Chapter 2 Bayesian Essentials

Using this Chapter

This chapter provides a self-contained introduction to Bayesian Inference. For those who need a refresher in distribution theory, section 2.0 provides an introduction to marginal, joint, and conditional distributions and their associated densities. We then develop the basics of Bayesian inference, discuss the role of subjective probability and priors and provide some of the most compelling arguments for adopting the Bayesian point of view. Regression models (both univariate and multivariate) are considered along with their associated natural conjugate priors. Asymptotic approximations and Importance Sampling are introduced as methods for non-conjugate models. Finally, a simulation primer for the basic distributions/models in Bayesian Inference is provided. Those who want a basic introduction to Bayesian inference without many details should concentrate on sections 2.1-2.5 and section 2.9.1.

2.0 Essential Concepts from Distribution Theory

Bayesian inference relies heavily on probability theory and, in particular, distributional theory. This section provides a review of basic distributional theory with examples designed to be relevant to Bayesian applications.

A basic starting point for probability theory is a discrete random variable, X. X can take on a discrete number of values, each with some probability. The classic example would be a Bernouli random variable. X = 1 with probability p and 0 with probability p-1. X denotes some event such as whether a company will sell a product tomorrow. p represents the probability of a sale. For now, let us set aside the question of whether this probability can represent a long run frequency or whether it represents a subject probability (note: it is hard to understand the long-run frequency argument for this example since it requires us to imagine an infinite number of "other-worlds" for the event of a sale tomorrow). We can easily extend this example to the number of units sold tomorrow. Then X is still discrete but can take on the values 0, 1, 2, …. , m with probabilities, $p_0, p_1, \ldots, p_m$. X now has a non-trivial probability distribution. With knowledge of this distribution, we can answer any

question such as the probability that there will be at least one sale tomorrow, the probability that there will be between 1 and 10 sales, etc. In general, we can compute the probability that sales will be in any set simply by summing over the probabilities of the elements in the set.

(2.0.1) $$\Pr(X \in A) = \sum_{x \in A} p_x$$

We can also compute the *expectation* of the number of units sold tomorrow as the average over the probability distribution.

(2.0.2) $$E[X] = \sum_{i=0}^{m} i p_i$$

If we are looking at aggregate sales of a popular consumer product, we might approximate sales as a *continuous* random variable which can take on any non-negative real number. For this situation, we must summarize the probability distribution of X by a probability density. A density function is a *rate* function which tells us the probability per volume or unit of X. X has a density function, $p_X(x)$; p_X is a positive-valued function which integrates to one. The probability that X takes on any set of values we must integrate $p_X(\,)$ over this set.

(2.0.3) $$\Pr(X \in A) = \int_A p_X(x|\theta)\,dx$$

This is very much the analogue of the discrete sum in (2.0.1). The sense in which p is a rate function is that the probability that $X \in (x_0, x_0 + dx)$ is approximately $p_X(x_0)\,dx$. Thus, the probability density function, $p_X(\,)$, plays the same role as the discrete probabilities (sometimes called probability mass function) in the discrete case. We can easily find the expectation of any function of X by computing the appropriate integral.

(2.0.4) $$E[f(X)] = \int f(x) p(x|\theta)\,dx$$

In many situations, we will want to consider the *joint* distribution of two or more random variables, both of which are continuous. For example, we might consider the joint distribution of sales tomorrow in two different markets. Let X denote the sales in market A and Y denote the sales in market B. For this situation, there is a bivariate density function, $p_{X,Y}(x,y)$. This density gives the probability rate per unit of area in the plane. That is, the probability that both $X \in (x_0, x_0 + dx)$ and $Y \in (y_0, y_0 + dy)$ is approximately, $p_{X,Y}(x_0, y_0)dxdy$. With the joint density, we compute the probability of any set of (X,Y) values. For example, we can compute the probability that both X and Y are positive. This is the area of under the density for the positive orthant.

$$\Pr(X > 0 \text{ and } Y > 0) = \int_0^\infty \int_0^\infty p_{X,Y}(x,y)dxdy \tag{2.0.5}$$

For example, the multinomial probit model, considered in chapter 4, has choice probabilities defined the integrals of a multivariate normal density over various cones. If $p_{X,Y}(\ )$ is a bivariate normal density, then (2.0.5) is one such equation.

Given the joint density, we can also compute the *marginal* densities of each of the variables X and Y. That is to say, if we know everything thing about the joint distribution, we certainly know everything about the marginal distribution. The way to think of this is via simulation. Suppose we were able to simulate from the joint distribution. If we look at the simulated distribution of either X or Y alone, we have simulated the marginal distribution.

To find the marginal density of X, we must average the *joint* density over all possible values of Y.

$$p_X(x) = \int p_{X,Y}(x,y)dy \tag{2.0.6}$$

A simple example will help make this idea clear. Suppose X, Y are uniformly distributed over the triangle, $\{X, Y : 0 < X < 1 \text{ and } Y < X\}$, depicted in figure 2.1. A uniform distribution means that the density is constant over the shaded triangle. The area of this triangle is ½ so this means that the density must be 2 in order to insure that the joint density integrates to 1.

$$\int_0^1 \int_0^y p_{X,Y}(x,y)\,dxdy = \int_0^1 \int_0^y 2\,dxdy = \int_0^1 \left(2x\Big|_0^y\right)dy = \int_0^1 2y\,dy = y^2\Big|_0^1 = 1$$

This means that the joint density is a surface over the triangle with height 2.

We can use (2.0.6) to find the marginal distribution of X by integrating out Y.

$$p_X(x) = \int p_{X,Y}(x,y)\,dy = \int_0^x 2\,dy = 2y\Big|_0^x = 2x$$

Thus, the marginal distribution of X is not uniform! The density increases as x increases toward 1. The marginal density of Y can easily be found to be of the "reverse" shape, $p_Y(y) = 2 - 2y$. This makes intuitive sense as the joint density is "widest" with X near one and with Y near 0.

We can also define the concept of a conditional distribution and conditional density. If X, Y have a joint distribution, we can ask what is the conditional distribution of Y given X=x? If X, Y are continuous random variables, then the conditional distribution of Y given X=x is also a continuous random variable. The conditional density of Y | X can be derived from the marginal and joint densities (the Borel paradox not withstanding).

(2.0.7)
$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

The argument of the conditional density on the left hand side of (2.0.7) is written $y|x$ to emphasize that there is a different density for every value of the conditioning argument x.

We note that the conditional density is proportional to the joint! The marginal only serves to get the right normalization.

Let's return to our simple example. The conditional distribution of $Y|X=x$ is simply a slice of the joint density along a vertical line at the point x. This is clearly uniform but only extends from 0 to x. We can use (2.0.7) to get the right normalization.

$$p_{Y|X}(y|x) = \frac{2}{2x};\ y \in (0,x)$$

Thus, if x=1, then the density is uniform over (0,1) with height 1. The dependence between X and Y is only evidenced by the fact the range of Y is restricted by the value of x.

In many statistics courses, we are taught that correlation is a measure of the dependence between two random variables. This stems from the bivariate normal distribution which uses correlation to drive the shape of the joint density. Let's start with two independent standard normal random variables, Z and W. This means that their joint density factors (this is because of the product rule for independent events).

(2.0.8) $$p_{Z,W}(z,w) = p_Z(z)\,p_W(w)$$

Each of the standard normal densities is given by:

(2.0.9) $$p_Z(z) = \tfrac{1}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}z^2\right)$$

If we create X and Y by an appropriate linear combination of Z and W, we can create correlated or dependent random variables.

$$X = Z$$
$$Y = \rho Z + \sqrt{\left(1-\rho^2\right)}W$$

X and Y have a correlated bivariate normal density with correlation coefficient ρ.

(2.0.10) $$p_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[x^2 - 2\rho xy + y^2\right]\right\}$$

It is possible to show that $\mathrm{cov}(X,Y) = E[XY] = \iint xy p_{X,Y}(x,y)\,dxdy$ is ρ. Both X, Y have marginal distributions which are standard normal and conditional distributions which are also normal but with a mean that depends on the conditioning argument.

$$X \sim N(0,1);\ Y \sim N(0,1);\ Y \mid X = x \sim N\left(\rho x, \left(1-\rho^2\right)\right)$$

We will return to this example when we consider methods of simulation from the bivariate and multivariate normal distributions. We also will consider this situation when introducing the Gibbs Sampler in chapter 3.

2.1 The Goal of Inference and Bayes Theorem

The goal of statistical inference is to use information to make inferences about unknown quantities. One important source of information is data but there is an undeniable role for non data-based information. Information can also come from theories of behavior (such as the information that, properly defined, demand curves slope downward). Information can also come from "subjective" views that there is a structure underlying the unknowns. For example, in situations with large numbers of different sets of parameters, an assumption that the parameters sets "cluster" or that they are drawn from some common distribution is often used in modeling. Less controversial might be the statement that we expect key quantities to be finite or even in some range (such as a price elasticity is not expected to be less than -50). Information can also be derived from prior analyses of other data, including data which is only loosely related to the dataset under investigation.

An unknown quantity is a generic term referring to any value not known to the investigator. Certainly, parameters can be considered unknown since these are purely

abstractions that index a class of models. In situations in which decisions are made, the unknown quantities can include the, as yet unrealized, outcomes of marketing actions. Even in a passive environment, predictions of "future" outcomes are properly regarded as unknowns. There should be no distinction between a parameter and an unknown such as an unrealized outcome in the sense that the system of inference should treat each symmetrically.

Our goal, then, is to make inferences regarding unknown quantities *given* the information available. We have concluded that the information available can be partitioned into information obtained from the data as well as other information obtained independently or *prior* to the data. Bayesian inference utilizes probability statements as the basis for inference. What this means is that our goal is to make probability statements about unknown quantities *conditional* on the sample and prior information.

In order to utilize the elegant apparatus of conditional probability, we must encode the prior information as a probability distribution. This requires the view that probability can represent subjective beliefs and is not some sort of long run frequency. There is much discussion in the statistics and probability theory literature as to whether or not this is a reasonable thing to do. We take a somewhat more practical view – there are many kinds of non-data based information to be incorporated into our analysis. A subjective interpretation of probability is a practical necessity rather than a philosophical curiosity.

It should be noted that there are several paths which lead to the conclusion that Bayesian inference is a sensible system of inference. Some start with the view that decision makers are expected utility maximizers. In this world, decision makers must be "coherent" or act in accordance with Bayes theorem in order to avoid exposing themselves to sure losses. Others start with the view that the fundamental primitive is not utility but subjective probability. Still others adhere to the view that the likelihood principle (section 2.2 below)

more or less forces you to adopt the Bayesian form of inference. We are more of the subjectivist stripe but we hope to convince the reader, by example, that there is tremendous practical value to the Bayesian approach.

Bayes Theorem

Denote the set of unknowns as θ. Our prior beliefs are expressed as a probability distribution, $p(\theta)$. $p(\bullet)$ is a generic notation for the appropriate density. In most cases, this represents a density with respect to standard Lebesgue measure but it can also represent a probability mass function for discrete parameter spaces or a mixed continuous-discrete measure. The information provided by the data is introduced via the probability distribution for the data, $p(D|\theta)$, where "D" denotes the observable data. In some classical approaches, modeling is the art of choosing appropriate probability models for the data. In the Bayesian paradigm, the model for prior information is also important. Much of the work in Bayesian statistics is focused on developing a rich class of models to express prior information and devices to induce priors on high dimensional spaces. In our view, the prior is very important and often receives insufficient attention.

To deliver on the goal of inference, we must combine the prior and likelihood to produce the distribution of the observables conditional on the data and the prior. Bayes Theorem is nothing more than an application of standard conditional probability to this problem.

(2.1.1) $$p(\theta|D) = \frac{p(D,\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{p(D)}$$

$p(\theta|D)$ is called the *posterior* distribution and reflects the combined data and prior information. (2.1.1) is often expressed using the likelihood function. Given D, any

function which is proportional to $p(D|\theta)$ is call the "likelihood," $\ell(\theta)$. The shape of the posterior is determined entirely by the likelihood and prior in the numerator of (2.1.1) and this is often emphasized by rewriting the equation.

(2.1.2) $$p(\theta|D) \propto \ell(\theta)p(\theta)$$

If $\ell(\theta) = p(D|\theta)$, then the constant of proportionality is the marginal distribution of the data, $p(D) = \int p(D,\theta)d\theta = \int p(D|\theta)p(\theta)d\theta$. Of course, we are assuming here that this normalizing constant exists. If $p(\theta)$ represents a proper distribution (i.e. it integrates to one), then this integral exists. With improper priors, it will be necessary to show that the integral exists which will involve the tail behavior of the likelihood functions.

2.2 Conditioning and the Likelihood Principle

The likelihood principle states that the likelihood function, $\ell(\theta)$, contains all relevant information from the data. Two samples (not necessarily even from the same "experiment" or method of sampling/observation) have equivalent information regarding θ if their likelihoods are proportional (see Berger and Wolpert (1984) for extensive discussion and derivation of the LP from conditioning and sufficiency principles). The likelihood principle, by itself, is not sufficient to build a method of inference but should be regarded as a minimum requirement of any viable form of inference. This is a controversial point of view for anyone familiar with the modern econometric literature. Much of this literature is devoted to methods that do not obey the likelihood principle. For example, the phenomenal success of estimators based on the Generalized Method of Moments procedure is driven by the ease of implementing these estimators even though, in most instances, GMM estimators violate the likelihood principle.

Adherence to the likelihood principle means that inferences are *conditional* on the observed data as the likelihood function is parameterized by the data. This is worth contrasting to any sampling-based approach to inference. In the sampling literature, inference is conducted by examining the sampling distribution of some estimator of $\theta, \hat{\theta} = f(D)$. Some sort of sampling experiment[1] results in a distribution of D and, therefore, the estimator is viewed as a random variable. The sampling distribution of the estimator summarizes the properties of the estimator *prior* to observing the data. As such, it is irrelevant to making inferences given the data we actually observe. For any finite sample, this distinction is extremely important. One must conclude that, given our goal for inference, sampling distributions are simply not useful.

While sampling theory does not seem to deliver on the inference problem, it is possible to argue that it is relevant to the choice of estimating procedures. Bayesian inference procedures are simply one among many possible methods of deriving estimators for a given problem. Sampling properties are relevant to choice of procedures before the data is observed. As we will see in section 2.5 below, there is an important sense in which one need never look farther than Bayes estimators even if the sole criterion is the sampling properties of the estimator.

2.3 Prediction and Bayes

One of the appeals of the Bayesian approach is that all unknowns are treated the same. Prediction is defined as making probability statements about the distribution of as yet unobserved data, denoted by D_f . The only real distinction between "parameters" and unobserved data is that D_f is potentially observable.

[1]In the standard treatment, the sampling experiment consists of draws from the probability model for the data used in the likelihood. However, many other experiments are possible including samples drawn from some

(2.3.1) $$p(D_f \mid D) = \int p(D_f, \theta \mid D)\, d\theta = \int p(D_f \mid \theta, D)\, p(\theta \mid D)\, d\theta$$

(2.3.1) defines the "predictive" distribution of D_f given the observed data. In many cases, we assume that D and D_f are independent, conditional on θ. In this case, the predictive distribution simplifies.

(2.3.2) $$p(D_f \mid D) = \int p(D_f \mid \theta)\, p(\theta \mid D)\, d\theta$$

In (2.3.2), we average the likelihood for the unobserved data over the posterior of θ. This averaging properly accounts for uncertainty in θ when forming predictive statements about D_f.

2.4 Summarizing the Posterior

For any problem of practical interest, the posterior distribution is a high dimensional object. Therefore, summaries of the posterior play an important role in Bayesian statistics. Most schooled in classical statistical approaches are accustomed to reporting parameter estimates and standard errors. The Bayesian analogue of this practice is to report moments of the marginal distributions of parameters such as the posterior mean and posterior standard deviations. It is far more useful and informative to produce the marginal distributions of parameters or relevant functions of parameters as the output of the analysis. Simulation methods are ideally suited for this. If we can simulate from the posterior distribution of the parameters and other unknowns, then we can simply construct the marginal of any function of interest. Typically, we describe these marginals graphically. As these distributions are often very non-normal, the mean and standard deviations are not particularly useful. One major purpose of this book is to introduce a set of useful simulation tools to achieve this goal of simulating from the posterior distribution.

---

other model for the data or various asymptotic experiments which involve additional assumptions regarding the

Prior to the advent of powerful simulation methods, attention focused on the evaluation of specific integrals of the posterior distribution as a way of summarizing this high dimensional object. The general problem can be written as finding the posterior expectation of a function of θ. (We note that marginal posteriors, moments, quantiles and probability of intervals are further examples of expectations of functions as in (2.4.1) below with suitably defined h). For any interesting problem, only the un-normalized posterior, $\ell(\theta)p(\theta)$ is available so that two integrals must be performed to obtain the posterior expectation of $h(\theta)$

$$E_{\theta|D}\left[h(\theta)\right] = \int h(\theta)p(\theta|D)d\theta = \frac{\int h(\theta)\ell(\theta)p(\theta)d\theta}{\int \ell(\theta)p(\theta)d\theta} \tag{2.4.1}$$

For many years, only problems for which the integrals in (2.4.1) could be performed analytically were analyzed by Bayesians. Obviously, this restricts the set of priors and likelihoods to a very small set that produces posteriors of known distributional form and for which these integrals can be evaluated analytically. One approach would be to take various asymptotic approximations to these integrals. We will discuss the Laplace approximation method in section 2.9 below. Unless these asymptotic approximations can be shown to be accurate, we should be very cautious about using them. In contrast, much of the econometrics and statistics literature uses asymptotic approximations to the sampling distributions of estimators and test statistics without investigating accuracy. In marketing problems, the combination of small amounts of sample information per parameter and the discrete nature of the data makes it very risky to use asymptotic approximations. Fortunately, we do not have to rely on asymptotic approximations in modern Bayesian inference.

---

data generation process.

2.5 Decision Theory, Risk and the Sampling Properties of Bayes Estimators

We started our discussion by posing the problem of obtaining a system of inference appropriate for marketing problems. We could just as well have started on the most general level – finding an appropriate framework for making decisions of any kind. Parameter estimation is only one of many such decisions which occur under uncertainty.

The general problem considered in decision theory is to search among possible actions for the action which minimizes expected loss. The loss function, $L(a,\theta)$, associates a loss with a state of nature (θ) and an action a. In chapter 6, loss functions are derived for marketing actions from the profit function of the firm. We choose a decision which performs well, on average, where the averaging is taken across the posterior distribution of states of nature.

$$\min_a \left\{ \bar{L}(a) = E_{\theta|D}\left[L(a,\theta)\right] = \int L(a,\theta)\, p(\theta|D)\, d\theta \right\} \tag{2.5.1}$$

In chapter 6, we will explore the implications of decision theory for optimal marketing decisions and valuing of information sets. At this point it is important to note that (2.5.1) involves the entire posterior distribution and not just the posterior mean. With non-linear loss functions, uncertainty or spread is just as important as location.

A special case of (2.5.1) is the estimation problem. If the action is the estimator and the state of nature is the unknowns to be estimated, then Bayesian decision theory produces a Bayes estimator. Typically, a symmetric function such as squared error or absolute error is used for loss. This defines the estimation problem as

$$\min_{\hat{\theta}} \left\{ L(\hat{\theta}) = E_{\theta|D}\left[L(\hat{\theta},\theta) \right] \right\}. \tag{2.5.2}$$

For squared error loss, the optimal choice of estimator is the posterior mean.

(2.5.3) $$\hat{\theta}_{\text{Bayes}} = \mathrm{E}\left[\theta \middle| \mathrm{D}\right] = \mathrm{f}\left(\mathrm{D}|\tau\right)$$

Here τ is the prior hyper-parameter vector (if any).

What are the sampling properties of the Bayes estimator and how do these compare to those of other competing general purpose estimation procedures such as Maximum Likelihood? Recall the sampling properties are derived from the fact that the estimator is a function of the data and therefore is a random variable whose distribution is inherited from the sampling distribution of the data. We can use the same loss function to define the "risk" associated with an estimator, $\hat{\theta}$, as

(2.5.4) $$r_{\hat{\theta}}(\theta) = \mathrm{E}_{\mathrm{D}|\theta}\left[\mathrm{L}\left(\hat{\theta},\theta\right)\right] = \int \mathrm{L}\left(\hat{\theta}(\mathrm{D}),\theta\right)\mathrm{p}\left(\mathrm{D}|\theta\right)\mathrm{dD}$$

Note that the risk function for an estimator is a function of θ. That is, we have a different "risk" at every point in the parameter space.

An estimator is said to be *admissible* if there exists no other estimator with a risk function that is less than or equal to the risk of the estimator in question. That is, we cannot find another estimator that does better (or at least as well, as measured by risk, for every point in the parameter space.[2] Define expected risk, $\mathrm{E}\left[\mathrm{r}(\theta)\right] = \mathrm{E}_{\theta}\left[\mathrm{E}_{\mathrm{D}|\theta}\left[\mathrm{L}\left(\hat{\theta},\theta\right)\right]\right]$. The outer expectation on the right hand side is taken with respect to the prior distribution of θ. With a proper prior that has support over the entire parameter space, we can apply Fubini's theorem and interchange the order of integration and show that Bayes estimators have the property of minimizing expected risk and, therefore, are admissible.

[2] Obviously, if we have a continuous parameter space, we have to be a little more careful but we leave those niceties for those more mathematically inclined.

$$
\begin{aligned}
E\left[r(\theta)\right] &= E_{\theta}\left[E_{D|\theta}\left[L\left(\hat{\theta},\theta\right)\right]\right] = \iint L\left(\hat{\theta}(D),\theta\right)p\left(D|\theta\right)p(\theta)\,dD\,d\theta \\
&= E_{D}\left[E_{\theta|D}\left[L\left(\hat{\theta},\theta\right)\right]\right] = \iint L\left(\hat{\theta}(D),\theta\right)p\left(\theta|D\right)p(D)\,dD\,d\theta
\end{aligned}
\tag{2.5.5}
$$

The complete class theorem (see Berger (1985), Chapter 8) says even more – all admissible estimators are Bayes estimators. This provides a certain level of comfort and moral superiority but little practical guidance. There can be estimators that outperform Bayes estimators in certain regions of, but not all, of the parameter space. Bayes estimators perform very well if you are in the region of the parameter space you expect to be in as defined by your prior. These results on admissibility also don't provide any guidance as to how to choose among infinite number of Bayes estimators which are equivalent from the point of view of admissibility.

Another useful question to ask is what is the relationship between standard classical estimators such as the MLE and Bayes estimators? At least the MLE obeys the likelihood principle. In general, the MLE is not admissible so there can be no exact sample relationship. However, Bayes estimators are consistent, asymptotically normal and efficient as long as mild regularity conditions[3] hold and the prior is non-dogmatic in the sense of giving support to the entire parameter space. The asymptotic "duality" between Bayes estimators and the MLE stems from the asymptotic behavior of the posterior distribution. As n increases, the posterior concentrates more and more mass in the vicinity of the "true" value of θ. The likelihood term dominates the prior and the prior becomes more and more uniform in appearance in the region in which the likelihood is concentrating. Thus, the prior has no asymptotic influence and the posterior starts to look more and more normal.

[3] It should be noted that, as the MLE is based on a maximum of a function while the Bayes estimator is based on an average, the conditions for asymptotic normality are different for the MLE than for the Bayes estimator.

(2.5.6) $$p(\theta|D) \sim N\left(\hat{\theta}_{MLE}, \left[-H_{\theta=\hat{\theta}_{MLE}}\right]^{-1}\right)$$

H_θ is the Hessian of the log-likelihood function. The very fact that, for asymptotics, the prior doesn't matter (other than its support) should be reason enough to abandon this method of analysis in favor of more powerful techniques.

2.6 Identification and Bayesian Inference

The set of models is only limited by the imagination of the investigator and the computational demands of the model and inference method. In marketing problems, we can easily write down a model that is very complex and may make extraordinary demands of data. A problem of identification is defined as the situation in which there is a set of different parameter values that give rise to the same distribution for the data. This set of parameter values are said to observationally equivalent in the sense that the distribution of the data is the same for any member of this set.

Lack of identification is a property of the model and holds over all possible values of the data rather than just the data observed. Lack of identification implies that there will be regions over which the likelihood function is constant for any given data set. Typically, these can be flats or ridges in the likelihood.

From a purely technical point of view, identification is not a problem for a Bayesian analysis. First of all, the posterior may not have a "flat" or region of constancy as the prior can modify the shape of the likelihood. Even if there are regions for which the posterior is constant, the Bayesian will simply report, correctly, that the posterior is uniform over these regions.

But both from a practical (i.e. computational) and theoretical perspective, averages behave more regularly than maxima.

Lack of identification is often regarded as a serious problem which must be dealt with by imposing some sort of restriction on the parameter space. Methods of inference, such as maximum likelihood, that rely on maximization will encounter severe problems with unidentified models. A maximizer may climb up and shut down anywhere in flat of the likelihood created by lack of identification. Bayesian computational methods which use simulation as the basis for exploring the posterior are not as susceptible to these computational problems.

Rather than imposing some sort of constraint on the parameter space, the Bayesian can deal with lack of identification through an informative prior. With the proper informative prior, the posterior may not have any region of constancy. However, it should be remembered that lack of identification means that there are certain functions of the parameters for which posterior is entirely driven by the prior. We can define a transformation function, $\tau = f(\theta)$, and a partition, $\tau' = (\tau_1', \tau_2')$ where $\dim(\tau_1') =$ r, such that $p(\tau_1|D) = p(\tau_1)$. This means that, for certain transformed coordinates, the posterior is the same as the prior and only prior information matters in a posteriori inference. r is the "dimension" of the redundancies in the θ parameterization. The "solution" to the identification problem can then be to report the marginal posterior for the parameters which are "identified," $p(\tau_2|D)$. We will see this idea is useful for analysis of certain non-identified models. It is important, however, to examine the marginal prior, $p_{\tau_2}(\bullet)$, which is induced by the prior on θ.

2.7 Conjugacy, Sufficiency and Exponential Families

Prior to modern simulation methods, a premium was placed on models that would allow analytical expressions for various posterior summaries. Typically, this means that we choose

models for which we can compute posterior moments analytically. What we want is for the posterior to be of a distributional form for which the posterior moments are available in analytical expressions. This requirement imposes constraints on both the choice of likelihood and prior distributions. One approach is to require that the prior distribution be conjugate to the likelihood. A prior is said to be *conjugate* to the likelihood if the posterior derived from this prior and likelihood is in the same class of distributions as the prior. For example, normal distributions have simple expressions for moments. We can get normal posteriors by combing normal priors and likelihoods based on normal sampling models as we will see in section 2.8. However, there can be conjugate priors for which no analytic expressions exist for posterior moments.

The key to conjugacy is the form of the likelihood since we can always pick priors with convenient analytic features. Likelihoods in the exponential family of distributions have conjugate priors (see Bernardo and Smith (1994), section 5.2). The exponential family is a family of distributions with a minimal sufficient statistic of fixed dimension equal to the dimension of the parameter space. The duality of sufficient statistics and the parameter vector is what drives this result. Moreover, the exponential form means that combining an exponential family likelihood and prior will result in an exponential family posterior that is tighter than either the likelihood or the prior. To see this, recall the form of the regular parameterization of the exponential family. If we have a random sample $y' = (y_1, \ldots, y_n)$ from the regular exponential family, the likelihood is given by

(2.7.1) $$p(y|\theta) \propto g(\theta)^n \exp\left\{\sum_{j=1}^{k} c_j \phi_j(\theta) \bar{h}_j(y)\right\}$$

$\bar{h}_j(y) = \sum_{i=1}^{n} h_j(y_i)$. $\{\bar{h}_1, \ldots, \bar{h}_k\}$ are the set of minimal sufficient statistics for θ. A prior in the same form would be given by:

(2.7.2) $$p(\theta|\tau) \propto g(\theta)^{\tau_0} \exp\left\{\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j\right\}$$

$\{\tau_0, \tau_1, \ldots, \tau_k\}$ are the prior hyperparameters. Clearly, the posterior is also in the exponential form with parameters $\tau_0^* = n + \tau_0, \tau_1^* = \tau_1 + \bar{h}_1, \ldots, \tau_k^* = \tau_k + \bar{h}_k$.

(2.7.3) $$p(\theta|y) \propto g(\theta)^{\tau_0 + n} \exp\left\{\sum_{j=1}^{k} c_j \phi_j(\theta)\left(\bar{h}_j + \tau_j\right)\right\}$$

Since the conjugate prior in (2.7.2) is of the same form as the likelihood, we can interpret the prior as the posterior from some other sample of data with τ_0 observations.

A simple example will illustrate the functioning of natural conjugate priors. Consider the Bernoulli probability model for the data, $y_i \sim \text{iidB}(\theta)$. θ is the probability of one of two possible outcomes for each $y_i = (0,1)$. The joint density of the data is given by

(2.7.4) $$p(y|\theta) = \theta^{\sum_i y_i} (1-\theta)^{n - \sum_i y_i}$$

The choice of prior for this simple Bernoulli problem has been the subject of numerous articles. Some advocate the choice of a "reference" prior or a prior that meets some sort of criteria for default scientific applications. Others worry about the appropriate choice of a "non-informative" prior. Given that this problem is one dimensional, one might want to choose a fairly flexible family of priors. The conjugate prior for this family is the Beta prior.

(2.7.5) $$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \sim \text{Beta}(\alpha, \beta)$$

This prior is reasonably flexible with regard to location, $E[\theta] = \alpha/(\alpha+\beta)$, spread, and shape (either uni-modal or "u-shaped"). We can interpret this prior as the posterior from another sample of $\alpha + \beta - 2$ observations and $\alpha - 1$ values of "1." The posterior is also of the Beta form

(2.7.6) $$p(\theta|y) \propto \theta^{\alpha+\sum_i y_i - 1}(1-\theta)^{\beta+n-\sum_i y_i - 1} \sim \text{Beta}(\alpha', \beta')$$

$\alpha' = \alpha + \sum_i y_i$; $\beta' = \beta + n - \sum_i y_i$. Thus, we can find the posterior moments from the Beta distribution.

Those readers who are familiar with numerical integration methods might regard this example as trivial and not very interesting since one could simply compute whatever posterior integrals are required by univariate numerical integration. This would allow for the use of any reasonable prior. However, the ideas of natural conjugate priors are most powerful when applied to vectors of regression parameters and covariance matrices which we develop in the next section.

2.8 Regression and Multivariate Analysis Examples

2.8.1 Multiple Regression

The regression problem has received a lot of attention in the Bayesian statistics literature and provides a very good and useful example of non-trivial natural conjugate priors. The standard linear regression model is a model for the conditional distribution of y given a vector of predictor variables in x.

(2.8.1) $$y_i = x_i'\beta + \varepsilon_i \qquad \varepsilon_i \sim \text{iid}N(0, \sigma^2)$$

or

(2.8.2) $$y \sim N(X\beta, \sigma^2 I_n)$$

$N(\mu, \Sigma)$ is the standard notation for a multivariate distribution with mean μ and variance-covariance matrix Σ.

We have only modeled the conditional distribution of y given x rather than the joint distribution. In non-experimental situations, it can be argued that we need to choose a

model for the joint distribution of both X and y. In order to complete the model, we would need a model for the marginal distribution of x.

(2.8.3) $$p(x,y) = p(x|\psi)p\left(y\middle|x,\beta,\sigma^2\right)$$

If ψ is a priori independent[4] of (β,σ^2), then the posterior factors into two terms, the posterior of the x marginal parameters and the posterior for the regression parameters and we can simply focus on the rightmost term of (2.8.4).

(2.8.4) $$p\left(\psi,\beta,\sigma^2\middle|y,X\right) \propto \left[p(\psi)\prod_i p(x_i|\psi)\right]\left[p\left(\beta,\sigma^2\right)p\left(y\middle|X,\beta,\sigma^2\right)\right]$$

What sort of prior should be used for the regression model parameters? There are many possible choices for family of prior distributions and, given the choice of family, there is also the problem of assessing the prior parameters. Given that the regression likelihood is a member of the exponential family, a reasonable starting place would be to consider natural conjugate priors. The form of the natural conjugate prior for the regression model can be seen by examining the likelihood function.

To start, let us review the likelihood function for the normal linear regression model. We start from the distribution of the error terms which are multivariate normal. If $x \sim N(\mu,\Sigma)$, then the density of x is given by

(2.8.5) $$p(x|\mu,\Sigma) = (2\pi)^{-k/2}|\Sigma|^{-1/2}\exp\left(-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

$\varepsilon \sim N\left(0,\sigma^2 I_n\right)$. The density of ε is then

(2.8.6) $$p\left(\varepsilon\middle|\sigma^2\right) = (2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2}\exp\left(-\tfrac{1}{2\sigma^2}\varepsilon'\varepsilon\right)$$

Given that the Jacobian from ε to y is 1 and we are conditioning on X, we can write the density of y given X, β and σ^2 as

$$p\left(y\middle|X,\beta,\sigma^2\right) \propto \left(\sigma^2\right)^{-n/2} \exp\left(-\tfrac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right) \tag{2.8.7}$$

The natural conjugate prior is a joint density for β and σ^2 which is proportional to the likelihood. Given that the likelihood has a quadratic form in β in the exponent, a reasonable guess is that the conjugate prior for β is normal. Rewrite the exponent so that the quadratic form is of the usual normal form. This can be done either by expanding out the existing expression and completing the square in β, or by recalling the usual trick of projecting y on the space spanned by the columns of X, $y = \hat{y} + e = X\hat{\beta} + \left(y - X\hat{\beta}\right)$ and $\hat{y}'e = 0$.

$$\begin{aligned}(y-X\beta)'(y-X\beta) &= \left(y-X\hat{\beta}\right)'\left(y-X\hat{\beta}\right) + \left(\beta-\hat{\beta}\right)'X'X\left(\beta-\hat{\beta}\right)\\ &= \upsilon s^2 + \left(\beta-\hat{\beta}\right)'X'X\left(\beta-\hat{\beta}\right)\end{aligned}$$

$\upsilon s^2 = \left(y-X\hat{\beta}\right)'\left(y-X\hat{\beta}\right)$, $\upsilon = n-k$, and $\hat{\beta} = (X'X)^{-1}X'y$. If we substitute this expression in the exponent, we can factor the likelihood into two components keeping the terms needed for a proper normal density for β and then the balance for σ

$$\begin{aligned}p\left(y\middle|X,\beta,\sigma^2\right) \propto \left(\sigma^2\right)^{-\upsilon/2} &\exp\left\{-\frac{\upsilon s^2}{2\sigma^2}\right\}\times\\ &\left(\sigma^2\right)^{-(n-\upsilon)/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\beta-\hat{\beta}\right)'(X'X)\left(\beta-\hat{\beta}\right)\right\}\end{aligned} \tag{2.8.8}$$

---

[4] This rules out deterministic relationships between ψ and $\left(\beta,\sigma^2\right)$ as well as stochastic dependence. In situations in which the x variables are set strategically by the marketer, this assumption can be violated. We explore this further in Chapter 7.

The first term in (2.8.8) suggests a form for the density[5] of σ^2 and the second for the density of β. We note that the normal density for β involves σ^2. This means that the conjugate prior will be specified as

(2.8.9) $$p\left(\beta,\sigma^2\right)=p\left(\sigma^2\right)p\left(\beta\middle|\sigma^2\right)$$

The first term in (2.8.8) suggests that the marginal prior on σ^2 has a density of the form $p(\theta)\propto\theta^{-\lambda}\exp\left(-\frac{\delta}{\theta}\right)$. This turns out to be in the form of the inverse-gamma distribution. The standard form of the inverse-gamma is

(2.8.10)
$$\theta\sim\text{Inv-Gamma}(\alpha,\beta)$$
$$p(\theta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{-(\alpha+1)}\exp\left(-\frac{\beta}{\theta}\right)$$

The natural conjugate priors for σ^2 is of the form:

(2.8.11) $$p\left(\sigma^2\right)\propto\left(\sigma^2\right)^{-(\upsilon_0/2+1)}\exp\left\{-\frac{\upsilon_0 s_0^2}{2\sigma^2}\right\}$$

The standard Inv-Gamma form with $\alpha=\upsilon_0/2$ and $\beta=\upsilon_0 s_0^2/2$. We note that Inv-Gamma density requires a slightly different power for σ^2 than found in the likelihood. There is an extra, $\left(\sigma^2\right)^{-1}$ in (2.8.11) which is not suggested by the form of the likelihood in (2.8.8). This term can be rationalized by viewing the conjugate prior as arising from the posterior of a

[5] Note that we can speak interchangeably about σ or $\theta=\sigma^2$. However, we must remember the Jacobian of the transformation from σ to θ when we re-write the density expressions converting from σ to θ (the Jacobian is 2σ).

sample of size υ_0 with sufficient statistics, $s_0^2, \bar{\beta}$, formed with the non-informative prior, $p\left(\beta,\sigma^2\right) \propto \frac{1}{\sigma^2}$.

The conjugate normal prior on β is given by:

$$p\left(\beta\middle|\sigma^2\right) \propto \left(\sigma^2\right)^{-k} \exp\left\{-\frac{1}{2\sigma^2}\left(\beta-\bar{\beta}\right)' A\left(\beta-\bar{\beta}\right)\right\} \tag{2.8.12}$$

(2.8.11) and (2.8.12) can be expressed in terms of common distributions using the relationship between the Inv-Gamma and the Inverse of a Chi-squared random variable.

$$\sigma^2 \sim \frac{\upsilon_0 s_0^2}{\chi^2_{\upsilon_0}} \quad \text{and} \quad \beta\middle|\sigma^2 \sim N\left(\bar{\beta}, \sigma^2 A^{-1}\right) \tag{2.8.13}$$

The notation υ_0, s_0^2 is suggestive of the interpretation of the natural conjugate prior as based on another sample with υ_0 degrees of freedom and sum of squared errors, $\upsilon_0 s_0^2$.

Given the natural conjugate priors for the regression problem (2.8.13), the posterior is in the same form.

$$\begin{aligned} p\left(\beta,\sigma|y,X\right) \propto p\left(y\middle|X,\beta,\sigma^2\right)p\left(\beta\middle|\sigma^2\right)p\left(\sigma^2\right) \propto \\ \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(y-X\beta\right)'\left(y-X\beta\right)\right\} \times \\ \left(\sigma^2\right)^{-k/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\beta-\bar{\beta}\right)' A\left(\beta-\bar{\beta}\right)\right\} \times \\ \left(\sigma^2\right)^{-(\upsilon_0/2+1)} \exp\left\{-\frac{\upsilon_0 s_0^2}{2\sigma^2}\right\} \end{aligned} \tag{2.8.14}$$

The quadratic forms in the first two exponents on the right hand side can be combined by "completing the square." To remove some of the mystery from this operation which appears to require that you know the answer prior to obtaining the answer, we rewrite the

problem as the standard problem of decomposing a vector into its projection on a subspace spanned by some column vectors and an orthogonal "residual." Since A is positive definite, we can find the upper triangular Cholesky root, U. A=U'U.

$$(y - X\beta)'(y - X\beta) + (\beta - \bar{\beta})' U'U(\beta - \bar{\beta}) = (v - W\beta)'(v - W\beta)$$

where $v = \begin{bmatrix} y \\ U\bar{\beta} \end{bmatrix}$ $W = \begin{bmatrix} X \\ U \end{bmatrix}$. We now project v onto the space spanned by the W columns using standard least squares.

$$(v - W\beta)'(v - W\beta) = ns^2 + (\beta - \tilde{\beta})' W'W(\beta - \tilde{\beta})$$

where

$$\tilde{\beta} = (W'W)^{-1} W'v = (X'X + A)^{-1}(X'X\hat{\beta} + A\bar{\beta}) \tag{2.8.15}$$

and

$$ns^2 = (v - W\tilde{\beta})'(v - W\tilde{\beta}) = (y - X\tilde{\beta})'(y - X\tilde{\beta}) + (\tilde{\beta} - \bar{\beta})' A(\tilde{\beta} - \bar{\beta}). \tag{2.8.16}$$

Using these results, we can write (2.8.14).

$$\begin{aligned} p(\beta, \sigma^2 | y, X) \propto (\sigma^2)^{-k/2} \exp\left\{ -\frac{1}{2\sigma^2}(\beta - \tilde{\beta})'(X'X + A)(\beta - \tilde{\beta}) \right\} \times \\ (\sigma^2)^{-((n+\upsilon_0)/2+1)} \exp\left\{ -\frac{(\upsilon_0 s_0^2 + ns^2)}{2\sigma^2} \right\} \end{aligned} \tag{2.8.17}$$

or

$$\begin{aligned} \beta | \sigma^2, y &\sim N(\tilde{\beta}, \sigma^2 (X'X + A)^{-1}) \\ \sigma^2 | y &\sim \frac{\upsilon_1 s_1^2}{\chi^2_{\upsilon_1}} \text{ with } \upsilon_1 = \upsilon_0 + n;\ s_1^2 = \frac{\upsilon_0 s_0^2 + ns^2}{\upsilon_0 + n} \end{aligned} \tag{2.8.18}$$

The Bayes estimator corresponding to the posterior mean is $\tilde{\beta}$.

$$E\left[\beta|y\right] = E_{\sigma^2|y}\left[E_{\beta|\sigma^2,y}\left[\beta\right]\right] = E_{\sigma^2|y}\left[\tilde{\beta}\right] = \tilde{\beta}$$

$\tilde{\beta}$ is a weighted average of the prior mean and the least squares estimator, $\hat{\beta}$. The weights depend on the prior precision and the sample information (recall that the information matrix for all n observations in the regression model is proportional to X'X). There are two important practical aspects of this estimator:

i. The Bayes estimator is a "shrinkage" estimator in the sense that the least squares estimator is "shrunk" toward the prior mean. Similarly, the posterior distribution of σ^2 is "centered" over s_1^2 which is a weighted average of the prior parameter and a sample quantity (however, the "sample" sum of squares includes a term (see (2.8.16)) which represents the degree to which the prior mean differs from the least square estimator).

ii. As we acquire more sample information in the sense of total X variation, the Bayes estimator converges to the least squares estimator (this insures consistency which we said is true, in general, for Bayes estimators).

These results for the posterior are special to the linear regression model and the natural conjugate prior set-up but if the likelihood function is approximately normal with mean equal to the MLE then these results suggest that the Bayes estimator will be a shrinkage estimator which is a weighted average of the MLE and prior mean.

While the posterior mean of β is available by direct inspection of the conditional distribution of β given σ^2, the marginal distribution of β must be computed. It turns out that the marginal distribution of β is in the multivariate t form. The multivariate t

distribution is in the elliptical class of distributions[6] and the marginal distribution of β will have the same density contours as the conditional distribution but will have fatter, algebraic tails rather than the thin exponential tapering normal tails. Moments of the multivariate t distribution are available[7]. However, the results in (2.8.18) can be used to devise a simple simulation strategy for making iid draws from the posterior of β. Simply draw a value of σ^2 from its marginal posterior, insert this value into the expression for the covariance matrix of the conditional normal distribution of $\beta|\sigma^2, y$ and draw from this multivariate normal.[8]

(2.8.18) illustrates one of the distinguishing characteristics of natural conjugate priors. The posterior is centered between the prior and the likelihood and is more concentrated than either. Thus, it appears that you always gain by combining prior and sample information. It should be emphasized that this is a special property of natural conjugate priors. The property comes, in essence, from the interpretation of the prior as the posterior from another sample of data. The posterior we calculate can be interpreted as the same as the posterior one would obtain by pooling both the prior "sample" and the actual sample and using a very diffuse prior. If the prior information is at odds with the sample information, the use of natural conjugate priors will gloss this over because of the implicit assumption that both prior and sample information are equally valid. Many investigators would like to be aware of the situation in which the prior information is very different from the sample information. It is often difficult to check on this if the parameter

---

[6] Note that linear combinations of independent t distributions are not in the multivariate t form.

[7] $\mathrm{Var}(\beta) = \frac{\upsilon_1}{\upsilon_1 - 2} s_1^2 (X'X + A)^{-1}$

[8] Note: you will simply multiply the appropriate draws from a $N\left(0, (X'X + A)^{-1}\right)$ by the square root of this draw, so that this will be computationally nearly the same as multivariate normal draws. In section 2.11, we will discuss some of the details for simulation from the distributions required for natural conjugate Bayes models.

space is high dimensional. This is particularly a problem in hierarchical models in which natural conjugate results like the ones given here are used. At this point, we call attention to this as a potential weakness of natural conjugate priors. After we develop computational methods that can work with non-conjugate priors, we will return to this issue.

2.8.2 Assessing Priors for Regression Models

(2.8.13) is one implementation of a particular view regarding prior information on the regression coefficients. The idea here is that our views about coefficients are dependent on the error variance or equivalently the scale of the regression error terms (σ). If σ is large, the prior on the regression coefficients spreads out to accommodate larger values or reflects greater uncertainty. This is very much driven by the view that prior information comes from data. In situations where the "sample" creating the prior information comes from a regression with highly variable error terms, this sample information will be less valuable and the prior should spread out. However, it is entirely possible to imagine prior information that directly addresses the size of the regression coefficients. After all, the regression coefficients are designed to measure the effects of changes in the x variables. We may have information about the size of effects of a one unit price change on sales, for example, that is independent of the percentage of variation in sales explained by price. For this reason, the natural conjugate prior may not be appropriate for some sorts of information.

The more serious practical problem with the prior in (2.8.13) is the large number of prior hyper-parameters that must be assessed, 2 for prior distribution of σ, and $k + (k(k+1))/2$ for the conditional prior on β. Moreover, as the likelihood is of order n, we may find that for any standard regression dataset with a modest number of observations, the hard work of prior assessment will have a low return as the prior may make little difference.

Obviously, the prior will only matter with a small amount of sample information which typically arises in situations with small datasets or lack of independent variation in the x variables. The full Bayesian machinery is rarely used directly on one dataset regression problems but is increasingly being used with sets of related regression equations. We will develop this further in Chapter 5 when we discuss hierarchical models.

One approach to the problem of assessment of the natural conjugate prior is to use prior settings that are very diffuse. This would involve a "large" value of s_0^2, "small" value of υ_0, and a "small" value of A, the prior precision. If the prior is diffuse, the mean of β, $\bar{\beta}$, is not very critical and can be set to zero. What constitutes a "diffiuse" or spread out prior is a relative statement – relative to the diffusion of the likelihood. We can be sure that the prior has "small" influence if we make the prior diffuse relative to the likelihood. If υ_0 = some small fraction of n like max(3, .01 n) and

$$A = \upsilon_0 S_X \text{ where } S_X = \text{diag}\left(s_1^2, \ldots, s_k^2\right) \text{ and } s_j^2 = 1/(n-1)\sum_i \left(x_{ij} - \bar{x}_j\right)^2 \tag{2.8.19}$$

then we have assessed a relatively diffuse prior. We have introduced the scaling of x into the prior as would seem reasonable. If we change the units of X, we might argue that our prior information should remain the same and this sort of prior is invariant to scale changes. The g-prior of Zellner (1986) takes this idea further and sets $A = gX'X$. This means that we are using not only the scale of the x variables but their observed correlation structure as well. There is some controversy as to whether this is "coherent" and/or desirable but for relatively diffuse priors, the difference between the prior given by (2.8.19) and the g-prior will be minimal.

Our scheme for assessing a relatively diffuse prior is not complete without method for assessing the value of s_0^2. s_0^2 determines the location of the prior in the sense that mean of σ^2 is

$$(2.8.20) \quad E\left[\sigma^2\right] = \frac{\upsilon_0 s_0^2}{\upsilon_0 - 2}, \text{ for } \upsilon_0 > 2; \quad \text{Var}\left(\sigma^2\right) = \frac{2\upsilon_0^2}{\left(\upsilon_0 - 2\right)^2\left(\upsilon_0 - 4\right)}\left(s_0^2\right)^2, \text{ for } \upsilon_0 > 4$$

As υ_0 increases, the prior becomes tighter and centered around s_0^2. Many investigators are rather cavalier about the value of s_0^2 and set it to some rather arbitrary number like 1. If the prior is barely proper then the argument is that there is sufficient prior mass over very large values of σ and this is all that is relevant. However, these prior settings often mean that little or no prior mass is put on small values of σ below .5 or so. Depending on the scale of y and the explanatory power of the X variables, this can be a very informative prior! A somewhat more reasonable, but still controversial view, is to take into account of the scale of y in assessing the value of s_0^2. For example, we could use the sample variance of y for s_0^2. At the most extreme, we could use the mean error sum of squares from the regression of y on X. Purists would argue that this is violating Bayes theorem in that the data is being "used" twice – once in the assessment of the prior and again in the likelihood. In the absence of true prior information on σ, it seems preferable to use some scale information on y rather than to put in an arbitrary value for s_0^2 and hope that the prior is diffuse enough so this doesn't affect the posterior inferences in an undesirable manner.

An alternative to assessment of an informative prior would be to use one of the many candidates for "non-informative" priors. There are many possible priors, each derived from different principles. For the case of continuous-valued parameters, it is

important to understand that there can be no such thing as an uninformative prior. For example, if we have a uniform but improper[9] prior on a uni-dimensional parameter θ, $p(\theta) \propto$ constant, then $\tau = e^{\theta}$ has a non-uniform density $1/\tau$. Various invariance principles have been proposed in which priors are formulated under the constraint that they are invariant to certain types of transformations. In many situations, prior information should not be invariant to transformations. Our view is that prior information is available and should be expressed through proper priors assessed in the parameterization which yields maximum interpretability. Moreover, the use of proper priors avoids mathematical pathologies and inadvertent rendering of extreme prior information. However, for the sake of completeness, we present the standard non-informative prior for the regression model.

(2.8.21) $$p(\beta, \sigma) = p(\beta)p(\sigma) \propto \frac{1}{\sigma}$$

Or

(2.8.22) $$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

In contrast to the natural conjugate prior, β is independent of σ. The uniform distribution of β over all of R^k means that the prior in (2.8.21) implies that we think β is large with very high prior probability. In our view, this exposes the absurdity of this prior. We surely have prior information that the regression coefficients are not arbitrarily large! The prior on σ^2 can be motivated by appeal to the notion of scale invariance – that is the prior should be unchanged when y is multiplied by any positive constant.

[9] Improper means non-integrable, i.e. the integral $\int p(\theta) d\theta$ diverges.

2.8.3 Bayesian Inference for Covariance Matrices

A key building block in many Bayesian models is some sort of covariance structure. The workhorse model in this area is a multivariate normal sample and the associated natural conjugate prior. Consider the likelihood for a random sample from an m-dimensional multivariate normal (for expositional purposes, we omit the mean vector; in section 2.8.5, we will consider the general case where the mean vector is a set of regression equations).

$$(2.8.23)\quad p\left(y_1,\ldots,y_n\middle|\Sigma\right)\propto\prod_{i=1}^{n}|\Sigma|^{-1/2}\exp\left\{-\tfrac{1}{2}y_i'\Sigma^{-1}y_i\right\}=|\Sigma|^{-n/2}\exp\left\{-\tfrac{1}{2}\sum y_i'\Sigma^{-1}y_i\right\}$$

We can rewrite the exponent of (2.8.23) in a more compact manner using the trace operator.

$$\sum tr\left(y_i'\Sigma^{-1}y_i\right)=\sum tr\left(y_iy_i'\Sigma^{-1}\right)=tr\left(S\Sigma^{-1}\right)$$

with $S=\sum y_iy_i'$. Using the notation, $etr(\ )\equiv\exp\left(tr(\ )\right)$.

$$(2.8.24)\qquad p\left(y_1,\ldots,y_n\middle|\Sigma\right)\propto|\Sigma|^{-n/2}etr\left\{-\tfrac{1}{2}S\Sigma^{-1}\right\}$$

(2.8.24) suggests that a natural conjugate prior for Σ is of the form

$$(2.8.25)\qquad p\left(\Sigma\middle|\upsilon_0,V_0\right)\propto|\Sigma|^{-(\upsilon_0+m+1)/2}etr\left(-\tfrac{1}{2}V_0\Sigma^{-1}\right)$$

with $\upsilon_0 > m$ required for an integrable density.

(2.8.25) is the expression for an Inverted Wishart density. We use the notation

$$(2.8.26)\qquad \Sigma\sim IW\left(\upsilon_0,V_0\right)$$

If $\upsilon_0\geq m+2$, then $E[\Sigma]=(\upsilon_0-m-1)^{-1}V_0$. The full density function for the IW distribution is given by

$$(2.8.27)\qquad \begin{aligned}p\left(\Sigma\middle|\upsilon_0,V_0\right)=&\left(2^{\upsilon_0m/2}\pi^{m(m-1)/4}\prod_{i=1}^{m}\Gamma\left(\frac{\upsilon_0-1-i}{2}\right)\right)^{-1}\\&\times\left|V_0\right|^{\upsilon_0/2}|\Sigma|^{-(\upsilon_0+m+1)/2}etr\left(-\tfrac{1}{2}V_0\Sigma^{-1}\right)\end{aligned}$$

This implies that Σ^{-1} has a Wishart distribution. $\Sigma^{-1} \sim W\left(\upsilon_0, V_0^{-1}\right)$. $E\left[\Sigma^{-1}\right] = \upsilon_0 V_0^{-1}$. We can interpret V_0 as determining the "location" of the prior and ν_0 as determining the spread of the distribution. However, some caution should be exercised in interpreting V_0 as a location parameter, particularly for small values of ν_0. The IW is a highly skewed distribution which can be thought of the matrix-valued generalization of the inverted chi-squared prior for the variance in the single parameter case.[10] As with all highly skewed distributions, there is a close relationship between the spread and the location. As we increase V_0 for small ν_0, then we also increase the spread of the distribution dramatically.

The Wishart and Inverted Wishart distributions have a number of additional drawbacks. The most important drawback is that the Wishart has only one tightness parameter. This means that we can't be very informative on some elements of the covariance matrix and less informative on others. We would have to use independent Inverted Wisharts or some more richly parameterized distribution to handle this situation. In some situations, we wish to condition on some elements of the covariance matrix and use a conditional prior. Unfortunately, the conditional distribution of a portion of the matrix given some other elements is not in the Inverted Wishart form.

As is usual with natural conjugate priors, we can interpret the prior as the posterior stemming from some other data set and with a diffuse prior. If we use the diffuse prior suggested by Zellner via the Jeffreys' invariance principle Zellner (1971), p. 225, $p(\Sigma) \propto |\Sigma|^{-(m+1)/2}$, then we can interpret υ_0 as the effective sample size of the sample

[10] IW distributions have the property that the marginal distribution of a square block along the diagonal is also IW with the same degrees of freedom parameter and the appropriate submatix of V_0. Therefore, an IW distribution with one degree of freedom is proportional to the inverse of a chi-squared variate.

underlying the prior and V_0 as the sum of squares and cross products matrix from this sample. This interpretation helps to assess the prior hyper-parameters. Some assess V_0 by appeal to the sum of squares interpretation. Simply scale up some sort of prior notation of the variance-covariance matrix by ν_0, $V_0 = \nu_0 \hat{\Sigma}$. A form of "cheating" would be to use the exact or a stylized version of covariance matrix of $\{y_1, \ldots, y_n\}$. In situations where a relatively diffuse prior is desired, many investigators use $\nu_0 = m + 3;\ V_0 = \nu_0 I$. This is an extremely spread out prior which should be used with some caution.

If we combine the likelihood in (2.8.24) with the conjugate prior (2.8.25), then we have a posterior in the IW form:

$$\begin{aligned} p(\Sigma|Y) &\propto p(\Sigma)p(Y|\Sigma) \\ &= |\Sigma|^{-(\nu_0+m+1)/2} \operatorname{etr}\left(-\tfrac{1}{2} V_0 \Sigma^{-1}\right) |\Sigma|^{-n/2} \operatorname{etr}\left(-\tfrac{1}{2} S \Sigma^{-1}\right) \\ &= |\Sigma|^{-(\nu_0+m+n+1)/2} \operatorname{etr}\left(-\tfrac{1}{2}(V_0 + S)\Sigma^{-1}\right) \end{aligned} \tag{2.8.28}$$

or

$$p(\Sigma|Y) \sim IW(\nu_0 + n, V_0 + S) \tag{2.8.29}$$

with mean $E[\Sigma] = \dfrac{V_0 + S}{\nu_0 + n - m - 1}$ which converges to the MLE as n approaches infinity.

2.8.4 Priors and the Wishart Distribution

Although it seems more natural and interpretable to place a prior directly on Σ, there is a tradition in some parts of the Bayesian literature of choosing the prior on Σ^{-1} instead. In the Σ^{-1} parameterization, the prior (2.8.26) is a Wishart distribution.

$$p\left(G = \Sigma^{-1}\right) \propto p(\Sigma|\upsilon_0, V_0)\Big|_{\Sigma = G^{-1}} \times J_{\Sigma \to G}$$

or

$$
\begin{aligned}
p(G) &\propto |G|^{(\nu_0+m+1)/2}\,\mathrm{etr}\left(-\tfrac{1}{2}V_0G\right)\times|G|^{-(m+1)} \\
&= |G|^{(\nu_0-m-1)/2}\,\mathrm{etr}\left(-\tfrac{1}{2}A_0^{-1}G\right)
\end{aligned}
\tag{2.8.30}
$$

where $A_0 = V_0^{-1}$. This is denoted $G \sim W(\nu_0, A_0)$. If $\nu_0 \geq m+2$, then $E[G] = \nu_0 A_0$.

The real value of the Wishart parameterization comes from its use to construct simulators. This comes from a result in multivariate analysis (c.f. Muirhead, theorem 3.2.1, p. 85). If we have a set of iid random vectors $\{e_1,\ldots,e_\nu\}$ and $e_i \sim N(0,I)$, then

$$
E = \sum_{i=1}^{\nu} e_i e_i' \sim W(\nu, I) \tag{2.8.31}
$$

if $\Sigma = U'U$, then

$$
W = U'EU \sim W(\nu, \Sigma) \tag{2.8.32}
$$

We note that we do not use (2.8.32) directly to simulate Wishart random matrices. In section 2.11, we will provide an algorithm for simulating from Wisharts.

2.8.5 Multivariate Regression

One useful way to view the multivariate regression model is as a set of regression equations related through common X variables and correlated errors.

$$
\begin{aligned}
y_1 &= X\beta_1 + \varepsilon_1 \\
&\vdots \\
y_c &= X\beta_c + \varepsilon_c \\
&\vdots \\
y_m &= X\beta_m + \varepsilon_m
\end{aligned}
\tag{2.8.33}
$$

The subscript "c" denotes a vector of n observations on equation c; we use the subscript "c" to suggest the columns of a matrix which will be developed later. The model in (2.8.33) is not complete without a specification of the error structure. The standard multivariate regression model specifies that the errors for observation c are correlated across equations.

Coupled with a normal error assumption, we can view the model as a direct generalization of the multivariate normal inference problem discussed in section 2.8.3. To see this, we will think of a row vector of one observation on each of the m regression equations. This row vector will have a multivariate normal distribution with means given by the appropriate regression equation.

(2.8.34)
$$\begin{aligned} y_r &= B' x_r + \varepsilon_r \\ \varepsilon_r &\sim \text{iidN}(0, \Sigma) \end{aligned}$$

The "r" subscript refers to observation r and B is a $k \times m$ matrix whose columns are the regression coefficients in (2.8.33). y_r and ε_r are m vectors of the observations on each of the dependent variables and error term.

(2.8.34) is a convenient way of expressing the model for the purpose of writing down the likelihood function for the model.

(2.8.35)
$$\begin{aligned} p(\varepsilon_1, \ldots, \varepsilon_n | \Sigma) &\propto |\Sigma|^{-n/2} \exp\left\{-\tfrac{1}{2} \text{tr} S_\varepsilon \Sigma^{-1}\right\} \\ S_\varepsilon &= \sum_{r=1}^{n} \varepsilon_r \varepsilon_r' \end{aligned}$$

Since the Jacobian from ε to y is 1, we can simply substitute in to obtain the distribution of the observed data. Clearly, an IW prior will be the conjugate prior for Σ. To obtain the form for the natural conjugate prior for the regression coefficients, it will be helpful to write down the likelihood function using the matrix form of the multivariate regression model.

(2.8.36)
$$\begin{aligned} Y &= XB + E \\ B &= [\beta_1, \ldots, \beta_c, \ldots, \beta_m] \end{aligned}$$

Both Y and E are $n \times m$ matrices of observations whose (i,j) elements are the ith observation on equation j. X is a $n \times k$ matrix of observations on k common independent variables. The columns of Y and E are $\{y_c\}$ and $\{\varepsilon_c\}$, respectively given in (2.8.33). The

rows of the E matrix are the $\{\varepsilon_r\}$ given in (2.8.35). Observing that $E'E = S_\varepsilon$, we can use (2.8.35) and (2.8.36) to write down the complete likelihood.

$$p(E|\Sigma) \propto |\Sigma|^{-n/2} \operatorname{etr}\left\{-\tfrac{1}{2}E'E\Sigma^{-1}\right\}$$
$$p(Y|X,B,\Sigma) \propto |\Sigma|^{-n/2} \operatorname{etr}\left\{-\tfrac{1}{2}(Y-XB)'(Y-XB)\Sigma^{-1}\right\}$$

Again, we can decompose the sum of squares using the least squares projection.

$$p(Y|X,B,\Sigma) \propto |\Sigma|^{-n/2} \operatorname{etr}\left\{-\tfrac{1}{2}\left(S+\left(B-\hat{B}\right)'X'X\left(B-\hat{B}\right)\right)\Sigma^{-1}\right\}$$

with $S = \left(Y-X\hat{B}\right)'\left(Y-X\hat{B}\right)$ and $\hat{B} = \left(X'X\right)^{-1}X'Y$. To suggest the form of the natural conjugate prior, we can break up the two terms in the exponent.

$$p(Y|X,B,\Sigma) \propto |\Sigma|^{-(n-k)/2} \operatorname{etr}\left\{-\tfrac{1}{2}S\Sigma^{-1}\right\} \times |\Sigma|^{-k/2} \operatorname{etr}\left\{-\tfrac{1}{2}\left(B-\hat{B}\right)'X'X\left(B-\hat{B}\right)\Sigma^{-1}\right\} \quad (2.8.37)$$

(2.8.37) suggests that the natural conjugate prior is a IW on Σ and a prior on B which is conditional on Σ. The term involving B is a density expressed as a function of an arbitrary $k \times m$ matrix. We can convert this density expression from a function of B to a function of $\beta = \text{vec}(B)$ using standard results on vec operators[11] (see Magnus and Neudecker p. 30).

$$\operatorname{tr}\left(\left(B-\hat{B}\right)'X'X\left(B-\hat{B}\right)\Sigma^{-1}\right) = \operatorname{vec}\left(B-\hat{B}\right)'\operatorname{vec}\left(X'X\left(B-\hat{B}\right)\Sigma^{-1}\right)$$

and

$$\operatorname{vec}\left(X'X\left(B-\hat{B}\right)\Sigma^{-1}\right) = \left(\Sigma^{-1}\otimes X'X\right)\operatorname{vec}\left(B-\hat{B}\right)$$

Thus,

[11] $\operatorname{tr}(A'B) = (\operatorname{vec}(A))'\operatorname{vec}(B)$ and $\operatorname{vec}(ABC) = (C'\otimes A)\operatorname{vec}(B)$

$$\operatorname{tr}\left(\left(B-\hat{B}\right)' X'X\left(B-\hat{B}\right)\Sigma^{-1}\right) = \operatorname{vec}\left(B-\hat{B}\right)'\left(\Sigma^{-1}\otimes X'X\right)\operatorname{vec}\left(B-\hat{B}\right)$$
$$= \left(\beta-\hat{\beta}\right)'\left(\Sigma^{-1}\otimes X'X\right)\left(\beta-\hat{\beta}\right)$$

Thus, the second term on the rhs of (2.8.37) is a normal kernel. This means that the natural conjugate prior for β is a normal prior conditional on this specific covariance matrix which depends on Σ.

The natural conjugate priors for the multivariate regression model are of the form.

$$\begin{aligned} &p(\Sigma, B) = p(\Sigma)\,p(B|\Sigma) \\ &\Sigma \sim IW(\nu_0, V_0) \\ &\beta|\Sigma \sim N\left(\bar{\beta}, \Sigma\otimes A^{-1}\right) \end{aligned} \tag{2.8.38}$$

Just as in the univariate regression model, the prior on the regression coefficients is dependent on the scale parameters and the same discussion applies. If we destroy natural conjugacy by using independent priors on β and Σ, we will not have analytic expressions for the posterior moments. The posterior can be obtained by combining terms from the natural conjugate prior to produce a posterior which is a product of an IW and a "matrix" normal kernel.

$$\begin{aligned} p(\Sigma, B|Y, X) \propto\ & |\Sigma|^{-(\upsilon_0+m+1)/2}\operatorname{etr}\left(-\tfrac{1}{2}V_0\Sigma^{-1}\right) \\ &\times|\Sigma|^{-k/2}\operatorname{etr}\left(-\tfrac{1}{2}\left(B-\bar{B}\right)'A\left(B-\bar{B}\right)\Sigma^{-1}\right) \\ &\times|\Sigma|^{-n/2}\operatorname{etr}\left(-\tfrac{1}{2}(Y-XB)'(Y-XB)\Sigma^{-1}\right) \end{aligned} \tag{2.8.39}$$

We can combine the two terms involving B, using the same device we used for the univariate regression model.

$$
\begin{aligned}
&\left(B-\bar{B}\right)'A\left(B-\bar{B}\right)+\left(Y-XB\right)'\left(Y-XB\right)\\
&=\left(Z-WB\right)'\left(Z-WB\right)\\
&=\left(Z-W\tilde{B}\right)'\left(Z-W\tilde{B}\right)+\left(B-\tilde{B}\right)'W'W\left(B-\tilde{B}\right)
\end{aligned}
\tag{2.8.40}
$$

with $W=\begin{bmatrix}X\\U\end{bmatrix};\ Z=\begin{bmatrix}Y\\U\bar{B}\end{bmatrix};\quad A=U'U$

The posterior density can now be written

$$
\begin{aligned}
&p\left(\Sigma,B\middle|Y,X\right)\propto\left|\Sigma\right|^{-\left(\upsilon_0+n+m+1\right)/2}\operatorname{etr}\left(-\tfrac{1}{2}\left(V_0+\left(V-W\tilde{B}\right)'(\bullet)\Sigma^{-1}\right)\right)\\
&\times\left|\Sigma\right|^{-k/2}\operatorname{etr}\left(-\tfrac{1}{2}\left(B-\tilde{B}\right)'W'W\left(B-\tilde{B}\right)\right)
\end{aligned}
\tag{2.8.41}
$$

with

$$
\begin{aligned}
&\tilde{B}=\left(X'X+A\right)^{-1}\left(X'X\hat{B}+A\bar{B}\right)\\
&\left(Z-W\tilde{B}\right)'\left(Z-W\tilde{B}\right)=\left(Y-X\tilde{B}\right)'\left(Y-X\tilde{B}\right)+\left(\tilde{B}-\bar{B}\right)'A\left(\tilde{B}-\bar{B}\right)
\end{aligned}
\tag{2.8.42}
$$

Thus, the posterior is in the form the conjugate prior: IW x conditional normal.

$$
\begin{aligned}
&\Sigma\middle|Y,X\sim IW\left(\nu_0+n,V_0+S\right)\\
&\beta\middle|Y,X,\Sigma\sim N\left(\tilde{\beta},\Sigma\otimes\left(X'X+A\right)^{-1}\right)\\
&\tilde{\beta}=\operatorname{vec}\left(\tilde{B}\right);\ \tilde{B}=\left(X'X+A\right)^{-1}\left(X'X\hat{B}+A\bar{B}\right);\\
&S=\left(Y-X\tilde{B}\right)'\left(Y-X\tilde{B}\right)+\left(\tilde{B}-\bar{B}\right)'A\left(\tilde{B}-\bar{B}\right)
\end{aligned}
\tag{2.8.43}
$$

2.8.6 The Limitations of Conjugate Priors

Up to this point, we have considered some standard problems for which there exist natural conjugate priors. Although the natural conjugate priors have some features which might not always be very desirable, convenience is a powerful argument. However, the set of problems for which there exist useable expressions for conjugate priors is very small as was pointed out in section 2. In this section, we will illustrate that a seemingly minor change in the multivariate regression model destroys conjugacy. We will also examine the logistic

regression model (one of the most widely used models in marketing) and see that there are no nice conjugate priors there. One should not conclude that conjugate priors are useless. Conjugate priors such as the normal and Wishart are simple useful representations of prior information which can be used even in non-conjugate contexts. Finally, many models are *conditionally conjugate* in the sense that conditional on some subset of parameters, we can use these conjugate results. This is exploited heavily in the hierarchical models literature.

The multivariate regression model in (2.8.33) with different regressors in each equations is called the Seemingly Unrelated Regression model (SUR) by Zellner. This minor change to the model destroys conjugacy. We can no longer utilize the matrix form in (2.8.36) to write the model likelihood as in (2.8.37)[12]. The best we can do is stack up the regression equations into one large regression.

$$y = X\beta + \varepsilon$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix};\ X = \begin{bmatrix} X_1 & 0 & 0 & 0 \\ 0 & X_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & X_m \end{bmatrix};\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix} \tag{2.8.44}$$

with

$$\mathrm{Var}(\varepsilon) = \Sigma \otimes I_n \tag{2.8.45}$$

Conditional on Σ, we can introduce a normal prior, standardize the observations to remove correlation and produce a posterior. However, we cannot find a convenient prior to integrate out Σ from this conditional posterior. We can also condition on β, use an IW prior on Σ and derive an IW posterior for Σ. The reason we can do this is that, given β, we "observe" the ε directly and we are back to the problem of Bayesian inference for a

[12] This should not be too surprising since we know from the standard econometrics treatment that the MLE for the SUR is not the same as equation by equation least squares.

correlation matrix with zero mean. In Chapter 3, we will illustrate a very simple simulation method based on the so-called Gibbs Sampler for exploring the marginal posteriors of the SUR model with a normal prior on β and an IW prior on Σ.

A workhorse model in the marketing literature is the Multinomial Logit model. The dependent variable is a multinomial outcome whose probabilities are linked to independent variables which are alternative specific. $y_i = \{0, 1, \ldots, J\}$ with probability p_{ij}.

$$p_{i,j} = \frac{\exp\left(x_{ij}'\beta\right)}{\sum_{k=0}^{J} \exp\left(x_{ik}'\beta\right)} \tag{2.8.46}$$

x_{ij} represent alternative specific attributes. Thus, the likelihood for the data (assuming independence of the observations) can be written as

$$p\left(y|\beta\right) = \prod_{i=1}^{n} p_{i,y_i} = \prod_{i=1}^{n} \frac{\exp\left(x_{iy_i}'\beta\right)}{\sum_{k=0}^{J} \exp\left(x_{ik}'\beta\right)} \tag{2.8.47}$$

Given that this model is in the exponential family, there should be a natural conjugate prior.[13] However, all this means is that the posterior will be in the same form as the likelihood. This does not assure us that we can integrate that posterior against any interesting functions. Nor does the existence of a natural conjugate prior insure that it is interpretable and, therefore, easily assessable. Instead, we might argue that a normal prior on β would be reasonable. The posterior, of course, is not in the form of a normal or, for that matter, any other standard distribution. While we might have solved the problem of prior assessment, we are left with the integration problem. Since it is known that the log-likelihood is globally concave, we might expect asymptotic methods to work reasonably well on this problem. We explore these in the next section.

2.9 Integration and Asymptotic Methods

Outside the realm of natural conjugate problems, we will have to resort to numerical methods to compute the necessary integrals for Bayesian inference. The integration problem is of the form

$$I = \int h(\theta)\, p(\theta)\, p(D|\theta)\, d\theta \tag{2.9.1}$$

This includes the computation of normalizing constants, moments, marginals, credibility intervals and expected utility.

There are three basic sorts of methods for approximating integrals of the sort given in (2.9.1): 1). Approximate the integrand by some other integrand which can be integrated numerically; 2). Approximate the infinite sum represented by the integral through some finite sum approximation such as a quadrature method; or 3). View the integral as an expectation with respect to some density and use simulation methods to approximate the expectation by a sample average over simulations.

The first approach is usually implemented via resort to an asymptotic approximation of the likelihood. We can expand the log-likelihood in a second order Taylor series about the MLE and use this as a basis of a normal kernel.

$$I \doteq \int h(\theta)\, p(\theta) \exp\left\{ L(\theta)\Big|_{\theta=\hat{\theta}} + \tfrac{1}{2}\left(\theta-\hat{\theta}\right)' H \left(\theta-\hat{\theta}\right)\right\} d\theta \tag{2.9.2}$$

$H = \left[\dfrac{\partial^2 L}{\partial\theta\partial\theta'}\right]$. This is the classic Laplace approximation. If we use the "asymptotic" natural conjugate normal prior, we can compute approximate normalizing constants and moments.

---

[13] See Robert and Casella (2004), p. 146, for an example of the conjugate prior for J =2.

(2.9.3) $$I \doteq e^{L(\hat{\theta})} \int h(\theta)(2\pi)^{-k/2} |A|^{1/2} \exp\left\{-\tfrac{1}{2}(\theta-\bar{\theta})' A(\bullet)\right\} \exp\left\{-\tfrac{1}{2}(\theta-\hat{\theta})' H^*(\bullet)\right\}$$

$H^* = -H$ (negative of the Hessian of the log-likelihood). Completing the square in the exponent,[14] we obtain

(2.9.4)
$$I \doteq e^{L(\hat{\theta})-\frac{1}{2}SS} |A|^{1/2} \left|A+H^*\right|^{-1/2} \times$$
$$\int h(\theta)(2\pi)^{-k/2} \left|A+H^*\right|^{1/2} \exp\left\{-\tfrac{1}{2}(\theta-\tilde{\theta})' (A+H^*)(\theta-\tilde{\theta})\right\} d\theta$$

where $SS = (\tilde{\theta}-\bar{\theta})' A(\bullet) + (\tilde{\theta}-\hat{\theta})' H^*(\bullet)$ and $\tilde{\theta} = (A+H^*)^{-1}(A\bar{\theta}+H^*\hat{\theta})$. (2.9.4) implies that the normalizing constant and first two posterior moments ($h(\theta)=\theta$; $h(\theta)=\theta^2$) are given by

(2.9.5)
$$\int p(\theta) p(\theta|D) d\theta \doteq e^{L(\hat{\theta})-\frac{1}{2}SS} |A|^{1/2} \left|A+H^*\right|^{-1/2}$$
$$E[\theta|D] \doteq \tilde{\theta}$$
$$Var(\theta|D) \doteq (A+H^*)^{-1}$$

While quadrature methods can be very accurate for integrands that closely resemble a normal kernel multiplied by a polynomial, many posterior distributions are not of this form. Moreover, quadrature methods suffer from the curse of dimensionality (the number of computations required to achieve a given level of accuracy increases to the power of the dimension of the integral) and are not useful for more than a few dimensions. For this reason, we will have to develop specialized simulation methods that will rely on

[14] Recall the standard completing the square result.

approximating (2.9.1) with a sample average of simulated values. One such method is importance sampling.

2.10 Importance Sampling

While we indicated that it is not feasible to devise efficient algorithms for producing iid samples from the posterior, it is instructive to consider how such samples could be used if available. The problem is to compute

$$E_{\theta|D}\left[h(\theta)\right] = \int h(\theta)\, p(\theta|D)\, d\theta .$$

If $\{\theta_1,\ldots,\theta_R\}$[15] are random sample from the posterior density, then we can approximate this integral by a sample average.

(2.10.1) $$\bar{h}_R = \tfrac{1}{R}\textstyle\sum_r h(\theta_r)$$

If $\mathrm{Var}(h(\theta))$ is finite, we can rely on the standard theory of sample averages to obtain an estimate of the accuracy of $\bar{h}_R$.

(2.10.2)
$$\mathrm{STD}(\bar{h}_R) = \frac{\sigma}{\sqrt{R}};\ \sigma^2 = E_{\theta|D}\left[\left(\bar{h}_R - E_{\theta|D}\left[h(\theta)\right]\right)^2\right]$$
$$\mathrm{STDERR}(\bar{h}_R) = \frac{s}{\sqrt{R}};\ s^2 = \tfrac{1}{R}\textstyle\sum_r\left(\bar{h}_r - \frac{\sum_r \bar{h}_r}{R}\right)^2$$

STD denotes standard devation and STDERR denotes the estimated standard deviations or standard error. The formula in (2.10.2) has tremendous appeal since the accuracy of the integral estimate is *independent* of the dimension of θ (k). This is true in a strict technical sense. However, in many applied contexts, the variation of the h function increases as the

---

$$(x-\mu_1)'A_1(x-\mu_1)+(x-\mu_2)'A_2(x-\mu_2) =$$
$$(x-\tilde{\mu})'(A_1+A_2)(x-\tilde{\mu})+(\mu_1-\tilde{\mu})'A_1(\mu_1-\tilde{\mu})+(\mu_2-\tilde{\mu})'A_2(\mu_2-\tilde{\mu})$$
where $\tilde{\mu} = (A_1+A_2)^{-1}(A_1\mu_1 + A_2\mu_2)$

[15] We use R to denote the size of the simulation sample in order to distinguish this from the sample size (n).

dimension of θ increases. This means that in some situations σ in (2.10.2) is a function of k. The promise of Monte Carlo integration to reduce or eliminate the curse of dimensionality is somewhat deceptive.

However, since we don't have any general purpose methods for generating samples from the posterior distribution, the classic method of Monte Carlo integration is not of much use in practice. The method of importance sampling uses an importance sampling density to make the Monte Carlo integration problem practical. Assume that we only have access to the un-normalized posterior density (the product of the prior and the likelihood). We then must estimate

$$E_{\theta|D}\left[h(\theta)\right] = \frac{\int h(\theta)p(\theta)p(D|\theta)d\theta}{\int p(\theta)p(D|\theta)d\theta}$$

Suppose we have an importance distribution with density g. We can rewrite the integrals in the numerator and denominator as expectations with respect to this distribution and then use a ratio of sample averages to approximate the integral.

(2.10.3)
$$E_{\theta|D}\left[h(\theta)\right] = \frac{\int h(\theta)\left[p(\theta)p(D|\theta)\Big/g(\theta)\right]g(\theta)d\theta}{\int\left[p(\theta)p(D|\theta)\Big/g(\theta)\right]g(\theta)d\theta}$$

This quantity can be approximated by sampling $\{\theta_1,\ldots,\theta_R\}$ from the importance distribution.

(2.10.4)
$$\bar{h}_{IS,R} = \frac{\frac{1}{R}\sum_r h(\theta_r)w_r}{\frac{1}{R}\sum_r w_r} = \frac{\sum_r h(\theta_r)w_r}{\sum_r w_r}$$

$w_r = p(\theta_r)p(D|\theta_r)\big/g(\theta_r)$. (2.10.4) is a ratio of weighted averages. Note that we do not have to use the normalized density g in computing the importance sampling estimate as the

normalizing constants will cancel out from the numerator and denominator. The kernel of g is often called the *importance function.*

Assuming that $E_{\theta|d}\left[h(\theta)\right]$ is finite, then the law of large numbers implies that $\bar{h}_{IS,R}$ converges to the value of (2.10.3). This is true for any importance density that has the same support as the posterior. However, the choice of importance sampling density is critical for proper functioning of importance sampling. In particular, the tail behavior of the posterior relative to the importance density is extremely important. If the ratio $p(\theta)p(D|\theta)/g(\theta)$ is unbounded, then the variance of $\bar{h}_{IS,R}$ can be infinite (see Geweke (1989), theorem 2, or p. 82 of Robert and Casella). Geweke (1989) gives sufficient conditions for a finite variance. These conditions are basically satisfied if the tails of the importance density are thicker than the tails of posterior distribution.[16] The standard error of the importance sampling estimate (sometimes called the "numerical standard error" can be computed as follows:

$$\text{STDERR}\left(\bar{h}_{IS,R}\right) = \sqrt{\frac{\sum_r \left(h(\theta_r) - \frac{\sum_r h(\theta_r)}{R}\right)^2 w_r^2}{\left[\sum_r w_r\right]^2}} \tag{2.10.5}$$

One useful suggestion for an importance density is to use the asymptotic approximation to the posterior developed in (2.9.4). The thin normal tails can be fattened by scale mixing to form a multivariate student t distribution with low degrees of freedom. This strategy was suggested by Zellner and Rossi (1984). Specifically, we develop an

[16] One of the conditions is that the ratio of the posterior to the importance density is finite. Robert and Casella point out that this means that g could be used in an rejection method. This is a useful observation only in very small dimensional parameter spaces. However, the real value of importance sampling methods are for moderate to reasonably high dimensional parameter spaces. In any more than a few dimensions, rejection methods based on the importance density would grind to a halt, rejecting a huge fraction of draws.

importance function as $\text{MSt} - \text{t}\left(\nu, \hat{\theta}_{\text{MLE}}, \text{s}\left(-\text{H}\Big|_{\theta=\hat{\theta}_{\text{MLE}}} \right)^{-1} \right)$. One could also use the posterior mode and Hessian evaluated at the posterior mode.

Draws from this distribution can be obtained via (2.11.6). "Trial" runs can be conducted with different degrees of freedom values. The standard error in (2.10.5) can be used to help "tune" the degrees of freedom and the scaling of the covariance matrix (s). The objective is to fatten the tails sufficiently to minimize variation in the weights. It should be noted that very low degree of freedom student t distributions ($\upsilon < 5$) become very "peaked" and do not make very good importance functions as the "shoulders" of the distribution are too narrow. We recommend moderate values of the degrees of freedom parameter and more emphasis on choice of the scaling parameter to broaden the "shoulders" of the distribution.

Geweke (1989) provides an important extension to this idea by developing a "split-t" importance density. The "split-t" importance function can handle posterior distributions which are highly skewed. The standard MSt importance function will be inefficient for skewed integrands as considerable fattening will be wasted in one "tail" or principal axis of varation.

A fat tailed importance function based on the posterior mode and Hessian can be very useful in solving integration problems in moderate to large dimension problems (2 to 20). For example, the extremely regular shape of the multinomial logit likelihood means that importance sampling will work very well for this problem. Importance Sampling has the side advantage of only requiring a maximizer and its associated output. Thus, importance sampling is almost a free good to anyone who is using the MLE.

But, there is also a sense that this exposes the limitations of importance sampling. If the posterior looks a lot like its asymptotic distribution, then importance sampling with the importance density proposed here will work well. However, it is in these situations that finite sample inference is apt to be least valuable.[17] If you want to tackle problems with very non-normal posteriors, then greater care must be taken in the choice of importance function. The variance formulas can be deceiving. For example, suppose we situate the importance density far from the posterior mass. Then the weights will not vary much and we may convince ourselves using the standard error formulas that we have estimated the integral very precisely. Of course, we have completely missed the mass we want to integrate against. In high dimensional problems, this situation can be very hard to detect. In very high dimensions, it would be useful to impose greater structure on the parameter space and break the integration problem down into more manageable parts. This is precisely what the methods and models discussed in Chapter 3 are designed to do. Finally, there are many models for which direct evaluation of the likelihood function is computationally intractable. Importance sampling will be of no use in these situations. However, importance sampling can be used as part of other methods to tackle many high dimensional problems, even those with intractable likelihoods.

2.10.1 GHK Method for Evaluation of Certain Integrals of MVN

In many situations, the evaluation of integrals of a multivariate normal distribution over a rectangular region may be desired. The rectangular region A is defined by $A = \{ x : a < x < b\}$ where a and b are vectors of endpoints which might include infinity. Let $P = \Pr(x \in A)$ with $x \sim N(0,\Sigma)$. The GHK method (Keane (1994) and Hajivassiliou, McFadden and Ruud

[17] One classical econometrician is rumored to have said – "importance sampling just adds fuzz to our standard asymptotics."

(1996)) uses an importance sampling method to approximate this integral. The idea of this method is to construct the importance function and draw from it using univariate truncated normals.

We can define P either in terms of the correlated normal vector, x, or in terms of a vector of z of uncorrelated unit normals, $z \sim N(0, I)$.

$$
\begin{aligned}
&x = Lz; \quad \Sigma = LL' \\
&P = \Pr(x \in A) = \Pr(a < x < b) = \Pr(a < Lz < b) \\
&\quad = \Pr\left(L^{-1}a < z < L^{-1}b\right) = \Pr(z \in B) \\
&B = \left\{z : L^{-1}a < z < L^{-1}b\right\}
\end{aligned}
$$

L is the lower triangular Cholesky root of Σ. We express the density of z as the product of a series of conditional densities which will allow us to exploit the lower triangular array which connects z and x.

$$
p(z|B) = p(z_1|B)\,p(z_2|z_1, B)\cdots p(z_m|z_1, \ldots, z_{m-1}, B) \tag{2.10.6}
$$

The region defined by A can be expressed in terms of elements of z as follows

$$
\begin{aligned}
&a_j < x_j < b_j \\
&a_j < l_{j,1}z_1 + l_{j,2}z_2 + \ldots + l_{j,j-1}z_{j-1} + l_{j,j}z_j < b_j
\end{aligned}
$$

$l_{i,j}$ are the elements of the lower Cholesky root, L. Given $z_{<j} = (z_1, z_2, \ldots, z_{j-1})$, this inequality implies that z is a univariate normal truncated to a particular interval given by

$$
\frac{a_j - \mu_j(z_{<j})}{l_{jj}} < z_j < \frac{b_j - \mu_j(z_{<j})}{l_{jj}} \tag{2.10.7}
$$

$\mu_j(z_{<j}) = l_{j,1}z_1 + l_{j,2}z_2 + \ldots + l_{j,j-1}z_{j-1}$. (2.10.7) provides an algorithm for drawing from $p(z|B)$. We simply draw $z_1 | B$ and then $z_2 | z_1, B$ and so on to fill out the z vector. The equation also provides us with a way to evaluate the conditional density of $z | B$.

$$p\left(z_j \middle| z_{<j}, B\right) = \frac{\phi\left(z_j\right)}{D_j\left(z_{<j}\right)} \text{ with } D_j\left(z_{<j}\right) = \Phi\left(\frac{b_j - \mu_j}{l_{jj}}\right) - \Phi\left(\frac{a_j - \mu_j}{l_{jj}}\right)$$

Using (2.10.6), $p\left(z|B\right) = \frac{\prod \phi\left(z_j\right)}{\prod D_j\left(z_{<j}\right)} = \frac{f\left(z\right)}{D\left(z\right)}$

We now write the integral defining P in terms of the density of z and use p(z|B) as an importance function.

$$\begin{aligned} P = \Pr\left(z \in B\right) &= \int_B f\left(z\right) dz \\ &= \int_B \frac{f\left(z\right)}{p\left(z|B\right)} p\left(z|B\right) dz \\ &= \int_B D\left(z\right) p\left(z|B\right) dz \end{aligned} \tag{2.10.8}$$

Thus, the GHK algorithm can be constructed as follows:

GHK ALGORITHM

Draw z from p(z|B) using (2.10.7) and truncated univariate normal draws

Evaluate D(z)

Repeat R times and form the estimate $\hat{P} = \frac{1}{R}\sum D\left(z_r\right)$

2.11 Simulation Primer for Bayesian Problems

If we could construct an iid sample directly from the posterior, the problem of summarizing the posterior could be solved to any desired degree of simulation accuracy. Unfortunately, the problem of generating random variables from an arbitrary (and possibly very high dimensional) distribution has no general purpose and computationally tractable solution. We will have to exploit the special structure of Bayesian models in order to develop useful methods. The basis for all of these methods are methods of simulating random variates from a set of frequently used distributions.

2.11.1 Uniform, Normal and Gamma Generation

All methods of continuous random variate generation start with a uniform pseudo random number generator. Univariate pseudo random number generators generate deterministic (conditional on a seed) sequences of numbers which pass various "tests" for distributional accuracy and appearance of randomness. In particular, the sequence should have a very long "period" before it repeats itself, exhibit minimal time dependence (sometimes measured by autocorrelation), and do a good job of "filling" a k-dimenisonal hyper-cube constructed from sub-sequences of length k. R uses the Mersenne Twister (Matsumoto and Nishimura (1998)) as the default method. Some might argue that the KISS generator by Marsaglia (1999) is faster.

Any standard computing environment will also supply methods for generating normal as well as gamma distributed random variates. By default, R uses the Inverse CDF method to generate normal random variates. This certainly produces draws with excellent properties but it may be slightly inefficient compared to method such as the Ziggurat method of Marsaglia and Tsang (2000). However, we have not found the computations required for normal draws to be a computational bottleneck in Bayesian computations.

R also provides methods to draw Gamma and chi-squared random variates using the Inverse CDF method. If your computing environment does not provide high quality Gamma random variates, the method of Marsalgia and Tsang (2000) can be programmed in a low level language such as C.

We have seen in section 2.8 that draws from the Inverted Gamma prior and posterior for σ^2 are needed. Recall that $\text{Inv Gamma}\left(\nu, s^2\right) \sim \nu s^2 \Big/ \chi^2_\nu$.

(2.11.1) $$\text{Inv-Gamma}\left(\upsilon, s^2\right) = \frac{\nu s^2}{2 \times \text{Gamma}\left(\frac{\nu}{2}, 1\right)}$$

The uniform, normal and Gamma methods available in R and many other computing environments can be used to construct simulators for many of the distributions needed for Bayesian inference. We will now provide a simulation primer for these distributions.

2.11.2 Truncated Distributions

The Inverse CDF method can be used to draw from truncated distributions, provided that a computationally efficient method is available to evaluate the inverse CDF or quantile function. To review the inverse CDF method, let $F_X^{-1}(x)$ be the inverse of the cdf of random variable X. [18] Then $X = F_X^{-1}(U)$, $U \sim \text{unif}(0,1)$ has distribution with cdf F_X. An important example is the truncated normal distribution. Consider the normal distribution truncated to the interval (a,b).

$$Y = X * I_{(a,b)}(X), \quad I_{(a,b)}(X) = 1 \text{ if } X \in (a,b), 0 \text{ otherwise}$$

The cdf of Y can be obtained from the cdf of X.

$$G_Y(y) = \frac{F(y) - F(a)}{F(b) - F(a)}$$

Let $p = G_Y(y)$ and solve for G^{-1}.

(2.11.2) $$p = F^{-1}\left(y\left(F(b) - F(a)\right) + F(a)\right)$$

[18] To avoid any technical difficulties, we must assume that there are no jumps in F or that the distribution of X is absolutely continuous wrt to Lebesgue measure.

Recall that $F(x) = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right)$ and $F^{-1}(p) = \mu_X + \sigma_X \Phi^{-1}(p)$. We have a simple algorithm for simulating from the truncated normal.

TRUNCATED NORMAL ALGORITHM**:**

Draw $U \sim \text{Unif}(0,1)$

$$y = \mu + \sigma\Phi^{-1}\left(U\left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right) + \Phi\left(\frac{a-\mu}{\sigma}\right)\right)$$

$$Y \sim \text{trun}_{(a,b)}N(\mu,\sigma)$$

There is a legitimate argument that this algorithm is computationally inefficient due to the evaluation of the normal CDF and inverse CDF functions. A combination of rejection sampling and other methods can be used to develop more efficient algorithms (c.f. McCulloch and Rossi (1994)). However, for a language such as R or MATLAB where vectorization is essential for efficiency, the algorithm based on the cdf method is more efficient. In our R package ***bayesm***, we provide the routine `rtrun` to simulate a vector of truncated normals.

2.11.3 Multivariate Normal and student-t Distributions

Given unit or standard normal draws, we can create the general Normal variate by scale and location transform. Multivariate normal draws can be calculated via the following algorithm:

(2.11.3)
$$\begin{aligned} z' &= (z_1,\ldots,z_k);\ z_i \sim \text{iidN}(0,1) \\ x &= U'z + \mu \sim N(\mu,\Sigma);\ \Sigma = U'U \end{aligned}$$

U is the upper triangular Cholesky root of Σ (the so-called "LU" decomposition). Computation of roots of positive definite real matrices has been studied closely and there are

reliable and computationally efficient algorithms for doing so (c.f. the implementation in LAPACK used by R).

In some cases, we must simulate from various the conditional distribution of a subvector of a multivariate normal given the remainder of the normal random vector. The fact that the conditional mean and conditional covariance matrix can be computed directly from the elements of the inverse of Σ is useful.

If
$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$
$$\Sigma^{-1} = V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$
Then

$$x_1 \left| x_2 \sim N\left(\mu_1 - V_{11}^{-1} V_{12} \left(x_2 - \mu_2 \right), V_{11}^{-1} \right) \right. \tag{2.11.4}$$

The multivariate t distribution is an elliptically symmetric distribution closely related to the multivariate normal distribution. The major difference is that the multivariate t has algebraic tails (not exponential like the normal) and it can be very peaked for very low degrees of freedom. A k dimensional multivariate t distribution with degrees of freedom parameter ν, location parameter μ and scale parameter Σ has density

$$p\left(x \left| \nu, \mu, \Sigma \right. \right) \propto \left| \Sigma \right|^{-1/2} \left[\nu + \left(x - \mu \right)' \Sigma^{-1} \left(x - \mu \right) \right]^{-(k+\nu)/2} \tag{2.11.5}$$

$E\left[x \right] = \mu;\ \mathrm{Var}\left(x \right) = \frac{\nu}{\nu - 2} \Sigma$. We can simulate from this distribution by fattening the tails of a multivariate normal.

$$\begin{aligned} X &= \frac{Y}{\left(Z / \nu \right)^{1/2}} + \mu;\ Y \sim N\left(0, \Sigma \right);\ Z \sim \chi_\nu^2 \\ X &\sim \mathrm{MSt}\left(\nu, \mu, \Sigma \right) \end{aligned} \tag{2.11.6}$$

Our R package, ***bayesm***, includes `rmvst` to simulate from the multivariate student t distribution.

2.11.4 The Wishart and Inverted Wishart Distributions

In section 2.8, we observed that prior and posteriors for the covariance matrix of the multivariate normal distribution are of the Inverted Wishart form. The correspondence between the Wishart and Inverted Wishart distribution can be exploited to develop a simulator. Let $G \sim W\left(\nu, V^{-1}\right)$ and factor V^{-1} into the product of upper Cholesky roots, $V^{-1} = U'U$. Then $G = U'BU$ where $B \sim W\left(\nu, I_m\right)$. To simulate from the standard Wishart, we construct the following lower triangular array of random variates (note all non-zero elements are independent. $Z_{i,j} \sim N(0,1)$.

$$T = \begin{bmatrix} \sqrt{\chi^2_\nu} & 0 & \cdots & 0 \\ Z_{2,1} & \sqrt{\chi^2_{\nu-1}} & & \vdots \\ \vdots & & \ddots & 0 \\ Z_{m,1} & \cdots & Z_{m,m-1} & \sqrt{\chi^2_{\nu-m+1}} \end{bmatrix} \tag{2.11.7}$$

$Z_{i,j} \sim N(0,1)$ and $\nu > m$.

$$TT' \sim W\left(\nu, I_m\right)$$

G can be constructed from the pieces.

$$\begin{aligned} G &= U'TT'U = C'C \\ C &= T'U \end{aligned} \tag{2.11.8}$$

The Inverted Wishart can be computed from the draw of the corresponding Wishart using (2.11.8) by taking the inverse of C.

$$\Sigma = G^{-1} = C^{-1}\left(C^{-1}\right)' \tag{2.11.9}$$

Note that (2.11.9) is the UL decomposition of Σ.

WISHART/INVERTED WISHART DRAW ALGORITHM**:**

To draw $\Sigma \sim IW(\nu, V)$ or $\Sigma^{-1} \sim W(\nu, V^{-1})$:

1. Factor $V^{-1} = U'U$
2. Draw random variates and compute T as in (2.11.7)
3. Compute $C^{-1} = (T'U)^{-1}$, $\Sigma = C^{-1}(C^{-1})'$, $\Sigma^{-1} = C'C$

bayesm includes the function, `rwishart`, to simulate from both the Inverted Wishart and Wishart distribution.

2.11.5 Multinomial Distributions

The most general discrete distribution is the multinomial distribution. If θ can take on a d values, $S = \{\theta^1, \theta^2, \ldots, \theta^d\}$, each with probability p_i, then $\theta \sim MN(p, S)$.

MULTINOMIAL DRAW ALGORITHM**:**

To draw from the multinomial, we only need to draw the index into S with the appropriate probability.

Draw $U \sim Unif(0,1)$

Find k such that $\sum_{i=0}^{k-1} p_i < U \leq \sum_{i=0}^{k} p_i$ with $p_0 = 0$

$\theta = S(k)$

2.11.5 Dirichlet Distribution

The natural conjugate prior for the multinomial distribution is called the Dirichlet distribution with density

(2.11.10) $$p(\theta|\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_1 + \ldots \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \quad \alpha_i > 0$$

θ is a k dimensional vector that must be in the unit simplex.

DIRICHLET DRAW ALGORITHM**:**

Draw $x_i \sim$ indep Gamma(α_i, α_i)

$$\theta_i = \frac{x_i}{\sum_j x_j}$$

Our R package, ***bayesm***, includes `rdirichlet` to simulate from the Dirichlet distribution.

2.12 Simulation from Posterior of Multivariate Regression Model

The Multivariate Regression model discussed in section 2.8.4 is a very useful model not only because of direct application to situations with sets of related linear regressions but also in various hierarchical model settings in which the model is used as part of the prior structure. To analyze these hierarchical models, it will be necessary to use simulation-based methods. For this reason, it will be useful to have an efficient algorithm for sampling from the posterior of this model given in (2.8.43). Recall the basic set-up.

Model:

(2.12.1) $$Y = XB + U \quad U = \begin{bmatrix} u_i' \end{bmatrix} \quad u_i \sim N(0, \Sigma)$$

Here Y is n x m, X is n x k, B is k x m where each row of B contains the regression coefficients for one of the m equations.

Prior:

(2.12.2) $$\begin{aligned} &\Sigma \sim IW(\nu, V) \\ &\beta = vec(B) | \Sigma \sim N\left(vec(\bar{B}), \Sigma \otimes A^{-1}\right) \end{aligned}$$

and

To draw from the posterior, we first draw Σ and then draw B given Σ.

Draw of Σ

$$
\begin{aligned}
&\Sigma \,|\, Y, X \sim IW\left(\nu + n, V + S\right) \\
&S = E'E \\
&E = Y - X\tilde{B} + \left(\tilde{B} - \bar{B}\right)' A \left(\tilde{B} - \bar{B}\right) \\
&\tilde{B} = \left(X'X + A\right)^{-1}\left(X'Y + A\bar{B}\right)
\end{aligned}
\tag{2.12.3}
$$

S and $\tilde{B}$ can be computed using the QR decomposition. The QR decomposition of a matrix X is X=QR where Q is a n x k matrix whose columns are orthogonal and R is a k x k upper triangular matrix that is the Cholesky root of X'X up to sign differences. We can compute the relevant quantities in (2.12.3) by forming following augmented matrices.

$$
W = \begin{bmatrix} X \\ R_A \end{bmatrix} \quad Z = \begin{bmatrix} Y \\ R_A \bar{B} \end{bmatrix}
\tag{2.12.4}
$$

If we take the QR decomposition of W, $W = Q_W R_W$, then

$$
\tilde{B} = R_W^{-1} Q_W Z \quad \text{and} \quad S = \left(Z - W\tilde{B}\right)'\left(Z - W\tilde{B}\right) = Z'\left(I - Q_W Q_W'\right)Z
\tag{2.12.5}
$$

However, timing experiments in R suggest that it is faster to compute these quantities by using the Cholesky root approach (between 30 and 100 per cent faster than using the LAPACK QR method in R). We take the Cholesky root of $W'W$, invert this, and use this to compute $\tilde{B}$ and E[19]. The inverse of the root can be computed by efficiently in R and the root can also be used to draw from the posterior as we see below.

$$
\begin{aligned}
&W'W = R_{W'W}' R_{W'W} \quad \text{and} \quad IR_{W'W} = R_{W'W}^{-1} \\
&\tilde{B} = IR_{W'W}\left(IR_{W'W}\right)'W'Z \\
&E = Z - W\tilde{B} \\
&S = E'E
\end{aligned}
\tag{2.12.6}
$$

[19] Golub and Van Loan (1989), p. 226, state that the QR decomposition will handle W'W arrays that are more nearly non-singular than the Cholesky root approach. Both are very stable numerically and, in Bayesian computations with proper priors, we never approach non-singularity.

$R_{W'W}$ is the upper triangular Cholesky root of $W'W$. We note that $IR_{W'W}$ defines the UL decomposition of $(W'W)^{-1}$ whereas $R_{W'W}$ forms the LU decomposition of $W'W$

To draw Σ, we draw from appropriate Wishart and then invert this matrix to obtain the Σ draw.

$$\Sigma^{-1} \sim W\left(\nu + n, (V + S)^{-1}\right)$$

The Bartlett strategy outlined in section 2.11.4 produces a draw of the Cholesky root of Σ^{-1}.

$$\begin{aligned} \Sigma^{-1} &= C'C \\ \Sigma &= \left(C^{-1}\right)\left(C^{-1}\right)' = CICI' \end{aligned} \tag{2.12.7}$$

We note that we have the LU decomposition of Σ^{-1} and the UL decomposition of Σ in (2.12.7).

Draw of $\beta|\Sigma$

The direct, but naïve, approach would be to draw β from a $N\left(\text{vec}(\bar{B}), \Sigma \otimes (X'X + A)^{-1}\right)$ distribution. We can exploit the special structure of the covariance matrix to develop an efficient strategy for making this draw.

$$\begin{aligned} \text{Var}(\beta) = \Omega &= \Sigma \otimes (X'X + A)^{-1} \\ &= CICI' \otimes \left(R'_{W'W} R_{W'W}\right)^{-1} \\ &= CICI' \otimes IR_{W'W} IR'_{W'W} \\ &= (CI \otimes IR_{W'W})(CI \otimes IR_{W'W})' \end{aligned} \tag{2.12.8}$$

$IR_{W'W} = R^{-1}_{W'W}$. Note that (2.12.8) is the UL decomposition of the covariance matrix. Thus, we can use this root directly to produce normal variates with the right covariance.

$$v = (CI \otimes IR_W) z \quad z \sim N(0, I_{m \times k})$$
$$Var(v) = \Omega$$

However, we can simplify this even further by using the identity $vec(ABC) = (C' \otimes A) vec(B)$.

(2.12.9) $$B = \tilde{B} + IR_{W'W} ZCI' \quad vec(Z) = z$$

`rmultireg` in our R package, ***`bayesm`***, implements this strategy.

Figure 2.1