Notes on Bayesian Analysis and Other Cool Stuff

Hillary Sanders

January 2, 2013

Contents

1 What is Bayesian Statistics **5**
1.1 Bayesian vs. Classical, Frequentist Statistics 5

2 Probability Review **9**
2.1 Random variables 10
2.2 Probability distributions 10
2.3 Moments: mean and variance 11
2.4 Likelihood Function 12
2.5 Joint distributions 13
2.6 Conditional distributions & independence 13
2.7 Important Probability Distributions 14
2.8 Bayes' Theorem 14
2.9 Vocab Review 15

3 The Story: Likelihoods, Priors, and Posteriors **17**
3.1 Wait, what? 17
3.2 Bayesian Inference 18
3.3 Hierarchical Models 21
3.4 Multi-Parameter Bayesian Models 21
3.5 Vocab Review 22

4 Priors **23**
4.1 What is a Prior? 23
4.2 Conjugate Priors 24
4.3 Non-Informative Priors 25
4.4 Well, Shit 26

5 Grid Approximation **27**

6 Model Checking & Model Comparison **29**

7 Bayesian Sampling **31**
7.1 Rejection Sampling 31
7.2 Sampling Importance Resampling (SIR) 32
7.3 MCMC Methods 32
7.3.1 Markov Chain 33
7.3.2 Gibbs Sampling 34
7.3.3 Metropolis 35
7.3.4 Metropolis Hastings 36
7.4 MCMC Diagnostics 37
7.4.1 Traceplots 37
7.4.2 Autocorrelation Plots 37
7.4.3 Multiple Sequence Diagnostic 37

8 Linear Models **39**
8.1 Simple Linear Models 40
8.2 Hierarchical Linear Models 40

9 Bayesian Model Averaging **43**

10 Mixture Models **45**

11 Bayesian Hypothesis Testing **47**
11.1 Frequentist (Classical) Hypothesis Testing 47
11.2 Bayesian Hypothesis Testing 47

12 Summary **51**

Chapter 1

What is Bayesian Statistics

1.1 Bayesian vs. Classical, Frequentist Statistics

Bayesian Statistics is a system for statistical inference and decision making which is based on expressing all uncertainty in terms of conditional probability statements. For example: in classical, frequentist statistics, a theory has no *probability* of being correct or incorrect - it is either correct or incorrect, we're just unsure about the truth. But in Bayesian statistics, we're allowed to construct probability models over both observations (data), and beliefs (unknown parameters and theories), and so come up with an actual $p(theory|data)$.

There are two aspects of Bayesian Statistics that are important to understand. The first is that unknown variables are not treated as they are in classical, frequentist statistics. In frequentist statistics, even if some value θ is unknown to us, it is assumed that θ has some true numeric value. Take the proportion of people in America who like Lady Gaga. A frequentist assumes that there is some true percentage of people in America who like Lady Gaga. And he can estimate that percentage by, for example, asking a (hopefully) random sample of Americans if they like Lady Gaga, and from that he can develop a confidence interval.

Say he gets a 95% confidence interval from 22% to 25%. It's not that he believes that there is a 95% chance that θ is somewhere inside the interval

[.22, .25], rather that there is a 95% chance that his confidence interval has overlapped with the fixed, real value of θ.

The Bayesian approach is a little different. A Bayesian might develop a similar confidence interval - say it's [.22, .25] again, and say that she is 95% sure that θ is equal to some value in [.22, .25].

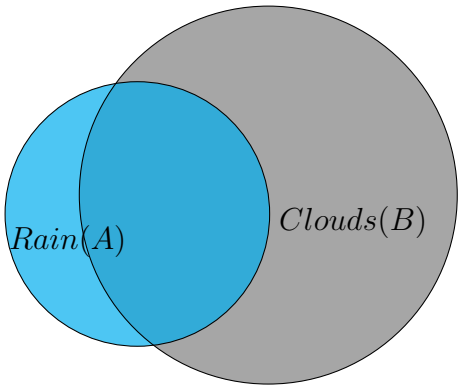
Foul, you say! Er, at least that's what I said. This is all well and good, but there actually *is* some fixed, true value of θ. We shouldn't place a distribution over θ - it's fixed (at least at a fixed time T)! Just unknown! Right? Well, sort of. Here's the thing: there's an amazing amount of very powerful methods and tools you can utilize *if* you treat unknown variables like a Bayesian does. And instead of thinking that you're placing a probability distribution over θ, which actually is a bit silly (in most circumstances), you can think of the distribution as a *distribution of your beliefs* about all possible values of θ - which, handily enough, can be altered by new, incoming data.

Okay, so that's the first important aspect of Bayesian Statistics: treat unknown variables as variables to be pulled from belief distributions that you define based on information, rather than fixed variables you attempt to estimate. The second important aspect of Bayesian Statistics is the handy result that comes from allowing unknown variables to have probability distributions. Basically, it is the ability to infer a new probability distribution over your unknown variable by using new evidence (data), through the use of Bayes Rule.

Ever heard of Bayes Rule? Yeah, I did to in my first statistics class (college, sophomore year), and it pretty much annoyed the heck out of me that something so obvious and intuitive could be named after - in all likelihood - an old white guy who probably believed in God. But, it's important to go over it, because the implications of this formula are very powerful, and very important to understand well.

$$p(A|B) = \frac{p(B|A) * p(A)}{p(B)} \tag{1.1}$$

(If this equation makes total sense to you, skip the following paragraph.)



In English, this says that the probability of event A, given that event B has happened, is the probability of event B happening, given A, multiplied by the probability of A in the first place (in other words, the probability of both A and B happening, also equal to $p(A|B) * p(B)$, all divided by the probability of B happening. So, the probability of B and A happening, over the probability of B happening, equals the probability of A happening, given B.

An example: say that A = rain, and B = a cloudy day. Then the probability of rain, given the fact that it's cloudy outside, is just the proportion of days that are both cloudy *and* rainy, divided by the proportion of days that are at least cloudy! So, given that we have a cloudy day (we're inside the grey circle above), the probability of it raining is just the area of the dark blue region (rain and clouds) divided by the entire clouds circle. This should make sense. Try to match up this cloud/rain example in your mind with the equation above.

Okay, got that? Good. So, what does this equation have to do with Bayesian Statistics in general? Well, it is very often the case that we want to estimate some parameter of a population (like the proportion of people who like Lady Gaga, θ, in the population of the United States). So if we allow our estimate of, say, θ, to be defined over a distribution, and we have some observed data, we can have:

$$p(\theta|data) = \frac{p(data|\theta)p(\theta)}{p(data)} \tag{1.2}$$

This equation is the source of a lot of what follows in this booklet thingy. Beginning from this equation, we can answer three main questions of inferential parametric statistics:

1. Parameter Estimation (given a model and data)
2. Prediction of New Data Values (given a model and data)
3. Model Checking (which model is best? Is it even good?)

It's very nice to be able to calculate these things with paper and pencil, or in more complex or difficult problems, estimate them (well!) by computer simulations.

Chapter 2

Probability Review

Most of the stuff in this chapter should be a pretty easy review for you - it should make you feel a little snide for already understanding, or at least easily understanding, what is being outlined. Good. If you don't feel even mildly snide about more than a few things, wikipedia that shiznats until it's your willing statistical servant.

1. **Random variables**
2. **Probability distributions**
3. **Moments: mean and variance**
4. **Likelihood Function**
5. **Maximum Likelihood Estimation**
6. **Joint distributions**
7. **Conditional distributions & independence**
8. **Important Probability Distributions**
9. **Bayes' Theorom**

2.1 Random variables

Random variables are random, but come from distributions. For more than that, go read Wikipedia, fool.

2.2 Probability distributions

The probability mass function (pmt) of a random variable X is:

$$f_X(x) = p(X = x) \tag{2.1}$$

The cumulative distribution function (cdf) of a random variable X is defined as:

$$F_X(x) = p(X \leq x), \forall x \in \mathbb{R} \tag{2.2}$$

- Can be discrete or continuous (or both, but that's ... ugly).
- Random variables are identically distributed if they have the same cdf.

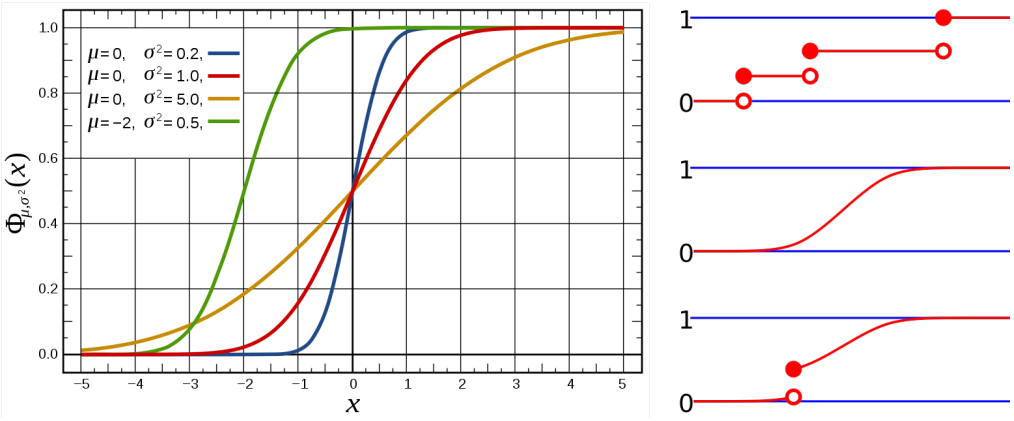


Figure 2.1: On the left, various normal CDFs. On the right, examples of a discrete, continuous, and a discrete-and-continuous CDF.

The probability density function (pdf) of a continuous random variable X is the function f_X that satisfies:

$$\int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = F_X(x) \tag{2.3}$$

Intuitively, f_X is the simply the probability of x occurring for any given point, (but since X is a continuous random variable, the probability of x occurring at any one point is zero). For discrete random variables, f_X is the function which outputs the probability that X will $= x$ for any input x, and it (similarly) satisfies $\sum_{-\infty}^{x} f_X(t)\,\mathrm{d}t = F_X(x)$.)

Note that $\sum_{-\infty}^{\infty} f_X(t)\,\mathrm{d}t = 1 = \int_{-\infty}^{\infty} f_X(t)\,\mathrm{d}t$.

For continuous random variables:

1. $p(X = x) = 0$
2. $p(X \in A) = \int_A f_X(t)\,\mathrm{d}t$
3. $\frac{d}{dx} F_X(x) = f_X(x)$
4. $p(a \leq X \leq b) = p(a < X < b)$
5. This is all just simple calculus stuff.

2.3 Moments: mean and variance

A moment is a sort of measure of a set of points. The n^{th} moment of a real valued continuous function $f(x)$ about a real value c is:

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx \tag{2.4}$$

A central moment is a moment with c equal to a variable's mean, $E(X)$. Generally, though, c=0.

So, the **expected value** or mean of a function $f(x)$ is the first moment of x, μ_1, equal to:

$\int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$ ... if x is continuous
$\sum_x x f_X(x)$ ... if x is discrete

For integer n, the n^{th} moment of X is $E[X^n]$, and the n^{th} central moment of X is $E[(X - E[X])^n]$.

The **variance** of a random variable is its second central moment, $E[(X - E[X])^2]$. In Bayesian statistics, the variance of an unknown parameter is the **uncertainty in belief** of its value.

$$E[aX + b] = aE[X] + b \tag{2.5}$$

$$Var(aX + b) = a^2 Var(X) \tag{2.6}$$

2.4 Likelihood Function

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ denote data and θ denote the unknown parameter(s) of the distribution that defines X.

Then the **likelihood function** is the joint pmf/pdf of the data, given the parameters:

$$f(\mathbf{X}|\theta) = \prod_{i=1}^{n} f(X_i|\theta) \tag{2.7}$$

But it's often handy to work with the **log likelihood**:

$$log(f(\mathbf{X}|\theta)) = \sum_{i=1}^{n} log(f(X_i|\theta)) \tag{2.8}$$

The maximum likelihood estimator (MLE), θ, maximizes $f(\mathbf{X}|\theta)$ to find an estimate of $\hat{\theta}$: the MLE of θ.

The MLE is asymptotically normal as $n \to \infty$, with mean 0.

2.5 Joint distributions

(X, Y) is a discrete bivariate random vector if it takes a countable number of possible values. Then $f_{X,Y}(x, y) = p(X = x, Y = y)$ is the **joint probability mass function** of (X, Y), and the **marginal pmf** of X can be computed as

$$f_X(x) = \sum_y f_{X,Y}(x, y) \tag{2.9}$$

... where the sum is taken over the values of y such that $f_{X,Y}(x, y) > 0$.

Similarly, (X, Y) is a continuous bivariate random vector if it takes an infinite number of possible values. Then $f_{X,Y}(x, y)$ is the **joint probability density function** of (X, Y) where for every A $\subset \mathbb{R}^2$:

$$p((X, Y) \in A) = \int \int_{(x,y)\in A} f_{X,Y}(x, y)dxdy \tag{2.10}$$

Then the **marginal pdf** of X can be computed as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \tag{2.11}$$

These ideas extend to n-dimensional random vectors with $n > 2$.

2.6 Conditional distributions & independence

Take a discrete random variable (X, Y): For any x such that $p(X = x) > 0$, the **conditional pmf of Y given that X = x** is defined as:

$$f_{Y|X}(y|x) = p(Y = y|X = x) = \frac{p(X = x, Y = y)}{p(X = x)} = \frac{f_{X,Y}(x, y)}{f_X(x)} \tag{2.12}$$

Continuous random variable (X, Y): For any set x such that $f_X(x) > 0$, the **conditional pdf of Y given that X = x** is defined as:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{joint(x, y)}{marginal(x)} \tag{2.13}$$

X and Y are independent if for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \tag{2.14}$$

2.7 Important Probability Distributions

Discrete: Discrete Uniform, Binomial, Poisson, Multinomial, Negative binomial, Hypergeometric
Continuous: Uniform, Exponential, (glorious) Normal, Beta, Gamma, Inverse Gamma, t

2.8 Bayes' Theorem

By the definition of conditional probability:

$$f(y|x)f(x) = f(x,y) = f(x|y)f(y) \Longrightarrow f(y|x) = \frac{f(x|y)f(y)}{f(x)} \tag{2.15}$$

Bayes Theorem is central to Bayesian statistics because it allows us to obtain a posterior for some parameter, given data $\mathbf{x}$:

$$\textit{posterior}(\theta) = f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})} \propto f(\mathbf{x}|\theta)f(\theta) \tag{2.16}$$

since the denominator, $f(x)$, does not involve θ in any way, and so is a constant.

Note that we can also rewrite $f(x)$ (by the Law of Total Probability):

$$f(y|x) = \frac{f(x|y)f(y)}{\sum_y f(x|y)f(y)} \qquad (\textit{discrete case}) \tag{2.17}$$

$$f(y|x) = \frac{f(x|y)f(y)}{\int_y f(x|y)f(y)dy} \qquad (\textit{continuous case}) \tag{2.18}$$

2.9 Vocab Review

Bayes Theorem	$f(\theta\|\mathbf{x}) = \frac{f(\mathbf{x}\|\theta)f(\theta)}{f(\mathbf{x})} \propto f(\mathbf{x}\|\theta)f(\theta)$
Conditional Probability	$f(y\|x)f(x) = f(x,y) = f(x\|y)f(y)$
Marginal Probability	$f_X(x) = \sum_y f_{X,Y}(x,y)$ OR $\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$
Likelihood Function	$f(\mathbf{X}\|\theta) = \prod_{i=1}^{n} f(X_i\|\theta)$

Okay, that's it for review. Now go drink a glass of orange juice. Or be a beast and read on immediately.

Chapter 3

The Story: Likelihoods, Priors, and Posteriors

1. **Wait, what?**
2. **Bayesian Inference**
3. **Hierarchical Models**

3.1 Wait, what?

So, we have this Bayes equation. But what the hell is $p(data|\theta), p(\theta)$, or $p(data)$, in the equation:

$$f(\theta|\mathbf{data}) = \frac{f(\mathbf{data}|\theta)f(\theta)}{f(\mathbf{data})} \propto f(\mathbf{data}|\theta)f(\theta) \tag{3.1}$$

Well,

1. $p(data|\theta)$ is simply the likelihood (the 'L' in the MLE Maximum Likelihood Estimate) of your data given θ. So, for observed data $x_i...x_n$, we have:

$$\prod_{i=1}^{n} f(x_i|\theta) \tag{3.2}$$

2. $p(data)$ is just a constant, so in practice we can just ignore it, and then make sure our resultant equation $p(\theta|data)$ sums to 1, since $p(\theta|data)$is a probability distribution. (It's a constant because it doesn't involve θ, which is the variable we're looking at in $p(\theta|data)$.)

3. $p(\theta)$ is a **prior distribution** that you may pick based on prior data or beliefs, before looking at your data X. Often, when you really don't have an idea of what θ is, you can choose a non-informative prior (more on this later - if you're a bit confused, no worries).

3.2 Bayesian Inference

In Bayesian Statistics, the story is basically this:

We have some data $\mathbf{x}$, and we want to fit a good model to it, generally either to be able to predict new future data or explain the past better. So first, we choose some model M (e.g. a binomial model for coin flips), and define our prior beliefs about the model M's parameter(s), θ (e.g. I think that the coin has a 50% chance of being fair: θ=.5, and a 25% chance each of θ being .2 or .8).

Then we calculate the posterior distribution of θ, given our prior beliefs, our chosen model, and the data $\mathbf{x}$ that we have using Bayes Rule:

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)f(\theta) \tag{3.3}$$

$$posterior \propto likelihood * prior \tag{3.4}$$

Using this posterior distribution of θ, we can now predict new values of $\mathbf{x}$: $\widetilde{\mathbf{x}}$, by determining $p(\widetilde{\mathbf{x}}|$ *info & assumptions & past data*$)$:

$$p(\widetilde{\mathbf{x}}|\mathbf{x}) = \int p(\widetilde{\mathbf{x}}|\theta)p(\theta|\mathbf{x})d\theta \tag{3.5}$$

From this, we can really make any type of inference about future data values, assuming our model is correct (e.g. confidence intervals, means, etc...). But is the model correct? That's the last part of the story. Model checking and model comparison! Basically, say you've done the above process with a few different priors, and even a few different models (perhaps a hypergeometric instead of binomial model, or with a 'hyper'-prior on the parameters

of your original prior). How do you know: 1) if a model makes sense, and is good, and 2) which of many models is best. To roughly determine which model looks best, often people use what's called a **Bayes Factor**:

$$B_{12} = \frac{p(\mathbf{x}|M_1)}{p(\mathbf{x}|M_2)} \tag{3.6}$$

Where M_i = Model i. But with many models, it's generally just 'who has the highest $p(\mathbf{x}|M_i)$'. Model checking will be discussed later on, but basically, you simulate k sets of length n values of $\widetilde{\mathbf{x}}$, calculate k test statistics (e.g. mean, sd, min, max), and check to see if the test statistic calculated from your data $\mathbf{x}$ is not an extreme value when compared to the distribution of k-test statistics that you have calculated from simulated values.

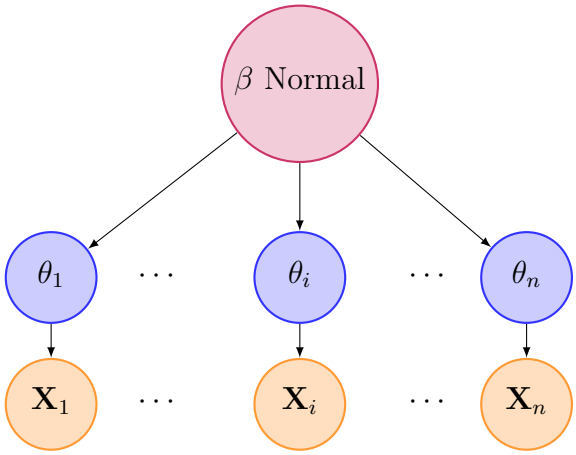
Below, you'll find a basic scenario in theory on the left, with an example of each step with real data on the right.

1. Specify a likelihood distribution and a prior *before* inspecting your data. If you let your data x influence your prior(s) in any significant way, you'll end up over fitting, and your resultant θ confidence intervals will be tighter than they should be. This is especially important when you don't have a lot of data, because your prior(s) will have a stronger influence on your posterior distribution(s).
2. Collect some data, $\mathbf{x}$.
3. Compute the posterior distribution for the parameter(s) of interest, θ, given the particular $\mathbf{x}$ that you saw: $f(\theta|\mathbf{x})$.
4. Make statistical inferences about your posterior distribution, about your model, and about future predicted data (you'll see that this is quite easy once you have (3).
5. Model Check. (Posterior Predictive Checks)
6. If model checking resulted in depression and despair at your future statistical success, repeat steps 1-5 again. Possibly use Bayes Factor to compare multiple models.
7. Refuse to get a job as an investment banker.

1. Example to comeeee.... yay a.
2.
3.
4.
5.
6.

3.3 Hierarchical Models

Okay, that's great and all, but it's hard to see why this is so much more powerful than classical frequentist statistical inference. Well, one of the strengths of Bayesian Analysis is that it's not to hard to build quite complex, hierarchical models.



(1) was used for prior probabilities, and the estimation of prior hyper-parameters. (3) was used to create 0-1 polling-like data from (2). The 0-1 data was used to estimate the posterior probability of each state's vote-share, which was then used to estimate the posterior probability of Obama winning the election.

3.4 Multi-Parameter Bayesian Models

Thus far, we have seen Bayesian inference in single-parameter models. Most real problems require multiple parameters to appropriately model the data. Example: Regression problems with multiple covariates. Advantages to the Bayesian approach to inference stand out in higher-dimensional problems.

Bayesian tools allow us to marginalize over the values of unknown nuisance parameters to obtain posterior distributions over the parameter(s) of interest:

$$p(\theta_1|\mathbf{x}) = \int p(\theta_1, \theta_2|\mathbf{x}) \; d\theta_2 \tag{3.7}$$

... where $p(\theta_1, \theta_2|\mathbf{x}) \propto p(\mathbf{x}|\theta_1, \theta_2)p(\theta_1, \theta_2)$.

3.5 Vocab Review

Bayes Theorem	$f(\theta\|\mathbf{x}) = \frac{f(\mathbf{x}\|\theta)f(\theta)}{f(\mathbf{x})} \propto f(\mathbf{x}\|\theta)f(\theta)$
Conditional Probability	$f(y\|x)f(x) = f(x,y) = f(x\|y)f(y)$
Marginal Probability	$f_X(x) = \sum_y f_{X,Y}(x,y)$ OR $\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$
Likelihood Function	$f(\mathbf{X}\|\theta) = \prod_{i=1}^{n} f(X_i\|\theta)$
$p(data\|\theta)$	The likelihood of your data, given parameter(s) θ: $f(\mathbf{X}\|\theta) = \prod_{i=1}^{n} f(X_i\|\theta)$
Prior Distribution	$p(\theta)$ ... The distribution from which your (model assumes) parameters are pulled. e.g. The distribution from which θ is pulled, where the distribution from which each X_i is determined by the value(s) of θ.
Hyperparameter	α ... A parameter that defines the distribution of your prior distribution. e.g. a hyper parameter α that defines the distribution from which θ is pulled, which then defines the distribution from which each X_i is pulled (according to the model).

Chapter 4

Priors

4.1 What is a Prior?

Philosophically, what is a prior?
- non informative priors (Why can't you just use 1? In some cases, I'm not even entirely sure.)
- informative priors (Only use these when you *are* informed)

Informative Priors:
State of Knowledge Interpretation: We express all knowledge and uncertainty about θ (prior to collecting data) as if its value could come from a random draw from the prior, $p(\theta)$
Population Interpretation: The prior distribution represents a population of possible parameter values, from which the θ of current interest is drawn.

You don't have to worry excessively about getting your prior exactly right when you're going to have a lot of data, because as more data is incorporated into the equation, the prior matters less and less (the weight of the data overpowers it). In my opinion, it is good practice to overestimate the variance on your priors, if your priors are coming from your gut, as opposed to another source of data (like history). (So, e.g., you could assume a normal distribution over some prior parameter θ with variance 10, or variance 15, the latter being 'safer'.) If you don't have a lot of data, and your prior is unjustly narrowly centered around your data, you'll be over-fitting.

4.2 Conjugate Priors

So, we have our handy dandy Bayes equation:

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)f(\theta) \tag{4.1}$$

... where our data is represented by $\mathbf{x}$. And we want to know the left side of the equation, the posterior distribution. Well, with certain assumed priors and data likelihoods, this is not so difficult. Others result in scary integrals that are literally un-integratable, which you'll need a computer (or 674 pencils) to solve well. We're going to start with the nice situations first, that don't require computers. Yay, us. These easy situations arise when your prior is 'conjugate' to your data's likelihood distribution (i.e. the type - normal, exponential, beta, etc., not its parameters).

With a conjugate prior, the posterior generally just pops out of the above equation when you plug everything in and sort some things out, and that posterior distribution will be from the same type of distribution as the prior (e.g. exponential, beta, etc..., but with slightly different parameters to take into account your data, $\mathbf{x}$). Computations of the posterior with a conjugate prior are pretty easy to do.

EXAMPLE: to come. LIST OF CONJUGATE PRIORS and LIKELIHOODS to come.

Often, a distribution's conjugate prior often is flexible enough to approximate a large number of prior beliefs. As we collect even more data (under the same likelihood), the posterior can be used as the next prior and it will be conjugate! This allows for easy updates.

However, with some likelihoods, conjugacy does not exist, especially in multi-dimensional parameter spaces. Additionally, some likelihood's conjugate priors will not be able to describe your prior beliefs well, in which case you'll need to use non-conjugate priors and more advanced methods.

Note that all likelihoods from the exponential family have conjugate pri-

ors. These include: normal, exponential, gamma, beta, Bernoulli, Poisson, etc.

When you have a conjugate prior, you can literally just plug in your information into the equation ...

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)f(\theta) \tag{4.2}$$

... to get $f(\theta|\mathbf{x})$.

A prior can express specific prior information about parameters.
- That the parameter is known to fall in some subspace (e.g. some proportion in [0,1]).
- Knowledge gained from previous experiments.
- Opinions from experts in the field (generally unsubstantiated with data that can be integrated into the model).

Generally, the more data a model is given, the less influence the prior will have on the final posterior estimation. So when you have a lot of data, it's less important to perfectly represent your prior, but it *is* always extremely important to make sure your prior probabilities are non-zero at any values that your parameter θ may take!

4.3 Non-Informative Priors

A prior can also express a lack of information. When little information about a parameter θ is known before looking at the data, it's important to use a non-informative prior, whose goal is basically to have a minimal impact on the posterior - let the data speak for themselves.
- We can use high variance priors if we have little confidence about our prior beliefs.
- Non-informative priors are great if we have no prior beliefs and/or do not want to influence the analysis with a prior.

EXAMPLE: *Non-informative Prior for Binomial Likelihood*
If $\mathbf{X}|\theta \sim Binomial(n, \theta)$, then

$\theta \sim Uniform[0,1]$ is non-informative because
$p(\theta|\mathbf{X} \propto p(\mathbf{X}|\theta)p(\theta) = p(\mathbf{x}|\theta)$

Principle of Indifference: All values are equally likely a priori. The simplest type of non-informative prior is a prior which is flat across its parameter space.
An improper prior does not integrate to 1.

The problem with flat priors... (–> Jeffrey's Prior).

- transformations of random variables

4.4 Well, Shit

In the previous section, we were able to use handy priors and likelihoods to calculate posterior distributions by hand. But - come on, it's usually not going to be the case that the data and story that you're looking at are represented ideally by a conjugate prior and likelihood pair.

Solution? Glorious computers, that's what.

Chapter 5

Grid Approximation

So, faced with both a computer, and an ugly non-conjugate problem, or a conjugate problem plus laziness, grid approximation is the most intuitive thing to do. It goes like this: instead of having an equation representing the prior probability for any value of θ, you can have a vector of actual θ values, and corresponding probability vector of the same length, where each element i in this probability vector is the probability that element i in the θ vector will occur, of all the θ given.

EXAMPLE: to come.

So. That's awesome and simple and intuitive. My kind of thing! But, wait, theeeeere's less! Grid approximation does not do so well in high dimensional problems. Say you have ten different parameters that affect your final posterior distribution: in order to maintain grid close-ness in your ten-dimensional sample space, the fineness of your grids will have to grow exponentially! If your initial parameter was defined over $[0, 1]$ and you had 100 equally spaced samples, to keep that fineness in the ten-dimensional parameter space, you'd now need 100^{10} samples! $(100,000,000,000,000,000,000,$ or $1e + 20.)$

Another problem is non-smooth distributions. If your priors are particularly bumpy, grid approximation might not work well.

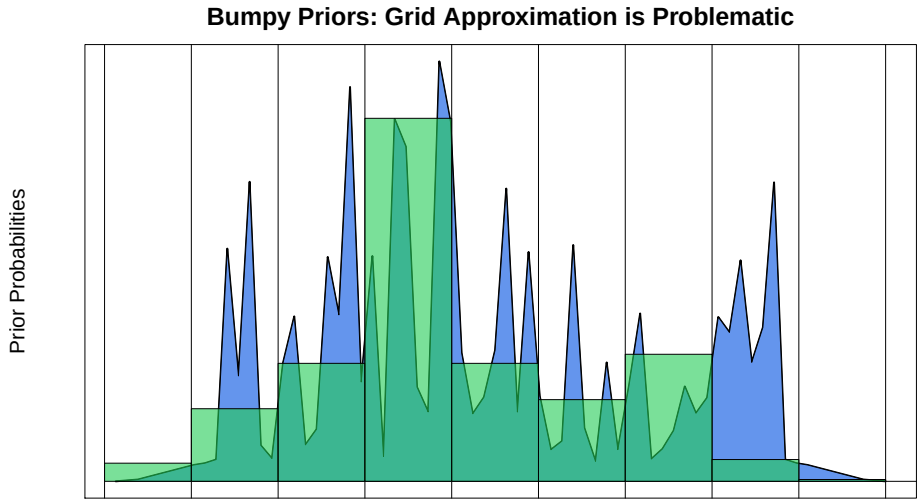


Figure 5.1: An exaggeration of how a 'bumpy' distribution may not be well estimated by a grid approximation.

But, worry not - there are yet cooler methods of estimating the posterior distribution of a parameter θ with good 'ol computers.

Chapter 6

Model Checking & Model Comparison

Chapter 7

Bayesian Sampling

This chapter deals with with situations when it's either impossible or just not easy to calculate a posterior distribution, and for whatever reason (high dimensionality, bumpy priors, or needed accuracy) grid approximation is not optimal. The following sections outline various algorithms and methods to sample from such posterior distributions.

7.1 Rejection Sampling

Say that you have a positive function $f(\theta)$ from which we want to draw samples. Unfortunately, you don't know how to draw samples from it. But you do know how to draw samples from another density $g(\theta)$, where $g(\theta)$¿0 for each θ where $f(\theta) > 0$.

1. Find a number M such that $\frac{f(\theta)}{g(\theta)} \leq M \quad \forall\theta$.
2. Generate a number θ from $g()$.
3. Generate a number u from Uniform[0,1].
4. If $u \leq \frac{f(\theta)}{Mg(\theta)}$ then accept θ as a pull from $f()$.
5. Repeat 2-4 until you have enough samples!

Note that, for efficiency's sake, you want f and Mg to be as close as possible, so you end up rejecting as few θ as possible. One easy approach is

to just set $g(\theta)$ equal to your prior, $p(\theta)$, so that:

$$\frac{f(\theta)}{g(\theta)} = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\theta)} = p(\mathbf{x}|\theta) \leq p(\mathbf{x}|\hat{\theta}_{MLE}) = M \tag{7.1}$$

This approach is a good one, although sometimes it is inefficient if $g = Mp(\mathbf{x}$ is far above $f(\theta)$ in most places. Additionally, M is not always easy to find. In which case, other approaches may be used...

7.2 Sampling Importance Resampling (SIR)

Sampling Importance Resampling is very similar to Rejection Sampling. Again, you start off with a function $f(\theta)$ that you can't pull samples from, and a function $g(\theta)$ such that $f(\theta) \leq g(\theta)$ for all θ. Then:

1. Draw a sample $\theta_1, \ldots, \theta_n$ from g.
2. For each θ, calculate $w_i = \frac{f(\theta)}{g(\theta}$
3. Calculate $q_i = \frac{w_i}{\sum_i w_i}$
4. Draw, with replacement, from $\theta_1, \ldots, \theta_n$ with probability q_i for each θ_i

The sample that you get is pretty much equivalent to a draw from f, especially when n is large and f and g are relatively 'non-bumpy'.

7.3 MCMC Methods

Markov Chain Monte Carlo methods are a set of stochastic methods that are great for sampling from difficult target distributions. They're useful in high dimensional problems (e.g. 100 different θs), when conjugacy does not hold and grid approximations are not useful. MCMC methods are iterative - they basically loop around some list of rules (algorithm) to eventually develop a posterior distribution. Three are three popular versions: Gibbs Sampling, Metropolis Hastings, and Metropolis.

7.3.1 Markov Chain

A Markov chain is a process $X_t = X_0, X_1, \ldots, X_T$ in discrete time $t = 0, 1, \ldots, T$ where $p(X_t|X_0, X_1, \ldots, X_{t-1}) = p(X_t|X_{t-1})$. In other words, X_t is only affected by the last X_{t-1}, but not by any other previous Xs. (Comparing this to Poisson's 'memoryless' attribute, you can think of Markov Chains as having very short term memories! Uhhh, *50 First Dates*, anyone?)

A Markov chain is **irreducible** if it is possible to reach all states (all possible 'X's) from any other state. So, if $p^n(j|i)$ represents the probability of getting from state i to state j in n steps, then irreducibility means that for any i and j,

$$p^n(j|i) < 0 \text{ and } p^m(i|j) < 0 \text{ for some m, n} \tag{7.2}$$

A Markov chain is **periodic** with period d if $p^n(i|j) = 0$ unless $n = kd$ for some integer k. If $d = 1$ then chain is **aperiodic**.

> **THEOREM**: A finite-state, irreducible, aperiodic Markov chain has a limiting stationary distribution: $\pi = lim_{n\to\infty} p^n(j|i)$. Then the limiting distribution, π, is reached no matter where we start the Markov chain.

The idea behind Markov Chain Monte Carlo methods is the following: you want to sample from some posterior distribution, $p(\theta|\mathbf{x})$, but it's too difficult, so instead you create a Markov Chain $\theta(t), t \in T$, with a stationary distribution $p(\theta|textbfx)$. In other words, as $t \to \infty$, then the samples given by the Markov Chain will be equivalent to samples taken from our wanted posterior distribution, $p(\theta|\mathbf{x})$.

In this scenario, we know the stationary distribution we want - $p(\theta|\mathbf{x})$ - and we want a Markov chain that will take us there: $\pi = p(\theta|\mathbf{x})$ as $t \to \infty$. Once you find that chain, all you have to do is choose some initial value(s) θ and run the chain for a large number of steps until it *converges* to π (more on how to tell if it's converged later). Then you can run it for a large number of more steps to get decent 'pull' from our wanted posterior distribution.

All MCMC methods stem from this idea. The only difference is how the

Markov chain transitions are made. Some algorithms are less prone than others to getting 'stuck' on a probability hill (local modes in the density curve of your posterior distribution), while some take longer to converge.

But, question: how the hell are you supposed to create out of thin air a Markov Chain like that? Well, the short answer is that you don't really have to. There are a lot of cool, free, and easy programs you can use (e.g. packages in R) that do it for you. But, we'll go over the theory anyways, and then some computer examples using R.

7.3.2 Gibbs Sampling

A Gibbs Sampler is an MCMC algorithm that uses conditional posterior distributions, $p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta i + 1, \ldots, \theta_n, \mathbf{x}) = p(\theta_i|rest, \mathbf{x})$, to develop the wanted posterior distribution, $p(\theta|\mathbf{x})$. All i conditional posterior distributions are cycled through for each iteration of the Gibbs sample. Say that you have three parameters $\theta = \theta_1, \theta_2, \theta_3$. To begin a Gibbs Sampler, you first have to guess initial values $\theta^0 = \theta_1^0, \theta_2^0, \theta_3^0$. Then one iteration of the Gibbs Sampler is as follows:

1. Draw θ_1^1 from $p(\theta_1|\theta_2 = \theta_2^0, \theta_3 = \theta_3^0, \mathbf{x})$
2. Draw θ_2^1 from $p(\theta_2|\theta_1 = \theta_1^0, \theta_3 = \theta_3^0, \mathbf{x})$
3. Draw θ_3^1 from $p(\theta_3|\theta_1 = \theta_1^0, \theta_2 = \theta_2^0, \mathbf{x})$

So you get $\theta^1 = \theta_1^1, \theta_2^1, \theta_3^1$, which you can use to replace θ^0 with, to then repeat steps 1-3 again, for a second iteration of the sampler. And so on and so forth, until you have converged your sampler, and later have a nice sample of θs from your wanted posterior distribution, $p(\theta|\mathbf{x})$.

Note that Gibbs Samplers can be created relatively easily for models that are either conjugate or **semi-conjugate** (?). A semi-conjugate prior is a prior that is the product of priors:

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2) \tag{7.3}$$

... each of which, conditioned on the other parameters, is conjugate. (?)

When the conditional distributions of a posterior distribution are easy to derive and sample from, Gibbs Sampling is a very efficient method, and it converges quickly. However, conditional distributions are not always easy to derive. Slower, but much more general algorithms like Metropolis Hastings work well in these scenarios.

7.3.3 Metropolis

While Gibbs Sampling requires that we can sample from full conditionals of $p(\theta)$, the Metropolis and Metropolis Hastings algorithms need only our model's likelihood and priors. (Gibbs Sampling is actually a less powerful, special case of the Metropolis Hastings MCMC method.)

The Metropolis algorithm (like Gibbs) begins with the selection of initial value(s), $theta^0 = \theta_1^0, \theta_2^0, \ldots, \theta_n^0$. Then one iteration of the Metropolis-Hastings algorithm is as follows:

1. Sample a proposal θ from a proposal (jumping) distribution, $J_t(\theta^*|\theta^{t-1})$. Note that the the proposal distribution must be symmetric, i.e.

$$J^t(\theta_a|\theta_b) = J^t(\theta_b|\theta_a) \quad \forall\ \theta_a, \theta_b, t \tag{7.4}$$

2. Compute the ratio of the posteriors:

$$r = \frac{p(\theta^*|\mathbf{x})}{p(\theta^{t-1}|\mathbf{x})} \tag{7.5}$$

3. Set the next point (θ_i in n^{th} dimensional space) in the Markov chain equal to:

$$\theta^t = \begin{cases} \theta^*, & \text{with probability } min(r, 1), \\ \theta^{t-1}, & \text{otherwise.} \end{cases} \tag{7.6}$$

In other words, the Metropolis algorithm always accepts proposals of θ^* for which $p(\theta^*|\mathbf{x}) \geq p(\theta^{t-1}|\mathbf{x})$, and we only sometimes accept proposals of θ^* for which $p(\theta^*|\mathbf{x}) < p(\theta^{t-1}|\mathbf{x})$, with acceptance probability corresponding to the probability of $p(\theta^*|\mathbf{x})$ compared to $p(\theta^{t-1}|\mathbf{x})$.

The result of this is that the Metropolis algorithm doesn't always go uphill, so it won't always get stuck in local probability maxima.

Although there are various options, the multivariate normal distribution, with a mean of θ^{t-1}, is the most often chosen for J_t.

I'm too lazy to write down the proof, for now.

7.3.4 Metropolis Hastings

The Metropolis Hastings algorithm is similar to, but even more flexible than the Metropolis algorithm. The proposal density of the Metropolis Hastings algorithm doesn't need to by symmetric, i.e. $J^t(\theta_a|\theta_b)$ need not equal $J^t(\theta_b|\theta_a)$. But to correct for the asymmetry, the ratio r is replaced by:

$$r = \frac{p(\theta^*|\mathbf{x})/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|\mathbf{x})/J_t(\theta^{t-1}|\theta^*)} \tag{7.7}$$

Allowing these asymmetric jumps can accelerate convergence.

The Metropolis and Metropolis-Hastings algorithms don't require that we can derive the conditional posteriors (as is necessary in Gibbs sampling), but it also means that Gibbs sampling, unlike the Metropolis and Metropolis-Hastings algorithms, requires that we can derive conditional posteriors. However, it has the benefit of never throwing out any proposed points of θ. As a result, Metropolis and Metropolis-Hastings algorithms tends to be much less efficient (slower).

The more narrow your proposal density is, the slower your model will converge. But the wider your proposal density is, the more crazy your proposal θs will be, leading to a lot of rejection. It's a fine line!

An often used proposal density is the multivariate normal with mean θ^{t-1}, but with correlation matrix $\sum$. The optimal proposal covariance is the true covariance of the posterior, but this is unknown, so we have to estimate it, and then adjust it over iterations. (Optimal acceptance rates of proposed values are generally in between 25 and 50%, although these percentages get smaller for higher dimensional problems). A solid estimate for $\sum$ is the inverse observed Fisher Information matrix. (?)

In general, when you're working on problems that don't take hours (or days,

months, or years!) to compute, Metropolis and Metropolis-Hastings algorithms are a great, flexible bet (and are particularly easy to use in R).

7.4 MCMC Diagnostics

So I've been talking about checking for 'convergence' this whole time without really explaining what that means. In a nutshell, as I mentioned before, a model has converged when the samples you are drawing from it are (more or less) equivalent to drawing from the true posterior distribution you are aiming for. As $t \to \infty$, this is guaranteed, but models generally converge much quicker than that, thankfully. But how can you tell? Well, there are both numeric and visual tests you can run to see if your model seems to have converged.

7.4.1 Traceplots

7.4.2 Autocorrelation Plots

7.4.3 Multiple Sequence Diagnostic

Chapter 8

Linear Models

Linear modeling is probably the most widely used statistical tool. It's also severely misunderstood by many, but hey, what to do expect? (A lot of people seem to forget that linear models are just that - *linear*! They only detect linear trends. It's always important to visualize your residuals (data - line) before making any conclusions. But, that's another rant. Additionally, the distance measure that people generally use is not some God-given formula - it's relatively arbitrary, used mostly because of mathematical convenience. But, that's neither here nor there...) The basic question to be answered is this: *how does a an 'outcome', y, vary with regards to a vector of covariates, $\boldsymbol{x} = x_1, \ldots, x_n$?*

$\mathbf{x}$ and outcome y can be discrete or continuous. The matrix of covariates, X, is called the **design matrix**. The most commonly used linear regression model goes something like this:

$$y_i \sim \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_n x_{i,n} \tag{8.1}$$

$$\text{error}_i = \epsilon_i \sim \text{Normal}(0, \sigma^2) \tag{8.2}$$

Then the likelihood for the normal linear model is then:

$$\mathbf{y}|\beta\sigma^2, \mathbf{X} \sim \text{Normal}(\mathbf{X}\beta, \sigma^2 I) \tag{8.3}$$

A handy non-informative prior on (β, σ^2) is $p(\beta, \sigma^2) \propto \sigma^{-2}$. It's not very hard to derive conditional posterior distributions of β and σ^2, which allows one to use a Gibbs sampler to determine the posterior distribution of $\mathbf{y}$!

$$\beta|\sigma^2, \mathbf{X}, \mathbf{y} \sim \text{Normal}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},\ \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right) \tag{8.4}$$

$$\sigma^2|\beta, \mathbf{X}, \mathbf{y} \sim \text{Inverse-Gamma}\left(\frac{n-k}{2}, \frac{n-k}{2}s^2\right) \tag{8.5}$$

It's also nice to be able to predict new value of $\widetilde{y}$ for a new set of covariates $\widetilde{x}_1, \ldots, \widetilde{x}_n$, i.e. draw from the posterior predictive distribution: $p(y|\mathbf{y}, \mathbf{X}, \widetilde{x}_1, \ldots, \widetilde{x}_n)$

8.1 Simple Linear Models

8.2 Hierarchical Linear Models

A **Simple Random Effects Model** is a linear model that may be described as follows:

$$y_i|\beta, \sigma^2, \mathbf{x}_i \sim \text{Normal}(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}, \sigma^2) \tag{8.6}$$

Where

$$\beta_j \sim \text{Normal}(\alpha, \sigma_\beta^2)\ for\ j \in 0, ..., p \tag{8.7}$$

Example:
$x_{i,j} = I(student\ i\ is\ in\ school\ j)$ (Where I is an indicator function.) Each school has its own effect on student performance, y_i, but the schools are all tied together via the common prior over $\beta_1, ..., \beta_p$.

A **Mixed Effects Model** is a linear model where:
$\beta_1, ..., \beta_m$ have independent non-informative priors ('fixed effects'), and
$\beta_m + 1, ..., \beta_p \sim \text{Normal}(\alpha, \sigma_\beta^2)$ ('random effects').

The term random effects refers to the deviations of randomly-selected entities (e.g., schools) from the average. When dealing with lots of predictor variables (large p), usually most of them will have no effect ($\beta = 0$)

In classical statistics, there are many methods to try to select a subset of covariates: forward / backward stepwise selection, least-angle regression, Lasso, ridge regression ,etc.

In Bayesian Regression, we can achieve similar results using hierarchical linear models by setting up a prior distribution over the β parameters that gives each parameter a high probability of being near zero and a small probability of being far from zero

Chapter 9

Bayesian Model Averaging

Chapter 10

Mixture Models

Chapter 11

Bayesian Hypothesis Testing

11.1 Frequentist (Classical) Hypothesis Testing

11.2 Bayesian Hypothesis Testing

Frequentist hypothesis testing generally follows this story: you have a likelihood (e.g. exponential or normal): $\mathbf{x}|\theta \sim p(x_1, ..., x_n|\theta)$, and you want to figure out if you should accept some original null hypothesis, H_0, or reject it in favor of another specific hypothesis, H_1.

For example, you may have:
• Simple: H_0: $\theta = 1$ versus H_1: $\theta = 2$
• One Sided: H_0: $\theta = 1$ versus H_1: $\theta \geq 2$
• Two Sided: H_0: $\theta = 1$ versus H_1: $\theta \neq 2$

The output of such a test is a p-value, which is the probability that, given the null hypothesis is true, you would have observed a test statistic (e.g. mean of data $\mathbf{x}$) as or more extreme than the test statistic of the data you observed. Very small p-values suggest that H_0 may not be correct.

What about simple tests??

This approach is good, and solid (despite what many mal-informed Bayesians

say). But it is quite inflexible, and people in the business world do have a stunning practice of falling in love with 95% significance tests and forgetting to actually use their brains. But that's another story.

Bayesian Hypothesis Testing is a little different. It's more flexible, but it's also more easily twisted and misunderstood (in my opinion).

You start off with a likelihood: $p(\mathbf{x}|\theta)$. You want to test basically the same thing as frequentist hypothesis testing: H_0: $\theta \in \Theta_0$ versus H_1: $\theta \in \Theta_1$, where H_0 and H_1 encompass all possibilities.

Here, though, we can assign prior distributions:
- $p(H_0)$
- $p(H_1)$
- $p(\theta|H_0)$
- $p(\theta|H_1)$

Then the marginal likelihood under each hypothesis is:

$$p(\mathbf{x}|H_i) = \int_\Theta p(\mathbf{x}|\theta)p(\theta|H_i)\ d\theta \quad i = 0, 1 \tag{11.1}$$

The **Bayes Factor** of H_0 to H_1 is defined as follows:

$$B_{01} = \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_1)} \tag{11.2}$$

Then the posterior probability of the null hypothesis (by Bayes) is:

$$p(H_0|\mathbf{x}) = \frac{p(H_0)p(\mathbf{x}|H_0)}{p(H_0)p(\mathbf{x}|H_0) + p(H_1)p(\mathbf{x}|H_1)} = [1 + \frac{p(H_1)}{p(H_0)}\frac{1}{B_{01}}]^{-1} = 1 - p(H_1|\mathbf{x}) \tag{11.3}$$

Since H_0 and H_1 encompass all hypothesis possibilities, the denominator in the second part of the sequence above, the sum of H_0 and H_1's joint probabilities with $\mathbf{x}$, is simply $p(\mathbf{x})$.

Conclusions may be drawn directly from the posterior odds: H_0 is 'accepted' if $p(H_0|\mathbf{x}) > p(H1|\mathbf{x})$

Instead of p-values, in Bayesian hypothesis testing, the Bayes Factor is typically reported, so that readers may place their prior beliefs about H_0 and H_1 into the $[1 + \frac{p(H_1)}{p(H_0)}\frac{1}{B_{01}}]^{-1}$ equation (though they don't get to choose their

own prior beliefs about $p(\theta|H_i)$.

As a very vague guideline (Jeffreys 1961):
1 - 3.3 $\Rightarrow$ Barely worth mentioning
3.3 - 10 $\Rightarrow$ Substantial
10 - 30 $\Rightarrow$ Strong
30 to 100 $\Rightarrow$ Very Strong
$> 100 \Rightarrow$ Decisive

But interpretation of B values will always depend on the situation!

You can also use sampling to estimate the Bayes Factor.

Chapter 12

Summary

1. Bayes Rule
2. Conjugate Priors
3. Informative and Non-Informative Priors
4. Grid Approximation
5. Hierarchical Models
6. Posterior Simulation
7. Model Checking
8. Gibbs: Metropolis Hastings
9. Gibbs: MCMC Methods
10. Hierarchical Bayesian Linear Modeling