally different, a point that is very important to philosophical anti-Bayesians. (It is much less important to most Bayesians, for whom the distinction between unknown parameters and unobserved auxiliary variables is quite unimportant, if not completely artificial.) In the r-sample model there are r unknown means and one unknown variance parameter. Here there is a single unknown mean and two unknown variance parameters.

6.5 Hierarchical Models (Normal and Others)

The kind of hierarchy of parameters and auxiliary variables just illustrated in the one-way random effects model can be generalized/extended in at least two directions. First, more levels of hierarchy can be appropriate. Second, the conditional distributions involved can be other than normal. This section provides a small introduction to these possibilities.

We consider a context where (more or less in "tree diagram fashion") each level of some factor A gives rise to levels (peculiar to the given level of A) of a factor B, which in turn each gives rise to levels (peculiar to the given level of B within A) of a factor C, etc. and at the end of each branch, there is some kind of observation. For example, heats of steel (A) could be poured into ingots (B), which are in turn are cut into specimens (C), on which carbon content is measured. Or work weeks (A) have days (B), which have in them hours of production (C), in which items (D) are produced and subjected to some final product test like a "blemish count." Notice that in the first of these examples, a normal measurement (of carbon content) might ultimately be made, while in the second, a Poisson model for each blemish count might be appropriate.

To be slightly more concrete, let us consider a hierarchical situation involving factors A, B, and C, with (possibly multivariate)

$$Y_{ijk} = \text{the data observed at level } k \text{ of factor C within level } j \text{ of factor B within level } i \text{ of factor A}$$

A hierarchical model for the entire set of observables is then constructed as follows. Suppose that the distribution of Y_{ijk} depends upon some parameter $\boldsymbol{\gamma}_{ijk}$ and possibly a parameter $\boldsymbol{c}$, and that conditional on the $\boldsymbol{\gamma}_{ijk}$'s and $\boldsymbol{c}$, the Y_{ijk}'s are independent. Then, in the obvious (abused) notation, a conditional

density for the observables becomes

$$f(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{c}) = \prod_{i,j,k} f\left(\boldsymbol{y}_{ijk}|\boldsymbol{\gamma}_{ijk}, \boldsymbol{c}\right)$$

Then we suppose that for some parameters $\boldsymbol{\beta}_{ij}$ and possibly a parameter $\boldsymbol{b}$, the $\boldsymbol{\gamma}_{ijk}$' are conditionally independent, the distribution of each $\boldsymbol{\gamma}_{ijk}$ governed by its $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{b}$. That is, the conditional density for the $\boldsymbol{\gamma}_{ijk}$'s becomes (again in obvious notation)

$$f(\boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{b}) = \prod_{i,j,k} f\left(\boldsymbol{\gamma}_{ijk}|\boldsymbol{\beta}_{ij}, \boldsymbol{b}\right)$$

And then we suppose that for some parameters $\boldsymbol{\alpha}_i$ and possibly a parameter $\boldsymbol{a}$ the $\boldsymbol{\beta}_{ij}$'s are conditionally independent, the distribution of each $\boldsymbol{\beta}_{ij}$ governed by its $\boldsymbol{\alpha}_i$ and $\boldsymbol{a}$. So the conditional density for the $\boldsymbol{\beta}_{ij}$'s becomes

$$f(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{a}) = \prod_{i,j} f\left(\boldsymbol{\beta}_{ij}|\boldsymbol{\alpha}_i, \boldsymbol{a}\right)$$

Finally, we suppose that conditional on a parameter vector $\boldsymbol{\theta}$, the $\boldsymbol{\alpha}_i$ are conditionally independent. So the conditional density for $\boldsymbol{\alpha}_i$'s becomes

$$f(\boldsymbol{\alpha}|\boldsymbol{\theta}) = \prod_{i} f(\boldsymbol{\alpha}_i|\boldsymbol{\theta})$$

The joint density for all of the Y_{ijk}'s, $\boldsymbol{\gamma}_{ijk}$'s, and $\boldsymbol{\beta}_{ij}$'s and $\boldsymbol{\alpha}_i$'s is then

$$f(\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}|\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{a}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{c})\, f(\boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{b})\, f(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{a})\, f(\boldsymbol{\alpha}|\boldsymbol{\theta}) \tag{58}$$

Notice that this form is consistent with a directed graph representing the joint distribution of the Y_{ijk}'s, $\boldsymbol{\gamma}_{ijk}$'s, and $\boldsymbol{\beta}_{ij}$'s and $\boldsymbol{\alpha}_i$'s where

1. each Y_{ijk} has parent $\boldsymbol{\gamma}_{ijk}$
2. each $\boldsymbol{\gamma}_{ijk}$ has parent $\boldsymbol{\beta}_{ij}$
3. each $\boldsymbol{\beta}_{ij}$ has parent $\boldsymbol{\alpha}_i$

This is illustrated in the small example in Figure 11.

The hierarchical form indicated in (58) has parameter $\boldsymbol{\theta}$ (and possibly parameters $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$). A Bayes analysis of a hierarchical data structure then requires specifying a prior for $\boldsymbol{\theta}$ and if relevant $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$. This would put $\boldsymbol{\theta}$ onto a directed graph like that in Figure 11 as a parent of all $\boldsymbol{\alpha}_i$'s, $\boldsymbol{a}$ as a parent of all $\boldsymbol{\beta}_{ij}$'s, $\boldsymbol{b}$ as a parent of all $\boldsymbol{\gamma}_{ijk}$'s, and $\boldsymbol{c}$ as a parent of all Y_{ijk}'s. The Bayes modeling thus breaks the independence of the two main branches of the directed graph in Figure 11 and makes all of the data relevant in inferences about all of the quantities represented on the figure and all the parameters.

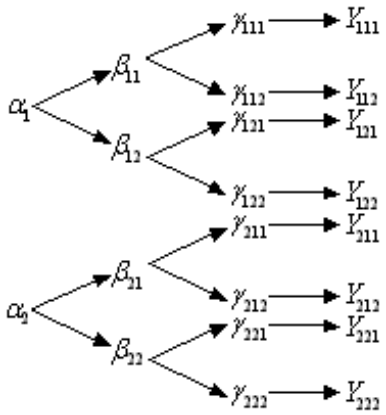

Figure 11: A Small Hierarchical Structure and Directed Graph

6.6 Mixed Linear Models (in General) (and Other MVN Models With Patterned Means and Covariance Matrices)

The normal one-way random effects model of Section 6.4 is not only a special case of the hierarchical modeling just discussed in Section 6.5, it is a special case of the so called mixed linear model of Stat 511. That is, it is also a special case of a model that is usually represented as

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times k}{\boldsymbol{X}}\underset{k\times 1}{\boldsymbol{\beta}} + \underset{n\times q}{\boldsymbol{Z}}\underset{q\times 1}{\boldsymbol{u}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known matrices, $\boldsymbol{\beta}$ is a parameter vector and

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \text{MVN}_{q+n}\left(\boldsymbol{0}, \begin{pmatrix} \underset{q\times q}{\boldsymbol{G}} & \underset{q\times n}{\boldsymbol{0}} \\ \underset{n\times q}{\boldsymbol{0}} & \underset{n\times n}{\boldsymbol{R}} \end{pmatrix}\right)$$

from which $\boldsymbol{Y}$ is multivariate normal with

$$\text{E}\boldsymbol{Y} = \boldsymbol{X\beta} \text{ and } \text{Var}\boldsymbol{Y} = \boldsymbol{ZGZ}' + \boldsymbol{R} \equiv \boldsymbol{V}$$

In typical applications of this model, the covariance matrix $\boldsymbol{V}$ is a patterned function of several variance components, say $\boldsymbol{\sigma}^2 = \left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2\right)$, and we might then write $\boldsymbol{V}\left(\boldsymbol{\sigma}^2\right)$. This then produces a likelihood based on the multivariate normal density

$$L\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2\right) \propto \left|\det \boldsymbol{V}\left(\boldsymbol{\sigma}^2\right)\right|^{-1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)' \boldsymbol{V}\left(\boldsymbol{\sigma}^2\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right)$$

As in Section 6.3 the most obvious priors for $\left(\boldsymbol{\beta}, \sigma^2\right)$ are of a product/independence form, where

1. $\boldsymbol{\beta}$ is either uniform on $\Re^k$ (or some large k-dimensional rectangle) or is MVN_k (typically with large covariance matrix), and

2. each $\ln \sigma_i^2$ is *a priori* either uniform on $\Re$ (or some large interval) or σ_i^2 is inverse gamma, i.e. scaled inverse χ^2 (typically with small degrees of freedom).

Notice that although only $\boldsymbol{Y}$ is observable, just as noted in the specific mixed model of Section 6.4, there may be good reasons to include the vector of random effects $\boldsymbol{u}$ in a posterior simulation. There may be independent interest in these. And since common models for $\boldsymbol{u}$ make its components independent (and thus $\boldsymbol{G}$ diagonal) and simply assembles linear combinations of these in the definition of the entries of $\boldsymbol{Y}$, the coding of a model that includes these variables may be operationally much simpler than coding of the multivariate normal form for $\boldsymbol{Y}$ alone.

One way to look at the mixed linear model is that it is a multivariate normal model with both mean vector and covariance matrix that are parametric functions. There are problems where one observes one or more multivariate normal vectors that don't fit the completely-unrestricted-mean-and-covariance-matrix context of Section 4.4 or the linear-mean-and-patterned-function-of-variances context of mixed linear models. Instead, for some parameter vector $\boldsymbol{\theta}$ and parametric forms for a mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, one observes $n \geq 1$ iid multivariate normal vectors $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ and has likelihood

$$L(\boldsymbol{\theta}) \propto |\det \boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))\right)$$

Consideration of the particulars of the situation being modeled and physical meanings of the coordinates of $\boldsymbol{\theta}$ can then sometimes be called on to produce a plausible prior for $\boldsymbol{\theta}$ and then a Bayes analysis.

6.7 Non-Linear Regression Models, etc.

A natural generalization of the linear model discussed in Section 6.3 (and a special case of the parametric mean vector and covariance matrix multivariate normal inference problem just alluded to) is a model where the means of n independent univariate normal observations y_i depend upon corresponding k-vectors of predictors $\boldsymbol{x}_i$ and some parameter vector $\boldsymbol{\beta}$ through a function $m(\boldsymbol{x}, \boldsymbol{\beta})$, and the variances are some constant σ^2. This is usually written as

$$y_i = m(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i$$

where the ϵ_i are iid $\text{N}(0, \sigma^2)$. This produces a likelihood that is

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_i - m(\boldsymbol{x}_i, \boldsymbol{\beta}))^2\right)$$

(The usual normal linear model is the case where $\boldsymbol{x}$ and $\boldsymbol{\beta}$ have the same dimension and $m(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{x}_i'\boldsymbol{\beta}$.)

For a Bayes analysis in this context, a prior distribution is needed for $\left(\boldsymbol{\beta}, \sigma^2\right)$. A product (independence between $\boldsymbol{\beta}$ and σ^2) form seems most obvious where

1. consideration of the particulars of the situation being modeled and physical meanings of the coordinates of $\boldsymbol{\beta}$ can then sometimes be called on to produce a plausible prior for $\boldsymbol{\beta}$, and

2. $\ln \sigma^2$ is *a priori* either uniform on $\Re$ (or some large interval) or σ^2 is inverse gamma, i.e. scaled inverse χ^2 (typically with small degrees of freedom).

The main point here is that operationally, where non-Bayesian analyses for the linear and non-linear regression models are quite different (for example different software and theory), Bayes analyses for the linear and non-linear regression models are not substantially different.

Notice too that mixed effects versions of non-linear regression models are available by assuming that $\boldsymbol{\epsilon}$ is a $\text{MVN}_n\left(\mathbf{0}, \boldsymbol{V}\left(\boldsymbol{\sigma}^2\right)\right)$ vector of random effects with patterned covariance matrix $\boldsymbol{V}\left(\boldsymbol{\sigma}^2\right)$ depending upon a vector of variance components $\boldsymbol{\sigma}^2 = \left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2\right)$. The parameters for which one needs to specify a prior are $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$, and this is a special case of the "Other MVN Models With Patterned Means and Covariance Matrices" discussion of the previous section.

6.8 Generalized Linear Models, etc.

The intent of the so-called "generalized linear model" introduced in Stat 511 is to extend regression/linear models type modeling of the effects of covariates beyond the realm of normal observations, particularly to cases of discrete (binomial and Poisson) responses. In the generalized linear model, one assumes that n independent univariate (binomial or Poisson) observations y_i have distributions depending upon corresponding k-vectors of predictors $\boldsymbol{x}_i$ and some parameter vector $\boldsymbol{\beta}$ through some appropriate "link" function $h(\cdot)$. That is, one assumes that

$$\text{E}y_i = h^{-1}\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)$$

Probably the most common Poisson version of the generalized linear model is the case where one assumes that

$$\text{E}y_i = \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)$$

which is the case of the so-called "log-linear model." Notice that the joint pmf for an n-vector of observations under the log-linear model is then

$$f\left(\boldsymbol{y}|\boldsymbol{\beta}\right) = \prod_{i=1}^{n} \frac{\exp\left(-\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)\left(\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{y_i}}{y_i!}$$

(more generally, one replaces $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ with $h^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})$). So the likelihood under the log-linear model is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))(\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))^{y_i}}{y_i!}$$

and upon making some choice of (proper) prior distribution for $\boldsymbol{\beta}$, say MVN_k with large covariance matrix or uniform on a large but bounded part of $\Re^k$ (one would need to think about whether an improper "flat" prior on $\Re^k$ for $\boldsymbol{\beta}$ will produce a proper posterior), a Bayes analysis will proceed as usual.

This could be easily extended to a mixed effects version by assuming that for $\boldsymbol{\epsilon}$ some $\text{MVN}_n\left(\mathbf{0}, \boldsymbol{V}\left(\boldsymbol{\sigma}^2\right)\right)$ vector of random effects with patterned covariance matrix $\boldsymbol{V}\left(\boldsymbol{\sigma}^2\right)$ depending upon a vector of variance components $\boldsymbol{\sigma}^2 = \left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2\right)$, conditional on $\boldsymbol{\epsilon}$ the y_i are independent with

$$y_i \sim \text{Poisson}(\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i))$$

This would produce a model with parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ that could be handled in `WinBUGS` by including in the analysis the auxiliary variables in $\boldsymbol{\epsilon}$ (or likely even more fundamental independent mean 0 normal random effects that when added appropriately produce $\boldsymbol{\epsilon}$ with the desired patterned covariance matrix). That is, in principle, there is no special difficulty involved in handling regression type or even mixed effects type modeling and analysis of Poisson responses from a Bayes viewpoint.

Common binomial versions of the generalized linear model set the binomial "success probability" parameter p to be

$$p_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

(the case of so-called "logistic regression") or

$$p_i = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$$

(the case of so-called "probit analysis") or

$$p_i = 1 - \exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$$

(the case of the "complimentary log log" link). Under any of these, a joint pmf for n independent binomial observations y_i is

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

and the likelihood is thus

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

Again upon making some prior assumption (like multivariate normal, uniform on a subset of $\Re^k$ or possibly uniform on all of $\Re^k$) on $\boldsymbol{\beta}$, a Bayes analysis is in principle straightforward. And just as discussed above in the Poisson case, the generalization to a random or mixed effects version is available by replacing $\boldsymbol{x}_i'\boldsymbol{\beta}$ with $\boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$ in any of the expressions for p_i above.

Finally, notice that from the point of view of simulation-based Bayes analysis that doesn't require that one develop specialized inference methods or distribution theory before doing statistical analysis, there is not even anything special about the linear form $\boldsymbol{x}_i'\boldsymbol{\beta}$ appearing in the expressions of this section. It is conceptually no more difficult to replace $\boldsymbol{x}_i'\boldsymbol{\beta}$ with an expression like $m(\boldsymbol{x}_i, \boldsymbol{\beta})$ than it was in the normal non-linear regression case of Section 6.7.

6.9 Models With Order Restrictions

The following is a bit of an amplification of the discussion of Section 3.3.3. As indicated in that section, if a natural parameter space $\Theta \subset \Re^k$ is of product form but some $\Theta_0 \subset \Theta$ that is not of a product form is of real interest, direct MCMC simulation from a posterior on Θ_0 may not be obvious. But if $h(\boldsymbol{\theta})$ specifies a posterior on Θ, one can sample from the posterior specified by

$$h(\boldsymbol{\theta})\, I[\boldsymbol{\theta} \in \Theta_0]$$

by sampling instead from $h(\boldsymbol{\theta})$ and simply "throwing away" those MCMC iterates $\boldsymbol{\theta}^j$ that do not belong to Θ_0. As indicated in Section 3.3.3 this can be done in `WinBUGS` using `coda` to transfer iterates to `R`.

The other way to address this kind of issue is to find a parameterization that avoids it altogether. Consider, for example, what is possible for the common type of order restriction

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$$

1. Where $\Theta = \Re^k$, one can define

   $$\delta_i = \theta_i - \theta_{i-1} \quad \text{for } i = 2, 3, \ldots, k$$

   (so that $\theta_j = \theta_1 + \sum_{i=2}^{j} \delta_i$ for $j \geq 2$) and replace the parameter vector $\boldsymbol{\theta}$ with the parameter vector $(\theta_1, \delta_2, \ldots, \delta_k) \in \Re \times [0, \infty)^{k-1}$. Placing a prior distribution of product form on $\Re \times [0, \infty)^{k-1}$ leads to a posterior on a product space and straightforward posterior simulation.

2. Where $\Theta = (0, \infty)^k$, one can do essentially as in 1., or parametrize in ratio form. That is, with

   $$r_i = \frac{\theta_i}{\theta_{i-1}} \quad \text{for } i = 2, 3, \ldots, k$$

   (so that $\theta_j = \theta_1 \cdot \prod_{i=2}^{j} r_i$ for $j \geq 2$), one may replace the parameter vector $\boldsymbol{\theta}$ with the parameter vector $(\theta_1, r_2, \ldots, r_k) \in (0, \infty) \times [1, \infty)^{k-1}$. Placing a prior distribution of product form on $(0, \infty) \times [1, \infty)^{k-1}$ leads to a posterior on a product space and straightforward posterior simulation.

3. Where $\Theta = (0,1)^k$, a modification of the ratio idea can be used. That is, with
$$d_i = \frac{\theta_i}{\theta_{i+1}} \quad \text{for } i = 1, 2, \ldots, k-1$$
(so that $\theta_j = \theta_k \cdot \prod_{i=j}^{k-1} d_i$ for $j \leq k-1$), one may replace the parameter vector $\boldsymbol{\theta}$ with the parameter vector $(d_1, d_2, \ldots, d_{k-1}, \theta_k) \in (0,1]^k$. Placing a prior distribution of product form on $(0,1]^k$ leads to a posterior on a product space and straightforward posterior simulation.

Of course the reparameterization ideas above are not specifically or essentially Bayesian, but they are especially helpful in the Bayes context.

6.10 One-Sample Mixture Models

For pdf's $f_1, f_2, \ldots, f_M$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_M)$ a probability vector (each $\alpha_i \geq 0$ and $\sum_{i=1}^M \alpha_i = 1$), a pdf
$$f_{\boldsymbol{\alpha}} = \sum_{i=1}^M \alpha_i f_i \tag{59}$$
specifies a so-called "mixture distribution." In cases where the f_i are completely specified and linearly independent functions, (frequentist or) Bayes estimation of $\boldsymbol{\alpha}$ is straightforward. On the other hand, where each f_i is parameterized by a (potentially multivariate) parameter $\boldsymbol{\gamma}_i$ and the whole vector
$$\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_M)$$
is unknown, the problem is typically technically more difficult.

In the first place, there are often identifiability problems (see Section 3.3.1) unless one is careful. For example, as suggested in Section 3.3.1, in a problem where $M = 2$, f_1 is $\mathrm{N}(\mu_1, \sigma_1^2)$ and f_2 is $\mathrm{N}(\mu_2, \sigma_2^2)$, with all of $\alpha_1, \mu_1, \sigma_1^2, \mu_2$, and σ_2^2 unknown, the parameter vectors
$$(\alpha_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = (.3, 1, 1, 2, 1)$$
and
$$(\alpha_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = (.7, 2, 1, 1, 1)$$
produce the same mixture distribution. In order to avoid this kind of difficulty, one must do something like parameterize not by the two means, but rather by the smaller of the two means and the difference between the larger and the smaller of the means.

The form (59) can be thought of as the density of an observable Y generated by first generating I from $\{1, 2, \ldots, M\}$ according to the distribution specified by $\boldsymbol{\alpha}$, and then conditional on I, generating Y according to the density f_I. Then given an iid sample from $f_{\boldsymbol{\alpha}}$, say $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, this motivates

associating with each Y_j a (potentially completely fictitious) auxiliary variable I_j indicating which f_i gave rise to Y_j. Bayes analyses of mixture samples typically make use of such variables. And this perspective begins to perhaps motivate a well known difficulty often encountered Bayes analyses of the one-sample mixture problem. That is that unless one constrains $\boldsymbol{\alpha}$ (either by use of a very strong prior essentially outlawing the possibility that any $\alpha_i = 0$, or by simply adopting a parameter space that is bounded away from cases where any $\alpha_i = 0$) to prevent "extreme" mixture parameters and cases where not all of the elements of $\{1, 2, \ldots, M\}$ are represented in $\{I_1, I_2, \ldots, I_n\}$, a posterior sampling algorithm can behave badly. Degenerate submodels of the full mixture model (59) that have one or more $\alpha_i = 0$ can act (at least for practical purposes) as "absorbing states" for MCMC algorithms. In the language of Definition 15 on page 79, the chains in effect fail to be "irreducible."

6.11 "Bayes" Analysis for Inference About a Function $g(t)$

An interesting application of random process theory and what are really Bayes ideas is to the estimation/interpolation of values of a function $g(t)$ for $t \in (a, b)$ (possibly observed with error) from values at some points in (a, b). (A $\boldsymbol{t} \in \Re^k$ version of what follows can be created using the ideas in the present "function of a single real variable" version, but the simplest case will suffice here.) This kind of material is very important in modern "analysis of computer experiments" applications, where in order to evaluate $g(t)$ a long and therefore expensive computer run is required. It is then desirable to get a few values of g and use them to derive some cheap/computationally simple interpolator/approximator/surrogate for g at other values of t.

So suppose that for

$$t_1 < t_2 < \cdots < t_k$$

one calculates or observes

$$g(t_1), g(t_2), \ldots, g(t_k) \tag{60}$$

or perhaps

$$g(t_1) + \epsilon_1, g(t_2) + \epsilon_2, \ldots, g(t_k) + \epsilon_k \tag{61}$$

where one might model the ϵ_i as iid $\mathrm{N}\left(0, \sigma^2\right)$ random variables, and that one wishes to estimate/predict $g(t^*)$ for some $t^* \in (a, b)$. A way to use Bayes machinery and the theory of stationary Gaussian random processes here is to model and calculate as follows.

I might invent a "prior" for the function $g(\cdot)$ by

1. specifying my best prior guess at the function $g(\cdot)$ as $\mu(\cdot)$,
2. writing
$$g(t) = \mu(t) + (g(t) - \mu(t)) = \mu(t) + \gamma(t)$$
and

3. modeling $\gamma(t)$ as a mean 0 stationary Gaussian process.

Item 3. here means that we assume that $\mathrm{E}\gamma(t) = 0\ \forall t$, $\mathrm{Var}\gamma(t) = \tau^2\ \forall t$, for some positive definite function $\rho(\Delta)$ taking values in $(0,1)$ (a mathematically valid correlation function)

$$\mathrm{Cov}(\gamma(t), \gamma(t')) = \tau^2 \rho(|t - t'|)$$

and that for any finite number of values t, the joint distribution of the corresponding $\gamma(t)$'s is multivariate normal. Standard choices of the function $\rho(\Delta)$ are $\exp(-\beta\Delta)$ and $\exp\left(-\beta\Delta^2\right)$. (The first tends to produce "rougher" realizations than does the second. In both cases, the positive parameter β governs how fast correlation dies off as a function of distance between to values t and t'.) In this model for $\gamma(t)$, τ^2 in some sense quantifies overall uncertainty about $g(t)$, the form of $\rho(\Delta)$ can be made to reflect what one expects in terms of smoothness/roughness of deviations of $g(t)$ from $\mu(t)$, and for a typical choice of $\rho(\Delta)$, a parameter β governs how fast (as one moves away from one of $t_1, t_2 \ldots, t_k$) a prediction of $g(t)$ should move toward $\mu(t)$.

Then, say in the case (60) where there are no errors of observation, with t^* different from any of $t_1, t_2 \ldots, t_k$, the model here (a prior for $g(t)$) implies that

$$\begin{pmatrix} g(t_1) \\ \vdots \\ g(t_k) \\ g(t^*) \end{pmatrix} \sim \mathrm{MVN}_{k+1}\left(\begin{pmatrix} \mu(t_1) \\ \vdots \\ \mu(t_k) \\ \mu(t^*) \end{pmatrix}, \boldsymbol{\Sigma}\right) \tag{62}$$

for

$$\underset{(k+1)\times(k+1)}{\boldsymbol{\Sigma}} = \tau^2 \left(\rho(|t_i - t_j|)\right)_{\substack{i=1,\ldots,k+1 \\ j=1,\ldots,k+1}} \tag{63}$$

with the understanding that we are letting $t_{k+1} = t^*$. Then multivariate normal theory gives fairly simple formulas for the conditional distribution of part of the multivariate normal vector given the value of the rest of the vector. That is, it is straightforward to find from (62) the normal conditional (posterior) distribution of

$$g(t^*) \,|\, (g(t_1), g(t_2), \ldots, g(t_k))$$

This, in turn, produces plausibility statements about the unevaluated $g(t^*)$.

The case (61) is much the same. The only differences are that the covariance matrix for $(g(t_1) + \epsilon_1, g(t_2) + \epsilon_2, \ldots, g(t_k) + \epsilon_k, g(t^*))$ is not $\boldsymbol{\Sigma}$ specified in (63), but rather

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma} + \mathbf{diag}\left(\sigma^2, \sigma^2, \ldots, \sigma^2, 0\right)$$

and that one is concerned with the conditional distribution of

$$g(t^*) \,|\, (g(t_1) + \epsilon_1, g(t_2) + \epsilon_2, \ldots, g(t_k) + \epsilon_k)$$

for purposes of prediction/interpolation at t^*.

7 Bayesian Nonparametrics

This section outlines an introduction to Bayesian analysis of some "nonparametric" and "semi-parametric" models. The standard textbook reference for such material is *Bayesian Nonparametrics* by Ghosh and Ramamoorthi and the topics here are discussed in Chapter 3 of that book. The basic concern is distributions on (and thus "priors" on) distributions F (a cdf, or P the corresponding "probability measure" that assigns probabilities to sets of outcomes, B) where F (or P) is not confined to any relatively simple parametric family (like the Gaussian, the Weibull, the beta, etc.)

7.1 Dirichlet and Finite "Stick-Breaking" Processes

Suppose that $\alpha(\cdot)$ is a multiple of a probability distribution on $\Re$ (meaning that it assigns "mass" to subsets of $\Re$ like a probability distribution does, but is not necessarily normalized to have total mass 1, i.e. potentially $\alpha(\Re) \neq 1$). (Much of what follows could be done in $\Re^d$, but for simplicity of exposition, we will here work in 1 dimension.) It turns out to be "mathematically OK" to invent a probability distribution for P (or F) (itself a probability distribution on $\Re$) by specifying that for any partition of $\Re$ into a finite number of disjoint sets $B_1, B_2, \ldots, B_k$ with $\bigcup_{i=1}^{k} B = \Re$, the vector of probabilities that P assigns to these sets is Dirichlet distributed with parameters specified by α, that is

$$(P(B_1), P(B_2), \ldots, P(B_k)) \sim \text{Dirichlet}_k(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k)) \quad (64)$$

When (64) holds for all such partitions of $\Re$, we'll say that P is a Dirichlet process on $\Re$ with parameter measure α, and write

$$P \sim \mathcal{D}_\alpha$$

Now the defining property (64) doesn't give one much feeling about what realizations from $\mathcal{D}_\alpha$ "look like," but it does turn out to be very tractable and enable the proof of all sorts of interesting and useful facts about Dirichlet processes, and particularly about models where

$$P \sim \mathcal{D}_\alpha \quad (65)$$

and conditioned on P,

$$Y_1, Y_2, \ldots, Y_n \sim \text{iid } P \text{ (or } F) \quad (66)$$

(this is the one-sample model where P or F has the "prior" $\mathcal{D}_\alpha$).

Some Dirichlet process facts are:

1. If P (or F)$\sim \mathcal{D}_\alpha$ there is the "neutral to the right property" that says for $t_1 < t_2 < \cdots < t_k$, the random variables

$$(1 - F(t_1)), \frac{1 - F(t_2)}{1 - F(t_1)}, \frac{1 - F(t_3)}{1 - F(t_2)}, \ldots, \frac{1 - F(t_k)}{1 - F(t_{k-1})}$$

are independent.

2. Under the assumptions (65) and (66), the posterior for P is also a Dirichlet process, that is

$$P|\left(Y_1, Y_2, \ldots, Y_n\right) \sim \mathcal{D}_{\left(\alpha + \sum_{i=1}^{n} \delta_{Y_i}\right)}$$

for δ_{Y_i} a unit point mass distribution located at Y_i. That is, the posterior for P is derived from the prior for P by updating the "parameter" measure by the addition of unit point masses at each observation.

3. It follows directly from 2. and (64) under the assumptions (65) and (66), that conditioned on $Y_1, Y_2, \ldots, Y_n$ the variable $F(t) = P((-\infty, t])$ is Beta and has mean

$$\begin{aligned} \mathrm{E}\left(F(t)\,|Y_1, Y_2, \ldots, Y_n\right) &= \frac{\alpha((-\infty, t]) + \sum_{i=1}^{n} \delta_{Y_i}((-\infty, t])}{\alpha(\Re) + n} \\ &= \frac{\alpha(\Re)}{\alpha(\Re) + n} \cdot \frac{\alpha((-\infty, t])}{\alpha(\Re)} + \frac{n}{\alpha(\Re) + n} \cdot \frac{\#\left[Y_i \leq t\right]}{n} \end{aligned}$$

That is, this conditional mean is a weighted average of the probability that a normalized version of α assigns to $(-\infty, t]$ and the relative frequency with which the observations Y_i are in $(-\infty, t]$, where the weights are $\alpha(\Re)$ (the prior mass) and n (the sample size).

4. It similarly follows from 2. (and the fact that if $P \sim \mathcal{D}_\alpha$ and $Y|P \sim P$ then $Y \sim \alpha/\alpha(\Re)$) that under the assumptions (65) and (66), (posterior) predictive distributions are tractable. That is

$$Y_{n+1}|\left(Y_1, Y_2, \ldots, Y_n\right) \sim \frac{\alpha + \sum_{i=1}^{n} \delta_{Y_i}}{\alpha(\Re) + n}$$

(Y_{n+1} has a predictive distribution that is a normalized version of the parameter measure of the posterior.)

5. Despite the fact that 4. has a "sequential" nature, under assumptions (65) and (66), the marginal of $(Y_1, Y_2, \ldots, Y_n)$ is "exchangeable"/symmetric. Every Y_i has the same marginal, every pair $(Y_i, Y_{i'})$ has the same bivariate distribution, etc.

6. Probably the initially least appealing elementary fact about Dirichlet processes is that with probability 1 their realizations are discrete. That is

$$\mathcal{D}_\alpha\left(\{\text{discrete distributions on } \Re\} = 1\right)$$

(P generated according to a Dirichlet "prior" is sure to be concentrated on a countable set of values.)

More insight into fact 6. above and ultimately motivation for other related nonparametric priors for probability distributions is provided by an important

representation theorem of Sethuraman. $\mathcal{D}_\alpha$ has a representation as a "stick-breaking prior" as follows. Suppose that

$$X_1, X_2, X_3, \ldots \text{ are iid according to } \frac{1}{\alpha(\Re)}\alpha$$

independent of

$$\theta_1, \theta_2, \theta_3, \ldots \text{ that are iid Beta}(1, \alpha(\Re))$$

Set

$$p_1 = \theta_1 \text{ and } p_m = \theta_m \prod_{i=1}^{m-1} (1-\theta_i) \ \forall m > 1$$

(these probabilities p_m are created in "stick-breaking" fashion). Then the (random) probability distribution

$$P \equiv \sum_{m=1}^{\infty} p_m \delta_{X_m} \sim \mathcal{D}_\alpha \tag{67}$$

This representation says that to simulate a realization from $\mathcal{D}_\alpha$, one places probability mass θ_1 at X_1, then places a fraction θ_2 of the remaining probability mass at X_2, then places a fraction θ_3 of the remaining probability mass at X_3, etc.

Representation (67) (involving as it does an infinite sum) is nothing that can be used in practical computations/data analysis. But it motivates the consideration of something that *can* be used in practice, namely a truncated version of P that has not a countable number of discrete components, but only a finite number, N, instead. That is, suppose that

$$X_1, X_2, X_3, \ldots, X_N \text{ are iid according to } \frac{1}{\alpha(\Re)}\alpha$$

independent of

$$\theta_1, \theta_2, \theta_3, \ldots, \theta_{N-1} \text{ that are iid Beta}(1, \alpha(\Re))$$

Set

$$p_m = \theta_m \prod_{i=1}^{m-1} (1-\theta_i) \ \forall 1 \le m < N \text{ and } p_N = \prod_{i=1}^{N-1} (1-\theta_i)$$

and define

$$P_N = \sum_{m=1}^{N} p_m \delta_{X_m}$$

Presumably, for "large" N, in some appropriate sense $P_N \dot{\sim} \mathcal{D}_\alpha$.

A natural generalization of this "truncated Dirichlet process" idea can be formulated as follows. Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{N-1})$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{N-1})$ be vectors of positive constants. For

$$X_1, X_2, X_3, \ldots, X_N \text{ iid according to a probability distribution } H \tag{68}$$

independent of

$$\theta_i \text{ for } i = 1, \ldots, N-1 \text{ independent Beta}(\gamma_i, \alpha_i) \text{ variables} \tag{69}$$

set

$$p_m = \theta_m \prod_{i=1}^{m-1} (1-\theta_i) \ \forall 1 \leq m < N \quad \text{and} \quad p_N = \prod_{i=1}^{N-1} (1-\theta_i)$$

and define

$$P_N = \sum_{m=1}^{N} p_m \delta_{X_m} \tag{70}$$

One might say that

$$P_N \sim \mathcal{SB}(N, H, \boldsymbol{\gamma}, \boldsymbol{\alpha})$$

i.e. that P_N is a general N component stick-breaking process.

The beauty of representation (70) is that it involves only the N ordinary random variables (68) and the $N-1$ ordinary random variables (69). So it can be used to specify nonparametric components of practically implementable Bayes models and thus be used in data analysis.

7.2 Polya Tree Processes

A second nonparametric way of specifying a distribution on distributions is through the use of so-called "Polya trees." We begin the exposition of Polya tree processes with a small relatively concrete example.

Suppose that one has in mind 8 real numbers $x_1 < x_2 < \cdots < x_8$ and is interested in distributions over distributions on these values. (Actually, it is not at all essential that these x's are real numbers rather than just arbitrary "points," but with an eye to the ultimate application we might as well think of them as ordered real numbers.) For convenience, we will rename the values with binary labels and think of them at the bottom of a binary tree as in Figure 12.

The p's marked on Figure 12 are meant to add to 1 in the obvious pairs ($p_0 + p_1 = 1, p_{00} + p_{01} = 1, p_{10} + p_{11} = 1$, etc.). For fixed values of these, the tree structure in the figure can be used to define a probability distribution over the 8 elements at the bottom of the tree according to the prescription "multiply p's on the branches you take to go from the top to a given final node." That is, with $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3) \in \{0,1\}^3$ the p's define a probability distribution on $\boldsymbol{\epsilon}$'s by

$$P(\{\boldsymbol{\epsilon}\}) = p_{\epsilon_1} p_{\epsilon_1 \epsilon_2} p_{\epsilon_1 \epsilon_2 \epsilon_3} \tag{71}$$

Then, if one places an appropriate probability distribution on the set of p's, one has placed a distribution on the distribution P. In fact, since the p's with labels ending in 0 are 1 minus the corresponding p's where the label is changed only by switching the last digit, one needs only to place a joint distribution on the p's with labels ending in 1.

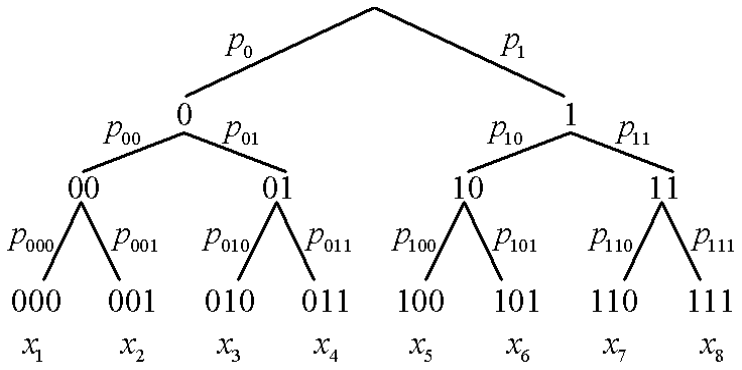

Figure 12: A 3-Level Binary Tree

A so-called "Polya tree" method of endowing the p's with a distribution is to assume that those with labels ending in 1 are independent Beta variables. That is, for $\boldsymbol{\epsilon}$ some string of $0, 1$, or 2 zeros and ones (so that $\boldsymbol{\epsilon}1$ and $\boldsymbol{\epsilon}0$ are strings of zeros and ones of length $1, 2$, or 3) suppose that the

$$p_{\boldsymbol{\epsilon}1} \sim \text{ind Beta}\left(\alpha_{\boldsymbol{\epsilon}1}, \alpha_{\boldsymbol{\epsilon}0}\right) \tag{72}$$

(for parameters $\alpha_{\boldsymbol{\epsilon}1}$ and $\alpha_{\boldsymbol{\epsilon}0}$). Letting $\boldsymbol{\alpha}$ stand for the whole collection of α's (two for each $p_{\boldsymbol{\epsilon}1}$) we will say that the distribution over distributions P of the form (71) produced by this assumption is a $\mathcal{PT}_3\left(\boldsymbol{\alpha}\right)$ process. (P is a simple "finite Polya tree" process.)

The form (71) and assumptions (72) immediately produce the result that for $P \sim \mathcal{PT}_3\left(\boldsymbol{\alpha}\right)$,

$$\text{E}P\left(\{\boldsymbol{\epsilon}\}\right) = \left(\frac{\alpha_{\epsilon_1}}{\alpha_1 + \alpha_0}\right)\left(\frac{\alpha_{\epsilon_1\epsilon_2}}{\alpha_{\epsilon_1 1} + \alpha_{\epsilon_1 0}}\right)\left(\frac{\alpha_{\epsilon_1\epsilon_2\epsilon_3}}{\alpha_{\epsilon_1\epsilon_2 1} + \alpha_{\epsilon_1\epsilon_2 0}}\right) \tag{73}$$

If conditioned on P variable Y is P distributed and $P \sim \mathcal{PT}_3\left(\boldsymbol{\alpha}\right)$, it is immediate that the marginal probabilities for Y are also given by (73).

Next observe that if $P \sim \mathcal{PT}_3\left(\boldsymbol{\alpha}\right)$ and conditional on P, $Y \sim P$, for $\boldsymbol{p}$ the set of $p_{\boldsymbol{\epsilon}1}$'s, a joint density for $\boldsymbol{p}$ and Y is proportional to the product

$$\left(\prod_{\boldsymbol{\epsilon}}\left(p_{\boldsymbol{\epsilon}1}^{\alpha_{\boldsymbol{\epsilon}1}-1}p_{\boldsymbol{\epsilon}0}^{\alpha_{\boldsymbol{\epsilon}0}-1}\right)\right)\left(p_{\epsilon_1}p_{\epsilon_1\epsilon_2}p_{\epsilon_1\epsilon_2\epsilon_3}\right) \tag{74}$$

(In (74) the product in the first term is over $\boldsymbol{\epsilon}$'s of length 0 through 2.) The first term of (74) is proportional to the joint density of p's and the second is the pmf for Y. It is then obvious that the posterior distribution of $P|Y$ is again a Polya tree process. That is because with $\boldsymbol{\epsilon}$ a vector of $1, 2$, or 3 zeros and ones and

$$\Delta_{\boldsymbol{\epsilon}}\left(Y\right) \equiv \begin{cases} 1 & \text{if the first part of } Y \text{ is } \boldsymbol{\epsilon} \\ 0 & \text{otherwise} \end{cases},$$

conditioned on $Y = (\epsilon_1, \epsilon_2, \epsilon_3)$

$$p_{\boldsymbol{\epsilon}1} \sim \text{ind Beta}\left(\alpha_{\boldsymbol{\epsilon}1} + \Delta_{\boldsymbol{\epsilon}1}(Y), \alpha_{\boldsymbol{\epsilon}0} + \alpha_{\boldsymbol{\epsilon}1} + \Delta_{\boldsymbol{\epsilon}0}(Y)\right)$$

That is, in order to update a $\mathcal{PT}_3(\boldsymbol{\alpha})$ "prior" for P to a posterior, one simply looks through the tree adding 1 to each α traversed to produce Y.

The conjugacy of the Polya tree process and the form (71) for a single observation obviously generalizes to $Y_1, Y_2, \ldots, Y_n$ iid according to P defined in (71) and further allows for easy identification of posterior predictive distributions. That is, adopt the notation

$$\boldsymbol{\alpha} \oplus \boldsymbol{\Delta}(Y_1, Y_2, \ldots, Y_n)$$

for a set of α's updated by adding to each $\alpha_{\boldsymbol{\epsilon}}$ a count of the number of Y_i's that involve use of the corresponding branch of the tree. If $P \sim \mathcal{PT}_3(\boldsymbol{\alpha})$ and conditioned on P the $Y_1, Y_2, \ldots, Y_n$ are iid according to P, the posterior of P is $\mathcal{PT}_3(\boldsymbol{\alpha} \oplus \boldsymbol{\Delta}(Y_1, Y_2, \ldots, Y_n))$. Further, if $P \sim \mathcal{PT}_3(\boldsymbol{\alpha})$ and conditioned on P the $Y_1, Y_2, \ldots, Y_n, Y_{\text{new}}$ are iid P, the posterior predictive distribution of $Y_{\text{new}} | (Y_1, Y_2, \ldots, Y_n)$ is specified by

$$\begin{aligned}\Pr\left[Y_{\text{new}} = (\epsilon_1, \epsilon_2, \epsilon_3) \,|\, (Y_1, Y_2, \ldots, Y_n)\right] &= \left(\frac{\alpha_{\epsilon_1} + \sum_{i=1}^n \Delta_{\epsilon_1}(Y_i)}{\alpha_1 + \alpha_0 + n}\right) \\ &\times \left(\frac{\alpha_{\epsilon_1\epsilon_2} + \sum_{i=1}^n \Delta_{\epsilon_1\epsilon_2}(Y_i)}{\alpha_{\epsilon_1 1} + \alpha_{\epsilon_1 0} + \sum_{i=1}^n \Delta_{\epsilon_1}(Y_i)}\right) \\ &\times \left(\frac{\alpha_{\epsilon_1\epsilon_2\epsilon_3} + \sum_{i=1}^n \Delta_{\epsilon_1\epsilon_2\epsilon_3}(Y_i)}{\alpha_{\epsilon_1\epsilon_2 1} + \alpha_{\epsilon_1\epsilon_2 0} + \sum_{i=1}^n \Delta_{\epsilon_1\epsilon_2}(Y_i)}\right)\end{aligned}$$

which is in some sense the generalization of the statement that (73) gives marginal probabilities for Y if given P, variable Y has distribution P and $P \sim \mathcal{PT}_3(\boldsymbol{\alpha})$.

An important question is how one might sensibly choose the parameters $\boldsymbol{\alpha}$ for a $\mathcal{PT}_3(\boldsymbol{\alpha})$ process (or differently put, what are the consequences of various choices). To begin, formula (73) shows how the mean of P distribution values depends upon the choice of $\boldsymbol{\alpha}$ If one has in mind some "best guess" distribution $H(\cdot)$ it is simple to choose $\boldsymbol{\alpha}$ to produce $\text{E}P(\{\boldsymbol{\epsilon}\}) = H(\{\boldsymbol{\epsilon}\})\ \forall \boldsymbol{\epsilon} \in \{0,1\}^3$. This is accomplished by choosing elements in $\boldsymbol{\alpha}$ so that

$$\begin{aligned}\frac{\alpha_1}{\alpha_0} &= \frac{H(\{100, 101, 110, 111\})}{H(\{000, 001, 010, 011\})}, \frac{\alpha_{01}}{\alpha_{00}} = \frac{H(\{010, 011\})}{H(\{000, 001\})}, \frac{\alpha_{11}}{\alpha_{10}} = \frac{H(\{110, 111\})}{H(\{100, 101\})}, \\ \frac{\alpha_{001}}{\alpha_{000}} &= \frac{H(\{001\})}{H(\{000\})}, \frac{\alpha_{011}}{\alpha_{010}} = \frac{H(\{011\})}{H(\{010\})}, \frac{\alpha_{101}}{\alpha_{100}} = \frac{H(\{101\})}{H(\{100\})}, \text{ and } \frac{\alpha_{111}}{\alpha_{110}} = \frac{H(\{111\})}{H(\{110\})}.\end{aligned}$$

Any pairs of α's with the correct ratios will do to produce a Polya tree process with mean $H(\cdot)$, and subject to these ratio relationships, one is still free to choose, say, the sums

$$\begin{aligned}&\alpha_0 + \alpha_1, \alpha_{00} + \alpha_{01}, \alpha_{10} + \alpha_{11}, \alpha_{000} + \alpha_{001}, \\ &\alpha_{010} + \alpha_{011}, \alpha_{100} + \alpha_{101}, \text{ and } \alpha_{110} + \alpha_{111}\end{aligned} \tag{75}$$

to govern how variable realizations P from $\mathcal{PT}_3(\boldsymbol{\alpha})$ are and "at what scale(s)" they vary the most.

To understand this last point, note that if $U \sim$Beta(γ, δ), if I fix γ/δ I have fixed

$$\frac{\gamma/\delta}{1+\gamma/\delta} = \frac{\gamma}{\gamma+\delta} = \mathrm{E}U \quad ,$$

but that the larger is $\gamma+\delta$, the smaller is Var$U$. So in the $\mathcal{PT}_3(\boldsymbol{\alpha})$ context, the larger are the sums $\alpha_{\boldsymbol{\epsilon}1}+\alpha_{\boldsymbol{\epsilon}0}$, the less variable are the realizations P. (And in the case where conditioned on P variables $Y_1, Y_2, \ldots, Y_n$ are iid P, the less strongly the posterior is pulled from the prior mean H toward the empirical distribution of the Y_i's.) Control of the scales at which P varies (or the posterior is pulled toward the empirical distribution of the Y_i's) can be made via control of the relative sizes of these sums at different levels in the tree. For example, for a given mean H for P,

1. the sum $\alpha_0 + \alpha_1$ being large in comparison to the other sums in (75) allows the "left and right half total P probability" to vary little from the corresponding "left and right half total H probability" but allows the details within those "halves" to vary relatively substantially (so posterior left and right half totals stay near H totals, but the specifics of posterior probabilities within the halves can become approximately proportional to the empirical frequency pattern of Y_i's within the halves), and

2. the sum $\alpha_0+\alpha_1$ being small in comparison to the other sums in (75) allows the "left and right half total P probability" to fluctuate substantially, but forces the patterns within those halves to be like those of the mean H (so posterior left and right half totals are pulled toward the sample fractions, but the specifics of posterior probabilities within the halves are pulled towards being in proportion to those of the prior mean distribution, H).

In keeping with the qualitative notion that a sample $Y_1, Y_2, \ldots, Y_n$ should provide less and less information about the finer and finer detail of $P \sim \mathcal{PT}_3(\boldsymbol{\alpha})$ as one goes further down the tree (and one must thus lean more and more heavily on prior assumptions) it is more or less conventional to make $\alpha_{\boldsymbol{\epsilon}0} + \alpha_{\boldsymbol{\epsilon}1}$ increase in the length of $\boldsymbol{\epsilon}$ (the depth one is at in the tree).

The issue of how to take the basic idea illustrated with the foregoing small 8 point example and make a general tool for Bayes data analysis for distributions on $\Re$ from it can be attacked in at least 2 ways. The simplest is to spread a power of 2 (say 2^k) values x_i across a part of $\Re$ thought to contain essentially all the probability of an unknown probability distribution and use a finite $\mathcal{PT}_k(\boldsymbol{\alpha})$ process on those points as a nonparametric prior over distributions (on the 2^k points) that might approximate the unknown one. A second more interesting one is to use (approximate truncated versions) of a general Polya tree process on $\Re$. In the balance of this discussion we consider this second possibility.

Definition of a Polya tree process on $\Re$ depends upon the choice of a nested set of partitions of $\Re$. That is, let

1. B_0 and B_1 be disjoint sets with $B_0 \cup B_1 = \Re$,

2. B_{00} and B_{01} be disjoint sets with $B_{00} \cup B_{01} = B_0$, and B_{10} and B_{11} be disjoint sets with $B_{10} \cup B_{11} = B_1$, and

3. $\forall k \geq 2$ and $\boldsymbol{\epsilon} \in \{0,1\}^k$ $B_{\boldsymbol{\epsilon}0}$ and $B_{\boldsymbol{\epsilon}1}$ be disjoint sets with $B_{\boldsymbol{\epsilon}0}$ and $B_{\boldsymbol{\epsilon}1} = B_{\boldsymbol{\epsilon}}$.

Then consider an infinite set of independent random variables

$$p_{\boldsymbol{\epsilon}1} \sim \text{Beta}\left(\alpha_{\boldsymbol{\epsilon}1}, \alpha_{\boldsymbol{\epsilon}0}\right)$$

for each $\boldsymbol{\epsilon}$ a (possibly degenerate) finite vector of zeros and ones (for positive constants $\alpha_{\boldsymbol{\epsilon}1}$ and $\alpha_{\boldsymbol{\epsilon}0}$). Define a (random) set of conditional probabilities

$$P\left(B_{\boldsymbol{\epsilon}1}|B_{\boldsymbol{\epsilon}}\right) = p_{\boldsymbol{\epsilon}1}$$

These specifications in turn imply (random) P probabilities for all of the sets $B_{\boldsymbol{\epsilon}}$. For example, much as in (71) for any $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_k) \in \{0,1\}^k$

$$P\left(B_{\boldsymbol{\epsilon}}\right) = p_{\epsilon_1} p_{\epsilon_1\epsilon_2} \cdots p_{\epsilon_1\epsilon_2\cdots\epsilon_k}$$

Theorem 3.3.2 of Ghosh and Ramamoorthi then says that under quite mild conditions on the set of α's this random set of probabilities on the sets $B_{\boldsymbol{\epsilon}}$ extends to a random distribution P giving probabilities not just to sets $B_{\boldsymbol{\epsilon}}$ but to all (measurable) subsets of $\Re$, and we can term that a realization from a $\mathcal{PT}\left(\boldsymbol{\alpha}\right)$ process.

Of course (involving as it does limits and "infinities") this definition of a $\mathcal{PT}\left(\boldsymbol{\alpha}\right)$ process on $\Re$ can not be used exactly in real data analysis. An approximation to it that can be used in practice is this. For H a desired mean distribution on $\Re$, suppose that for $\boldsymbol{\epsilon}$'s vectors of zeros and ones of lengths no more than k, we have defined $B_{\boldsymbol{\epsilon}}$'s, $p_{\boldsymbol{\epsilon}}$'s and $P\left(B_{\boldsymbol{\epsilon}}\right)$'s as above with α's chosen to make

$$\text{E}P\left(B_{\boldsymbol{\epsilon}}\right) = H\left(B_{\boldsymbol{\epsilon}}\right)$$

Then for any $\boldsymbol{\epsilon}$ of length k and $A \subset B_{\boldsymbol{\epsilon}}$, define

$$P\left(A|B_{\boldsymbol{\epsilon}}\right) = \frac{H\left(A\right)}{H\left(B_{\boldsymbol{\epsilon}}\right)}$$

This works to define a random process for assigning probabilities to all subsets of $\Re$. In particular, in cases where H has pdf h, this random distribution on $\Re$ has (random, because the $P\left(B_{\boldsymbol{\epsilon}}\right)$ are random) pdf defined piecewise by

$$f_k\left(y\right) = P\left(B_{\boldsymbol{\epsilon}}\right)\frac{h(y)}{H\left(B_{\boldsymbol{\epsilon}}\right)} \text{ for } y \in B_{\boldsymbol{\epsilon}} \text{ (for all } \boldsymbol{\epsilon} \in \{0,1\}^k\text{)}$$

(This is $P\left(B_{\boldsymbol{\epsilon}}\right)$ times the H conditional density over $B_{\boldsymbol{\epsilon}}$.)

If this model is used to do data analysis based on $Y_1, Y_2, \ldots, Y_n$ iid according to P, to get a posterior for P, the original α's are simply updated as in the first

$\mathcal{PT}_3(\boldsymbol{\alpha})$ example above according to counts of Y_i's falling into the $B_{\boldsymbol{\epsilon}}$ (for all $\boldsymbol{\epsilon} \in \{0,1\}^k$). The posterior is then of the "approximate $\mathcal{PT}(\boldsymbol{\alpha})$" type just described. The "posterior mean pdf" (and posterior predictive density for Y_{new}) under this structure is defined piecewise by

$$\mathrm{E}\left[P\left(B_{\boldsymbol{\epsilon}}\right)|Y_1, Y_2, \ldots, Y_n\right] \frac{h(y)}{H\left(B_{\boldsymbol{\epsilon}}\right)} \text{ for } y \in B_{\boldsymbol{\epsilon}}$$

where $\mathrm{E}[P\left(B_{\boldsymbol{\epsilon}}\right)|Y_1, Y_2, \ldots, Y_n]$ is of a form similar to (73) but based on the updated α's.

A particularly attractive choice of the partitions leading to a $\mathcal{PT}(\boldsymbol{\alpha})$ process on $\Re$ (and thus to a truncation of it) in the case that H has density h that is positive exactly at every point of the (potentially infinite) interval (a, b) is

1. $B_0 = \left(a, H^{-1}\left(\frac{1}{2}\right)\right]$ and $B_1 = \left(H^{-1}\left(\frac{1}{2}\right), b\right)$,
2. $B_{00} = \left(a, H^{-1}\left(\frac{1}{4}\right)\right], B_{01} = \left(H^{-1}\left(\frac{1}{4}\right), H^{-1}\left(\frac{1}{2}\right)\right], B_{10} = \left(H^{-1}\left(\frac{1}{2}\right), H^{-1}\left(\frac{3}{4}\right)\right]$, and $\left(H^{-1}\left(\frac{3}{4}\right), b\right)$,

and so on. This not only provides a natural set of partitions, but suggests the choices $\alpha_{\boldsymbol{\epsilon}1} = \alpha_{\boldsymbol{\epsilon}0}$ and thus all $\mathrm{E}p_{\boldsymbol{\epsilon}1} = 1/2$ under the Polya scheme (making $\mathrm{E}P\left(B_{\boldsymbol{\epsilon}}\right) = H\left(B_{\boldsymbol{\epsilon}}\right)$ the power of 1/2 corresponding to the length of $\boldsymbol{\epsilon}$). In keeping with a desire for (relatively) smooth (though admittedly only piecewise continuous) posterior mean densities and the kind of considerations discussed for choice of the α's for $\mathcal{PT}_3(\boldsymbol{\alpha})$, it is conventional to make $\alpha_{\boldsymbol{\epsilon}0} + \alpha_{\boldsymbol{\epsilon}1}$ increase in the length of $\boldsymbol{\epsilon}$ (increase with depth in the tree) and in particular, growth of the sum at a "squared length of $\boldsymbol{\epsilon}$ rate" is often recommended.

8 Some Scraps (WinBUGS and Other)

8.1 The "Zeroes Trick"

WinBUGS is a very flexible/general program. But it obviously can not automatically handle *every* distribution one could invent and want to use as part of one of its models. That is, there is sometimes a need to include some non-standard factor $h_1(\boldsymbol{\eta})$ in an expression

$$h(\boldsymbol{\eta}) = h_1(\boldsymbol{\eta})\, h_2(\boldsymbol{\eta})$$

from which one wishes to sample. (For example, one might invent a nonstandard prior distribution specified by $g(\boldsymbol{\theta})$ that needs to be combined with a likelihood $L(\boldsymbol{\theta})$ in order to create a product $L(\boldsymbol{\theta})\, g(\boldsymbol{\theta})$ proportional to a posterior density from which samples need to be drawn.) Obviously, in such cases one will somehow need to code a formula for $h_1(\boldsymbol{\eta})$ and get it used as a factor in a formula for $h(\boldsymbol{\eta})$. The WinBUGS method of doing this is to employ "the zeroes trick" based on the use of a fictitious Poisson observation with fictitious observed value 0.

That is, a Poisson variable X with mean λ has probability of taking the value 0

$$P_\lambda [X = 0] = \exp(-\lambda)$$

So if in addition to all else one does in a `WinBUGS` model statement, one specifies that a variable Y is Poisson with mean

$$c - \ln(h_1(\boldsymbol{\eta}))$$

and gives the program "data" that says $Y = 0$, the overall effect is to include a multiplicative factor of

$$\exp(-c + \ln(h_1(\boldsymbol{\eta}))) = \exp(-c)\, h_1(\boldsymbol{\eta})$$

in the $h(\boldsymbol{\eta})$ from which `WinBUGS` samples. Notice that since `WinBUGS` expects a positive mean for a Poisson variable, one will typically have to use a non-zero value of c when employing the "trick" with

$$c > \max_{\boldsymbol{\eta}} \ln(h_1(\boldsymbol{\eta}))$$

in order to prevent `WinBUGS` from balking at some point in its Gibbs iterations.

8.2 Convenient Parametric Forms for Sums and Products

In building probability models (including "Bayes" models that treat parameters as random variables) it is often convenient to be able to think of a variable as a sum or product of independent pieces that "combine nicely," i.e. to be able to model Y as either

$$X_1 + X_2 + \cdots + X_k \tag{76}$$

or as

$$X_1 \cdot X_2 \cdot \cdots \cdot X_k \tag{77}$$

for independent X_i with suitable marginal distributions. It is thus useful to review and extend a bit of Stat 542 probability that is relevant in accomplishing this.

For the case of (76) recall that

1. $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ independent of $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ implies that $Y = X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ producing a fact useful when one is modeling a continuous variable taking values in all of $\Re$,

2. $X_1 \sim$Poisson(λ_1) independent of $X_2 \sim$Poisson(λ_2) implies that $Y = X_1 + X_2 \sim$Poisson$(\lambda_1 + \lambda_2)$ producing a fact useful when one is modeling a discrete variable taking values in $\{0, 1, 2, \ldots\}$, and

3. $X_1 \sim \Gamma(\alpha_1, \beta)$ independent of $X_2 \sim \Gamma(\alpha_2, \beta)$ implies that $Y = X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \beta)$ producing a fact useful when one is modeling a continuous variable taking values in $(0, \infty)$.

And, of course, the facts 1) that independent Binomial variables the same success probability add to give another Binomial variable with that success probability and 2) that independent negative Binomials (or geometrics) with a common success probability add to give a negative Binomial variable with that success probability are sometimes helpful.

For purposes of convenient modeling of products (77), one can make use of facts about sums and the exponential function (i.e. the fact that exponentiation turns sums into products). That is, if

$$X_i = \exp(X_i')$$

then

$$Y = X_1 \cdot X_2 \cdot \cdots \cdot X_k = \exp(X_1' + X_2' + \cdots + X_k')$$

so facts about convenient forms for sums have corresponding facts about convenient product forms. Using 1. and 3. above, one has for example

1. $X' \sim \mathrm{N}(\mu, \sigma^2)$ implies that $X = \exp(X')$ has what is typically called a "lognormal" distribution on $(0, \infty)$ and the product of independent lognormal variables is again lognormal,
2. $X' \sim \Gamma(\alpha, \beta)$ implies that $X = \exp(X')$ has what is might be called a "log-gamma" distribution on $(1, \infty)$ and the product of independent log-gamma variables with a common β is again log-gamma, and
3. $X' \sim \Gamma(\alpha, \beta)$ implies that $X = \exp(-X')$ has what is might be called a "negative log-gamma" distribution on $(0, 1)$ and the product of independent negative log-gamma variables with a common β is again negative log-gamma.

This last fact can be useful, for example, when modeling the reliability of series systems (where system reliability is assumed to be the product of component reliabilities).

9 Some Theory of MCMC for Discrete Cases

The following exposition for discrete cases of $h(\boldsymbol{\eta})$ is based on old ISU lecture notes obtained from Noel Cressie. (Parallel theory for general cases can be found in Luke Tierney's December 1994 *Annals of Statistics* paper.)

9.1 General Theory

The question addressed here is how the theory of Markov Chains is useful in the simulation of realizations from a (joint) distribution for $\boldsymbol{\eta}$ specified by a function $h(\boldsymbol{\eta})$ proportional to a "density" (here a pmf).

Definition 12 *A (discrete time/discrete state space) Markov Chain is a sequence of random quantities $\{\boldsymbol{\eta}^k\}$, each taking values in a (finite or) countable set $\mathcal{X}$, with the property that*

$$P[\boldsymbol{\eta}^k = x_k | \boldsymbol{\eta}^1 = x_1, ..., \boldsymbol{\eta}^{k-1} = x_{k-1}] = P[\boldsymbol{\eta}^k = x_k | \boldsymbol{\eta}^{k-1} = x_{k-1}] \quad .$$

Definition 13 *A Markov Chain is stationary provided $P[\boldsymbol{\eta}^k = x_k | \boldsymbol{\eta}^{k-1} = x_{k-1}]$ is independent of k.*

WOLOG we will for the time being name the elements of $\mathcal{X}$ with the integers $1, 2, 3, ...$ and call them "states."

Definition 14 *With $p_{ij} \doteq P[\boldsymbol{\eta}^k = j | \boldsymbol{\eta}^{k-1} = i]$, the square matrix $P \doteq (p_{ij})$ is called the transition matrix for a stationary Markov Chain and the p_{ij} are called transition probabilities.*

Note that a transition matrix has nonnegative entries and row sums of 1. Such matrices are often called "stochastic" matrices. As a matter of further notation for a stationary Markov Chain, let

$$p_{ij}^t = P[\boldsymbol{\eta}^{t+k} = j | \boldsymbol{\eta}^k = i]$$

(this is the i, j entry of the tth power of P, $P^t = \overbrace{P \cdot P \cdot \cdots \cdot P}^{t \text{ factors}}$) and

$$f_{ij}^t = P[\boldsymbol{\eta}^{k+t} = j, \boldsymbol{\eta}^{k+t-1} \neq j, ..., \boldsymbol{\eta}^{k+1} \neq j | \boldsymbol{\eta}^k = i] \quad .$$

(These are respectively the probabilities of moving from i to j in t steps and *first* moving from i to j in t steps.)

Definition 15 *We say that a MC is irreducible if for each i and j $\exists$ t (possibly depending upon i and j) such that $p_{ij}^t > 0$.*

(A chain is irreducible if it is possible to eventually get from any state i to any other state j.)

Definition 16 *We say that the ith state of a MC is transient if $\sum_{t=1}^{\infty} f_{ii}^t < 1$ and say that the state is persistent if $\sum_{t=1}^{\infty} f_{ii}^t = 1$. A chain is called persistent if all of its states are persistent.*

(A state is transient if once in it, there is some possibility that the chain will never return. A state is persistent if once in it, the chain will with certainty be in it again.)

Definition 17 *We say that state i of a MC has period t if $p_{ii}^s = 0$ unless $s = \nu t$ (s is a multiple of t) and t is the largest integer with this property. The state is aperiodic if no such $t > 1$ exists. And a MC is called aperiodic if all of its states are aperiodic.*

Many sources (including Chapter 15 of the 3rd Edition of Feller Volume 1) present a number of useful simple results about MC's. Among them are the following.

Theorem 18 *All states of an irreducible MC are of the same type (with regard to persistence and periodicity).*

Theorem 19 *A finite state space irreducible MC is persistent.*

Theorem 20 *Suppose that a MC is irreducible, aperiodic, and persistent. Suppose further that for each state i the mean recurrence time is finite, i.e.*

$$\sum_{t=1}^{\infty} t f_{ii}^t < \infty \quad .$$

Then an invariant/stationary distribution for the MC exists, i.e. $\exists\ \{u_j\}$ with $u_j > 0$ and $\sum u_j = 1$ such that

$$u_j = \sum_i u_i p_{ij} \quad .$$

(If the chain is started with distribution $\{u_j\}$, after one transition it is in states $1, 2, 3, \ldots$ with probabilities $\{u_j\}$.) Further, this distribution$\{u_j\}$ satisfies

$$u_j = \lim_{t\to\infty} p_{ij}^t \ \ \forall\, i \quad ,$$

and

$$u_j = \frac{1}{\sum_{t=1}^{\infty} t f_{jj}^t} \quad .$$

There is a converse of this theorem.

Theorem 21 *An irreducible, aperiodic MC for which $\exists\ \{u_j\}$ with $u_j > 0$ and $\sum u_j = 1$ such that $u_j = \sum_i u_i p_{ij}$ must be persistent with $u_j = \frac{1}{\sum_{t=1}^{\infty} t f_{jj}^t}$.*

And there is an important "ergodic" result that guarantees that "time averages" have the right limits.

Theorem 22 *Under the hypotheses of Theorem 20, if g is a real-valued function such that*

$$\sum_j |g(j)| u_j < \infty$$

then for any j, if $\boldsymbol{\eta}^0 = j$

$$P\left[\frac{1}{n}\sum_{k=1}^{n} g(\boldsymbol{\eta}^k) \to \sum_j g(j) u_j\right] = 1$$

(Note that the choice of g as an indicator provides approximations for stationary probabilities.)

With this background, the basic idea of MCMC for Bayes computation is the following. If we wish to simulate from a distribution $\{u_j\}$ with

$$u_j \propto h(j) \tag{78}$$

or approximate properties of the distribution that can be expressed as moments of some function $g(j)$, we find a convenient P whose invariant distribution is $\{u_j\}$. From a starting state $\boldsymbol{\eta}^0 = i$, we use P to generate a value for $\boldsymbol{\eta}^1$. Using the realization of $\boldsymbol{\eta}^1$ and P, we generate $\boldsymbol{\eta}^2$, etc. Then one applies Theorem 22 to approximate the quantity of interest.

In answer to the question of "How does one argue that the common algorithms of Bayes computation have P's with invariant distribution proportional to $\{h(j)\}$?" there is the following useful sufficient condition (that has application in the original motivating problem of simulating from high dimensional distributions) for a chain to have $\{u_j\}$ for an invariant distribution.

Lemma 23 *If $\{\boldsymbol{\eta}^k\}$ is a MC with transition probabilities satisfying*

$$u_i p_{ij} = u_j p_{ji} \quad , \tag{79}$$

then it has invariant distribution $\{u_j\}$.

Note then that if a candidate P satisfies (79) and is irreducible and aperiodic, Theorem 21 shows that it is persistent. Theorem 20 then shows that any arbitrary starting value can be used and yields approximate realizations from $\{u_j\}$ and Theorem 22 implies that "time averages" can be used to approximate properties of $\{u_j\}$. Of course, in the Bayes context, it is distributions (78) that are of interest in MCMC from a posterior.

9.2 Application to the Metropolis-Hastings Algorithm

Sometimes MCMC schemes useful in Bayes computation can be shown to have the "correct" invariant distributions by observing that they satisfy (79). For example, Lemma 23 can be applied to the Metropolis-Hastings Algorithm. That is, let $T = (t_{ij})$ be any stochastic matrix corresponding to an irreducible aperiodic MC. This specifies, for each i, a jumping distribution. Note that in a finite case, one can take $t_{ij} = 1/(\text{the number of states})$. As indicated in Section 2.4, the Metropolis-Hastings algorithm operates as follows:

- Supposing that $\boldsymbol{\eta}^{k-1} = i$, generate J (at random) according to the distribution over the state space specified by row i of T (that is, according to $\{t_{ij}\}$).
- Then generate $\boldsymbol{\eta}^k$ based on i and (the randomly generated) J according to

$$\boldsymbol{\eta}^k = \begin{cases} J & \text{with probability} \quad \min\left(1, \frac{u_J t_{Ji}}{u_i t_{iJ}}\right) \\ i & \text{with probability} \quad \max\left(0, 1 - \frac{u_J t_{Ji}}{u_i t_{iJ}}\right) \end{cases} \tag{80}$$

Note that for $\{\boldsymbol{\eta}^k\}$ so generated, for $j \neq i$

$$p_{ij} = P[\boldsymbol{\eta}^k = j | \boldsymbol{\eta}^{k-1} = i] = \min\left(1, \frac{u_j t_{ji}}{u_i t_{ij}}\right) t_{ij}$$

and

$$p_{ii} = P[\boldsymbol{\eta}^k = i | \boldsymbol{\eta}^{k-1} = i] = t_{ii} + \sum_{j \neq i} \max\left(0, 1 - \frac{u_j t_{ji}}{u_i t_{ij}}\right) t_{ij} \quad .$$

So, for $i \neq j$

$$u_i p_{ij} = \min(u_i t_{ij}, u_j t_{ji}) = u_j p_{ji} \quad .$$

That is, (79) holds and the MC $\{\boldsymbol{\eta}^k\}$ has stationary distribution $\{u_j\}$. (Further, the assumption that T corresponds to an irreducible aperiodic chain implies that $\{\boldsymbol{\eta}^k\}$ is irreducible and aperiodic.)

As indicated in Section 2.4, in order to use the Metropolis-Hastings Algorithm one only has to know the u_j's up to a multiplicative constant. If (78) holds

$$\frac{u_J}{u_i} = \frac{h(J)}{h(i)}$$

we may write (80) as

$$\boldsymbol{\eta}^k = \begin{cases} J & \text{with probability} \quad \min\left(1, \frac{h(J) t_{Ji}}{h(i) t_{iJ}}\right) \\ i & \text{with probability} \quad \max\left(0, 1 - \frac{h(J) t_{Ji}}{h(i) t_{iJ}}\right) \end{cases} \tag{81}$$

Notice also that if T is symmetric, (i.e. $t_{ij} = t_{ji}$ and the jumping distribution is symmetric), (81) reduces to the Metropolis algorithm with

$$\boldsymbol{\eta}^k = \begin{cases} J & \text{with probability} \quad \min\left(1, \frac{h(J)}{h(i)}\right) \\ i & \text{with probability} \quad \max\left(0, 1 - \frac{h(J)}{h(i)}\right) \end{cases}$$

A variant of the Metropolis-Hastings algorithm is the "Barker Algorithm." The Barker algorithm modifies the above by replacing

$$\min\left(1, \frac{u_J t_{Ji}}{u_i t_{iJ}}\right) \text{ with } \frac{u_J t_{Ji}}{u_i t_{iJ} + u_J t_{Ji}}$$

and

$$\max\left(0, 1 - \frac{u_J t_{Ji}}{u_i t_{iJ}}\right) \quad \text{with} \quad \frac{u_i t_{iJ}}{u_i t_{iJ} + u_J t_{Ji}}$$

in (80). Note that for this algorithm, for $j \neq i$

$$p_{ij} = \left(\frac{u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}}\right) t_{ij} \ ,$$

so

$$u_i p_{ij} = \frac{(u_i t_{ij})(u_j t_{ji})}{u_i t_{ij} + u_j t_{ji}} = u_j p_{ji} \quad .$$

That is, (79) holds and thus Lemma 23 guarantees that under the Barker algorithm $\{\boldsymbol{\eta}^k\}$ has invariant distribution $\{u_j\}$. (And T irreducible and aperiodic continues to imply that $\{\boldsymbol{\eta}^k\}$ is also irreducible and aperiodic.)

Note also that since

$$\frac{u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}} = \frac{\frac{u_j}{u_i} t_{ji}}{t_{ij} + \frac{u_j}{u_i} t_{ji}}$$

once again it suffices to know the u_j up to a multiplicative constant in order to implement Barker's algorithm. In the Bayes posterior simulation context, this means that the Barker analogue of the Metropolis-Hastings form (81) is

$$\boldsymbol{\eta}^k = \begin{cases} J & \text{with probability } \frac{h(J)t_{Ji}}{h(i)t_{iJ}+h(J)t_{Ji}} \\ i & \text{with probability } \frac{h(i)t_{iJ}}{h(i)t_{iJ}+h(J)t_{Ji}} \end{cases} \tag{82}$$

and if T is symmetric, (82) becomes

$$\boldsymbol{\eta}^k = \begin{cases} J & \text{with probability } \frac{h(J)}{h(i)+h(J)} \\ i & \text{with probability } \frac{h(i)}{h(i)+h(J)} \end{cases}$$

9.3 Application to the Gibbs Sampler

Consider now the Gibbs Sampler (of Section 2.2). For sake of concreteness, consider the situation where the distribution of a discrete 3-dimensional random vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)$ with probability mass function proportional to $h(\boldsymbol{\eta})$ is at issue. One defines a MC $\{\boldsymbol{\eta}^k\}$ as follows. For an arbitrary starting state $\boldsymbol{\eta}^0 = \left(\eta_1^0, \eta_2^0, \eta_3^0\right)$ once one has $\boldsymbol{\eta}^{k-1} = \left(\eta_1^{k-1}, \eta_2^{k-1}, \eta_3^{k-1}\right)$:

- Generate $\boldsymbol{\eta}_1^{k-1} = \left(\eta_1^k, \eta_2^{k-1}, \eta_3^{k-1}\right)$ by generating η_1^k from the conditional distribution of $\eta_1 | \eta_2 = \eta_2^{k-1}$ and $\eta_3 = \eta_3^{k-1}$, i.e. from the (conditional) distribution with probability mass function $h_{\eta_1|\eta_2,\eta_3}(\eta_1 | \eta_2^{k-1}, \eta_3^{k-1}) \doteq \frac{h(\eta_1, \eta_2^{k-1}, \eta_3^{k-1})}{\sum_{\eta_1} h(\eta_1, \eta_2^{k-1}, \eta_3^{k-1})}$.

- Generate $\boldsymbol{\eta}_2^{k-1} = (\eta_1^k, \eta_2^k, \eta_3^{k-1})$ by generating η_2^k from the conditional distribution of $\eta_2 | \eta_1 = \eta_1^k$ and $\eta_3 = \eta_3^{k-1}$, i.e. from the (conditional) distribution with probability function $h_{\eta_2|\eta_1,\eta_3}(\eta_2 | \eta_1^k, \eta_3^{k-1}) \doteq \frac{h(\eta_1^k, \eta_2, \eta_3^{k-1})}{\sum_{\eta_2} h(\eta_1^k, \eta_2, \eta_3^{k-1})}$.

- Generate $\boldsymbol{\eta}^k = (\eta_1^k, \eta_2^k, \eta_3^k)$ by generating η_3^k from the conditional distribution of $\eta_3 | \eta_1 = \eta_1^k$ and $\eta_2 = \eta_2^k$, i.e. from the (conditional) distribution with probability function $h_{\eta_3|\eta_1,\eta_2}(\eta_3 | \eta_1^k, \eta_2^k) \doteq \frac{h(\eta_1^k, \eta_2^k, \eta_3)}{\sum_{\eta_3} h(\eta_1^k, \eta_2^k, \eta_3)}$.

Note that with this algorithm, a typical transition probability (for a step where a $\boldsymbol{\eta}_1^{k-1}$ is going to be generated) is

$$P[\boldsymbol{\eta}_1^{k-1} = \left(\eta_1, \eta_2^{k-1}, \eta_3^{k-1}\right) | \boldsymbol{\eta}^{k-1} = \left(\eta_1^{k-1}, \eta_2^{k-1}, \eta_3^{k-1}\right)] = \frac{h(\eta_1, \eta_2^{k-1}, \eta_3^{k-1})}{\sum_{\eta_1} h(\eta_1, \eta_2^{k-1}, \eta_3^{k-1})}$$

so if $\boldsymbol{\eta}^{k-1}$ has distribution specified by h, the probability that $\boldsymbol{\eta}_1^{k-1} = (\eta_1, \eta_2, \eta_3)$ is

$$\sum_{\gamma} \frac{h(\gamma, \eta_2, \eta_3)}{\sum_{\gamma, \eta_2, \eta_3} h(\gamma, \eta_2, \eta_3)} \frac{h(\eta_1, \eta_2, \eta_3)}{\sum_{\eta_1} h(\eta_1, \eta_2, \eta_3)} = \frac{h(\eta_1, \eta_2, \eta_3)}{\sum_{\eta_1, \eta_2, \eta_3} h(\gamma, \eta_2, \eta_3)}$$

so $\boldsymbol{\eta}_n$ also has distribution h And analogous results hold for the other two types of transitions (where $\boldsymbol{\eta}_2^{k-1}$ and $\boldsymbol{\eta}^k$ are to be generated). That is, direct calculation (as opposed to the use of Lemma 23) shows that if P_1, P_2 and P_3 are the 3 (different) transition matrices respectively for transitions $\boldsymbol{\eta}^{k-1} \to \boldsymbol{\eta}_1^{k-1}, \boldsymbol{\eta}_1^{k-1} \to \boldsymbol{\eta}_2^{k-1}$, and $\boldsymbol{\eta}_2^{k-1} \to \boldsymbol{\eta}^k$, they each have the distribution specified by h as their invariant distributions. This means that the transition matrix for $\boldsymbol{\eta}^{k-1} \to \boldsymbol{\eta}^k$, namely

$$P = P_1 P_2 P_3$$

also has the distribution specified by h as its invariant distribution, and describes a whole cycle of the Gibbs/Successive Substitution Sampling algorithm. $\left\{\boldsymbol{\eta}^k\right\}$ is thus a stationary Markov Chain with transition matrix P. So one is in a position to apply Theorems 21 and 22. **If P is irreducible and aperiodic** (this has to be checked), Theorem 21 says that the chain $\left\{\boldsymbol{\eta}^k\right\}$ is persistent and then Theorems 20 and 22 say that observations from h can be simulated using an arbitrary starting state.

9.4 Application to Metropolis-Hastings-in-Gibbs Algorithms

Consider now the kind of combination of Metropolis-Hastings and Gibbs sampling algorithms considered in Section 2.5. For sake of concreteness, suppose again that a discrete 3-dimensional random vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)$ with probability mass function proportional to $h(\boldsymbol{\eta})$ is at issue. Suppose further that it is clear how to make η_2 and η_3 updates using conditionals of $\eta_2|\eta_1, \eta_3$ and $\eta_3|\eta_1, \eta_2$ (these are recognizable as of a standard form) but that a "Metropolis-Hastings step" is to be used to make η_1 updates.

For an arbitrary starting state $\boldsymbol{\eta}^0 = \left(\eta_1^0, \eta_2^0, \eta_3^0\right)$ once one has $\boldsymbol{\eta}^{k-1} = \left(\eta_1^{k-1}, \eta_2^{k-1}, \eta_3^{k-1}\right)$, one first makes an η_1 update as follows. Suppose that for every pair (η_2, η_3),

$$t(\eta_1, \eta_1' | \eta_2, \eta_3)$$

specifies a transition matrix on the set of corresponding possible η_1's (for transitions $\eta_1 \to \eta_1'$), and for safety sake, let's require that all $t(\eta_1, \eta_1' | \eta_2, \eta_3) > 0$. Then

- sample η_1^{k*} from $t\left(\eta_1^{k-1}, \cdot | \eta_2^{k-1}, \eta_3^{k-1}\right)$, and

- set

$$\eta_1^k = \begin{cases} \eta_1^{k*} & \text{with probability } \min\left(1, \frac{h\left(\eta_1^{k*},\eta_2^{k-1},\eta_3^{k-1}\right)t\left(\eta_1^{k*},\eta_1^{k-1}|\eta_2^{k-1},\eta_3^{k-1}\right)}{h\left(\eta_1^{k-1},\eta_2^{k-1},\eta_3^{k-1}\right)t\left(\eta_1^{k-1},\eta_1^{k*}|\eta_2^{k-1},\eta_3^{k-1}\right)}\right) \\ \eta_1^{k-1} & \text{otherwise} \end{cases}$$

  and then just as in Section 9.3,

- generate η_2^k from the conditional distribution of $\eta_2|\eta_1 = \eta_1^k$ and $\eta_3 = \eta_3^{k-1}$, i.e. from the (conditional) distribution with probability function $h_{\eta_2|\eta_1,\eta_3}(\eta_2|\eta_1^k,\eta_3^{k-1}) \doteq \frac{h(\eta_1^k,\eta_2,\eta_3^{k-1})}{\sum_{\eta_2} h(\eta_1^k,\eta_2,\eta_3^{k-1})}$ and

- generate $\boldsymbol{\eta}^k = (\eta_1^k, \eta_2^k, \eta_3^k)$ by generating η_3^k from the conditional distribution of $\eta_3|\eta_1 = \eta_1^k$ and $\eta_2 = \eta_2^k$, i.e. from the (conditional) distribution with probability function $h_{\eta_3|\eta_1,\eta_2}(\eta_3|\eta_1^k,\eta_2^k) \doteq \frac{h(\eta_1^k,\eta_2^k,\eta_3)}{\sum_{\eta_3} h(\eta_1^k,\eta_2^k,\eta_3)}$.

We have already argued that the two "straight Gibbs updates" above have the distribution specified by h as their invariant distribution. We need to argue that the first (Metropolis-Hastings) step leaves the distribution specified by h invariant (notice that this is not obviously covered by the argument for the *overall* Metropolis-Hastings algorithm offered in Section 9.2). So suppose that $\boldsymbol{\eta}^{k-1}$ has distribution specified by h and consider the distribution of $\boldsymbol{\eta}_1^{k-1} = \left(\eta_1^k, \eta_2^{k-1}, \eta_3^{k-1}\right)$ obtained from $\boldsymbol{\eta}^{k-1}$ by employing the Metropolis-Hastings step to replace η_1^{k-1}.

$$P\left[\boldsymbol{\eta}_1^{k-1} = (\eta_1', \eta_2, \eta_3)\right] = \sum_{\eta_1} P\left[\boldsymbol{\eta}^{k-1} = (\eta_1, \eta_2, \eta_3)\right] \cdot P\left[(\eta_1, \eta_2, \eta_3) \to (\eta_1', \eta_2, \eta_3)\right]$$

where $(\eta_1, \eta_2, \eta_3) \to (\eta_1', \eta_2, \eta_3)$ is shorthand for the Metropolis-Hastings step resulting in the indicated transition. Then if $\kappa = \sum_{\boldsymbol{\eta}} h\left(\boldsymbol{\eta}\right)$ so that it is $\frac{1}{\kappa}h$ that

is the pmf of interest,

$$\begin{aligned}
& \kappa \cdot P\left[\boldsymbol{\eta}_1^{k-1} = (\eta_1', \eta_2, \eta_3)\right] \\
= & \sum_{\eta_1 \neq \eta_1'} h(\eta_1, \eta_2, \eta_3)\, t(\eta_1, \eta_1'|\eta_2, \eta_3) \min\left(1, \frac{h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1|\eta_2, \eta_3)}{h(\eta_1, \eta_2, \eta_3)\, t(\eta_1, \eta_1'|\eta_2, \eta_3)}\right) \\
& + h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1'|\eta_2, \eta_3) \cdot 1 \\
& + h(\eta_1', \eta_2, \eta_3) \sum_{\eta_1 \neq \eta_1'} t(\eta_1', \eta_1|\eta_2, \eta_3) \max\left(0, 1 - \frac{h(\eta_1, \eta_2, \eta_3)\, t(\eta_1, \eta_1'|\eta_2, \eta_3)}{h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1|\eta_2, \eta_3)}\right) \\
= & \sum_{\eta_1 \neq \eta_1'} \min\left(h(\eta_1, \eta_2, \eta_3)\, t(\eta_1, \eta_1'|\eta_2, \eta_3), h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1|\eta_2, \eta_3)\right) \\
& + h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1'|\eta_2, \eta_3) \\
& + \sum_{\eta_1 \neq \eta_1'} \max\left(0, h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1|\eta_2, \eta_3) - h(\eta_1, \eta_2, \eta_3)\, t(\eta_1, \eta_1'|\eta_2, \eta_3)\right) \\
= & h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1'|\eta_2, \eta_3) + \sum_{\eta_1 \neq \eta_1'} h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1|\eta_2, \eta_3) \\
= & \sum_{\eta_1} h(\eta_1', \eta_2, \eta_3)\, t(\eta_1', \eta_1|\eta_2, \eta_3) \\
= & h(\eta_1', \eta_2, \eta_3)
\end{aligned}$$

and $\boldsymbol{\eta}_1^{k-1} = \left(\eta_1^k, \eta_2^{k-1}, \eta_3^{k-1}\right)$ also has the distribution specified by h. That is, the Metropolis-Hastings step leaves the distribution specified by h invariant. This can be represented by some transition matrix P_1 for the $\boldsymbol{\eta}^{k-1} \to \boldsymbol{\eta}_1^{k-1}$ transition. Then if as in Section 9.3, P_2 and P_3 represent respectively $\boldsymbol{\eta}_1^{k-1} \to \boldsymbol{\eta}_2^{k-1}$, and $\boldsymbol{\eta}_2^{k-1} \to \boldsymbol{\eta}^k$ transitions, the whole transition matrix for $\boldsymbol{\eta}^{k-1} \to \boldsymbol{\eta}^k$

$$P = P_1 P_2 P_3$$

has the distribution specified by h as its invariant distribution, and describes a complete cycle of the Metropolis-Hastings in Gibbs algorithm. $\left\{\boldsymbol{\eta}^k\right\}$ is thus a stationary Markov Chain with transition matrix P. So again one is in a position to apply Theorems 21 and 22. **If P is irreducible and aperiodic** (this has to be checked), Theorem 21 says that the chain $\left\{\boldsymbol{\eta}^k\right\}$ is persistent and then Theorems 20 and 22 say that observations from h can be simulated using an arbitrary starting state.

9.5 Application to "Alternating" Algorithms

The kind of logic used above in considering the Gibbs and Metropolis-Hastings-in-Gibbs algorithms suggests another variant on MCMC for Bayes computation. That is, one might think about alternating in some regular way between two or more basic algorithms. That is, if P_{Gibbs} is a transition matrix for a complete cycle of Gibbs substitutions and $P_{\text{M-H}}$ is a transition matrix for an iteration of

a Metropolis-Hastings algorithm, then

$$P = P_{\text{Gibbs}} P_{\text{M-H}}$$

is a transition matrix for an algorithm that can be implemented by following a Gibbs cycle with a Metropolis-Hastings iteration, followed by a Gibbs cycle and so on. It's possible that in some cases that such an alternating algorithm might avoid difficulties in "mixing" that would be encountered by either of the component algorithms applied alone.