Bayesian statistics: a concise introduction

Kevin P. Murphy
murphyk@cs.ubc.ca

Last updated October 5, 2007

1 Bayesian vs frequentist statistics

In Bayesian statistics, probability is interpreted as representing the **degree of belief** in a proposition, such as "the mean of X is 0.44", or "the polar ice cap will melt in 2020", or "the polar ice cap would have melted in 2000 if we had not...", etc. Thus we see it can be applied to reasoning about one time events (ice cap melting), counterfactual events (ice cap would have melted), as well as more "traditional" statistical questions, such as computing distributions over random variables. **Bayes rule** provides the mechanism by which **prior** beliefs are converted into **posterior** beliefs when new data arrives. (Bayes rule is sometimes called the rule of **inverse probability**.) For example, to estimate a parameter θ from data $\mathcal{D}$, one can write $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$, where $p(\theta)$ is the prior and $p(\mathcal{D}|\theta)$ is the **likelihood**. **Decision theory** can be used to decide how to convert beliefs into actions. For example, if we want to summarize our belief state with a single number (called a **point estimate**), we often use the **posterior mean** or **posterior mode**, depending on our loss function. There are various compelling arguments (see e.g., [Jay03]) that Bayesian statistics is the only consistent way to reason under uncertainty.

In **frequentist statistics** (also called **classical statistics** or **orthodox statistics**), probability is interpreted as representing long run frequencies of repeatable events. Thus it cannot be used to reason about one time events or counterfactual events. One can talk about the probability of data having a certain value, $p(\mathcal{D}|\theta)$ (this is the likelihood function), since one can imagine repeating the experiment and observing different data. But one cannot talk about the probability of a parameter having a certain value, $p(\theta|\mathcal{D})$, since parameters are assumed to be fixed (but unknown) constants, which do not have probability distributions associated with them. However, one can use decision theory to design **estimators**, which are functions that map directly from the data to point estimates of the parameters, $\hat{\theta} = f(\mathcal{D})$. These are designed to work well over repeated trials. Uncertainty estimates in frequentist statistics are based on the **sampling distribution** of the estimator, i.e., how much variation there will be in the estimate when different data is used. This is not necessarily the same as uncertainty about θ given the actual data you have observed. An estimator is said to have good **frequentist properties** it it works well in the long run (i.e., over repeated trials); however, such estimators are not necessarily optimal for any given problem.

The Bayesian approach is often criticized because the interpretation of probability in terms of beliefs seems **subjective**. In particular, the dependence on the prior (which can differ from one person to the next) is seen as "unscientific". However, all statistical modeling depends on prior assumptions (e.g., the form of the model); Bayesians just make such assumptions explicit. As I. J. Good said (quoted in [Ber85]), "The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science."

As you can see, there has been much heated debate between frequentists and Bayesians. However, these days there is a growing consensus that both approaches are useful. For example, the method of **empirical Bayes** is an approach which sets the prior based on the data. Although not strictly Bayesian, such approaches work well in practice, and have provably good frequentist properties. We will see examples of this in later chapters.

In this chapter, we will avoid philosophical arguments, and present a brief overview of the Bayesian approach to statistics. As we will see, it is intuitive and conceptually elegant. More importantly, the Bayesian approach allows us to model complex probabilistic dependencies amongst the parameters, using **hierarchical Bayesian models**. This is not possible using point estimation/ optimization methods, such as maximum likelihood, since probabilistic information can only "flow" between random variables, not between constants. (This remark will become clearer later.)

The main disadvantage of Bayesian methods is their computational expense. Some frequentist methods, such as (penalized) maximum likelihood estimation, can be thought of as simply computationally cheap approximations to full Bayesian inference. We will discuss some other simple techniques at the end, but the topic of approximate Bayesian inference is beyond the scope of this chapter.

2 Conjugate analysis

Bayes rule tells us how to combine the prior $p(\theta)$ and the likelihood $p(\mathcal{D}|\theta)$ to get the posterior $p(\theta|\mathcal{D})$:

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})}. \tag{1}$$

where the normalizing constant is

$$p(\mathcal{D}) = \int p(\theta)p(\mathcal{D}|\theta)d\theta \tag{2}$$

This is often computationally difficult to compute. However, when the prior has a certain nice mathematical form, we can work out the answer in closed form. In particular, we say a prior is **conjugate** to a likelihood if, when multiplied together, the posterior has the same functional form as the prior. (The prior is called a **natural conjugate prior** if it has the same functional form as the likelihood.) If the prior is conjugate, then the model is **closed under Bayesian updating**, which lets us easily perform sequential (recursive) updating. Below we will see some examples that should make these concepts clearer.

2.1 The beta-binomial model

Let us start out introduction to Bayesian statistics by looking at a simple example: analysing coin tosses. Suppose we toss a coin N times, and observe $x \in \{0, 1, \ldots, n\}$ heads. The probability of this happening is given by the **binomial** distribution:

$$p(x|\theta) = Bin(x|\theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{3}$$

where θ is the probability of heads. This is called the **likelihood** of the data x given unknown parameter θ.

A closely related distribution is the **Bernoulli** distribution, which is a special case of the binomial when $n = 1$ (so $x \in \{0, 1\}$).

$$p(x|\theta) = Ber(x|\theta) = \theta^x (1-\theta)^{1-x} = \theta^{I(x=1)}(1-\theta)^{I(x=0)} \tag{4}$$

Suppose we toss the coin n times; let $D = (x_1, \ldots, x_n)$ represent the sequence of heads/ tails. The likelihood of generating this data sequence (assuming independent coin tosses) is

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n} \theta^{I(x_i=1)}(1-\theta)^{I(x_i=0)} = \theta^{N_1}(1-\theta)^{N_0} \tag{5}$$

where $N_1 = \sum_i I(x_i = 1)$ is the number of heads and $N_0 = \sum_i I(x_i = 0)$ is the number of tails. Thus the difference between the binomial likelihood (which is suitable for modeling *count* data, $x \in \{0, \ldots, n\}$) and the bernoulli likelihood (which is suitable for modeling *binary* data, $x_i \in \{0, 1\}$) is just the $\binom{n}{x}$ term. Since this is a constant with respect to θ, we can drop it from the likelihood function. Thus most of the following analysis applies to both situations, although we shall focus on the Bernoulli case, since we will often be interested in modeling sequences of bits (binary data).

2.1.1 Problems with the MLE

In frequentist statistics, one estimates θ by constructing estimators, such as the **maximum likelihood estimate** (MLE), which in this case is just the empirical fraction of heads:

$$\hat{\theta} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N} \tag{6}$$

One problem with the MLE is that it can **overfit** when the sample size is small. For example, suppose we have seen 3 tails out of 3 trials. Then the MLE estimates that the probability of heads is zero:

$$\hat{\theta} = \frac{0}{0+3} = 0 \tag{7}$$

In this context, this problem is called the **sparse data** problem: if we fail to see something in the training set, we predict that it can never happen in the future, which seems a little extreme. To consider another example, suppose we have seen 3 white swans and 0 black swans; can we infer all swans are white? No! On visiting Australia, we may encounter a black swan. This is called the **black swan paradox**, and is an example of the famous **problem of induction** in philosophy. Below we will see how a Bayesian approach solves this problem.

2.1.2 Prior

Since the Binomial likelihood has the form

$$p(D|\theta) \quad \propto \quad [\theta^{N_1}(1-\theta)^{N_0}] \tag{8}$$

we see that the natural conjugate prior has the form of a **beta distribution**

$$p(\theta|\alpha_1, \alpha_0) = \text{Beta}(\theta|\alpha_1, \alpha_0) = \frac{1}{B(\alpha_1, \alpha_0)}\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \tag{9}$$

where $B(\alpha_1, \alpha_0)$ is the **beta function**, defined as

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{10}$$

and the gamma function is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du \tag{11}$$

Note that $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$. Also, for integers, $\Gamma(x+1) = x!$. Note also that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. The normalization constant $1/B(\alpha_0, \alpha_1)$ ensures

$$\int_0^1 \text{Beta}(x|\alpha_1, \alpha_0)dx = 1 \tag{12}$$

α_1, α_0 are called **hyperparameters**, since they are parameters of the prior; we will discuss these in more detail below. This prior is suitable since it defines a density on the $[0,1]$ interval, and $\theta \in [0,1]$.

If $x \sim Beta(\alpha_1, \alpha_0)$, then we have the following properties

$$\begin{aligned} \text{mean} &= \frac{\alpha_1}{\alpha_1 + \alpha_0} && (13) \\ \text{mode} &= \frac{\alpha_1 - 1}{\alpha_1 + \alpha_0 - 2},\ \alpha_0 + \alpha_1 > 2 && (14) \\ \text{Var} &= \frac{\alpha_1\alpha_0}{(\alpha_1+\alpha_0)^2(\alpha_1+\alpha_0+1)},\ \ \alpha_0 + \alpha_1 > 1 && (15) \end{aligned}$$

See Figure 1 for plots of some beta distributions. Notice that the mode of the distribution is not unique unless $\alpha_0 + \alpha_1 > 1$. For example, if $\alpha_0 = \alpha_1 = 1$, we get the uniform distribution, and if α_0 and α_1 are both less than 1, we get a bimodal distribution with "spikes" at 0 and 1. We require $\alpha_0 > 0$ and $\alpha_1 >$ to ensure the distribution is integrable (i.e., to ensure $B(\alpha_1, \alpha_0)$ exists.

To set the hyper parameters of the beta distribution, suppose your prior is that the probability of heads should be about p, and you believe this prior with strength equivalent to about N samples. Then you just solve the following equations for α_1, α_0:

$$\begin{aligned} p &= \frac{\alpha_1}{\alpha_1 + \alpha_0} && (16) \\ N &= \alpha_1 + \alpha_0 && (17) \end{aligned}$$

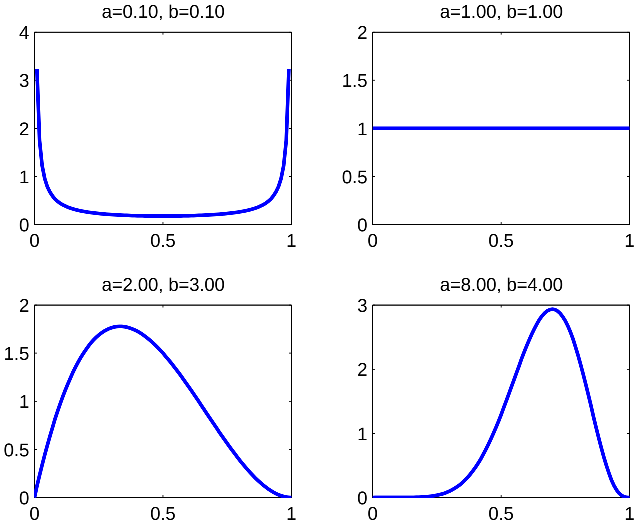


Figure 1: Some beta $Be(a, b)$ distributions. This figure was produced by `betaDistPlot`.

So we find the very intuitive result that we set α_1 to the expected number of heads, $\alpha_1 = Np$, and α_0 to the expected number of tails, $\alpha_0 = N - Np$. In other words, we can interpret the hyper-parameters of the prior in terms of **virtual data** or **fictitious data**.

2.1.3 Posterior

Multiplying prior and likelihood yields the posterior:

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) && (18) \\ &= [\theta^{N_1}(1-\theta)^{N_0}][\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}] && (19) \\ &= \theta^{N_1+\alpha_1-1}(1-\theta)^{N_0+\alpha_0-1} && (20) \\ &\propto \text{Beta}(\theta|N_1+\alpha_1, N_0+\alpha_0) && (21) \end{aligned}$$

We see that the hyper-parameters play a role analogous to N_1 and N_0, so they can be thought of as “virtual” heads/tails; they are often called **pseudo counts**. $\alpha = \alpha_1 + \alpha_0$ is called the **effective sample size** (strength) of the prior, and plays a role analogous to $N = N_1 + N_0$. The posterior is another Beta distribution with updated parameters. For example, suppose we start with $Beta(\theta|\alpha_1 = 2, \alpha_0 = 2)$ and observe $x = 1$, so $N_1 = 1, N_0 = 0$; then the posterior is $Beta(\theta|\alpha_1 = 3, \alpha_0 = 2)$. So the mean shifts from $E[\theta] = 2/4$ to $E[\theta|D] = 3/5$. We can plot the prior and posterior, as in Figure 2. We can continue to sequentially update the distribution (converting prior into posterior) as more data streams in; this is useful for **online learning** and for processing large datasets, since we don’t need to store the original data.

Let us re-write the hyper-parameters of the prior in such a way that the posterior mean becomes a convex combination of the prior mean and the MLE. Let $N = N_1 + N_0$ be number of samples (observations). Let N' be the number of pseudo observations (strength of prior) and define the prior means as fractions of N':

$$\alpha_1 = N'\alpha'_1, \quad \alpha_0 = N'\alpha'_2, \qquad (22)$$

where

$$\begin{aligned} 0 < \alpha'_1, \alpha'_0 < 1 && (23) \\ \alpha'_1 + \alpha'_0 = 1 && (24) \end{aligned}$$

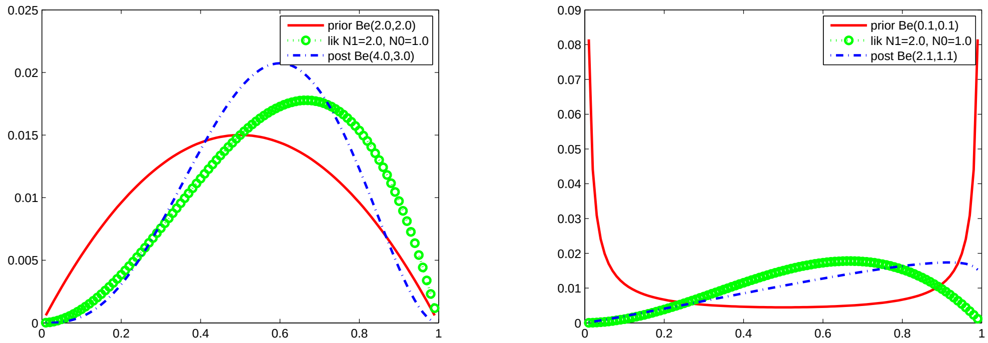


Figure 2: Updating a Beta(α_1, α_0) prior with a Bernoulli likelihood with sufficient statistics $N_1 = 2$,$N_0 = 1$ to yield a Beta$(\alpha_1 + N_1, \alpha_0 + N_0)$ posterior. (The distributions have been normalized to sum to one for plotting purposes, to make the vertical scale of the likelihood and prior comparable.) Left: $\alpha_1 = \alpha_0 = 2$. The posterior mean is shifted slightly leftwards away from the MLE of $2/3$ towards the prior mean of $2/2$. The posterior is also narrower than the prior. Right: $\alpha_1 = \alpha_0 = 0.1$. The posterior mean is strongly shifted to the right, since the prior encodes our belief that the coin is biased towards heads or tails. Figure made by `betaDistPlot2`.

Thus our new prior is

$$p(\theta) = Beta(N'\alpha_1', N'\alpha_0') = Beta(\alpha_1, \alpha_0) \tag{25}$$

where N' is the strength of our prior, and α_1' and α_0' are fractions. Then posterior mean is a **convex combination** of the prior mean and the MLE

$$\begin{aligned} E[\theta|\alpha_1, \alpha_0, N_1, N_0] &= \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_0 + N_0} && (26) \\ &= \frac{N'\alpha_1' + N_1}{N + N'} && (27) \\ &= \frac{N'}{N + N'}\alpha_1' + \frac{N}{N + N'}\frac{N_1}{N} && (28) \\ &= w\alpha_1' + (1 - w)\frac{N_1}{N} && (29) \end{aligned}$$

where $w = N'/(N + N')$ is the number of virtual samples relative to the total number of samples (total plus actual).

2.1.4 Posterior predictive distribution

Ultimately the only way we can be sure our beliefs are valid is if they help us predict the future well. We can compute such predictions by integrating out the parameters of the model (after all, parametric models are just a tool we use to predict observables). For a single Bernoulli trial, we have

$$\begin{aligned} p(X = 1|D) &= \int_0^1 p(X = 1|\theta) p(\theta|D) d\theta && (30) \\ &= \int_0^1 \theta \text{ Beta}(\theta|\alpha_1', \alpha_0') d\theta = E[\theta] = \frac{\alpha_1'}{\alpha_0' + \alpha_1'} && (31) \end{aligned}$$

where $\alpha_1' = \alpha_1 + N_1$ and $\alpha_0' = \alpha_0 + N_0$ are the parameters of the posterior.[1] With a uniform prior $\alpha_1 = \alpha_0 = 1$, we get **Laplace's rule of succession**

$$p(X = 1|N_1, N_0) = \frac{N_1 + 1}{N_1 + N_0 + 2} \tag{32}$$

[1] We are redefining α_1' and α_0' from their previous role as rescaled parameters of the prior.

This avoids the sparse data problem we encountered earlier.

If we have a large sample size, the posterior will converge to a point centered on the MLE:

$$p(\theta|D) \rightarrow \delta(\theta - \hat{\theta}_{mle}) \tag{33}$$

In this case, the posterior predictive distribution can be gotten by simply using a **plug-in estimate**

$$p(X|D) = \int p(X|\theta)p(\theta|D)d\theta \approx \int p(X|\theta)\delta(\theta - \hat{\theta})d\theta = p(X|\hat{\theta}) \tag{34}$$

In the case of the Beta-Bernoulli model, if we plug in the posterior mean estimate, $\hat{\theta} = E[\theta|D]$, we get the same result as the exact posterior predictive density:

$$p(X = 1|N_1, N_0) = p(X = 1|\hat{\theta}^{mean}) = \frac{\alpha_1'}{\alpha_0' + \alpha_1'} \tag{35}$$

2.1.5 Marginal likelihood

The **marginal likelihood** is the expected value of the likelihood, where the expectations are with respect to the prior:

$$p(D) \quad \stackrel{\text{def}}{=} \quad \int p(\theta)p(D|\theta)d\theta \tag{36}$$

It is called *marginal* likelihood because we are marginalizing out θ. In Section 6.1, we will see that the marginal likelihood forms the basis of **Bayesian model selection**.

Let us know discuss how to compute $p(D)$. Since we know $p(\theta|D) = \text{Beta}(\theta|\alpha_1', \alpha_0')$, where $\alpha_1' = \alpha_1 + N_1$ and $\alpha_0' = \alpha_0 + N_0$, we know the normalization constant of the posterior. Hence

$$\begin{aligned}
p(\theta|D) &= \frac{p(\theta)p(D|\theta)}{p(D)} && (37)\\
&= \frac{1}{p(D)}\left[\frac{1}{B(\alpha_1, \alpha_0)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}\right]\left[\theta^{N_1}(1-\theta)^{N_0}\right] && (38)\\
&= \frac{1}{p(D)}\frac{1}{B(\alpha_1, \alpha_0)}\left[\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}\theta^{N_1}(1-\theta)^{N_0}\right] && (39)\\
&= \frac{1}{B(\alpha_1', \alpha_0')}[\theta^{\alpha_1'-1}(1-\theta)^{\alpha_0'-1}] && (40)
\end{aligned}$$

Matching up the constant terms gives

$$\frac{1}{p(D)}\frac{1}{B(\alpha_1, \alpha_0)} = \frac{1}{B(\alpha_1', \alpha_0')} \tag{41}$$

so

$$p(D) \quad = \quad \frac{B(\alpha_1', \alpha_0')}{B(\alpha_1, \alpha_0)} \tag{42}$$

(The Beta-Binomial model has an extra factor in front.)

2.2 The Dirichlet-multinomial model

Let $X_n \sim Mult(\theta, 1)$ have K possible values. (The generalization to multiple trials, $X_n \sim Mult(\theta, M_n)$, is straightforward.) The analysis of the multinomial distribution is very similar to the Binomial case, so we summarize it here without proof.

2.2.1 Likelihood

$$p(D|\vec{\theta}) \quad = \quad \prod_{j=1}^{K} \theta_j^{N_j} \tag{43}$$

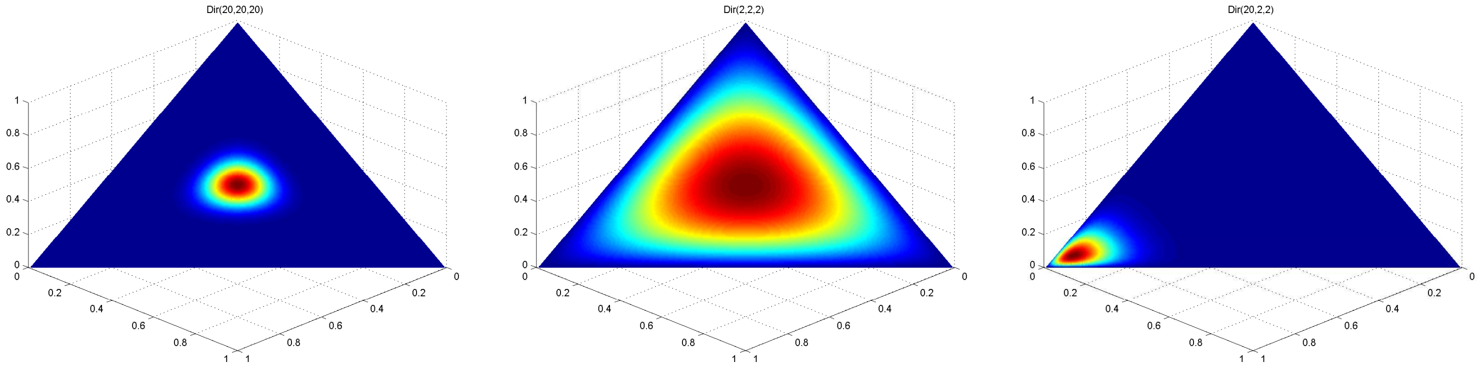


Figure 3: Some Dirichlet distributions in 3D, defined over the 3D simplex. (Such points satisfy $0 \leq x_k \leq 1$ and $\sum_{k=1}^{3} x_k = 1$.) If we put a lot of prior mass on one of the components, we select out one of the vertices. Figure produced by `dirichletPlot3d`.

2.2.2 Prior

The Beta generalized to K states is called the **Dirichlet** distribution:

$$p(\theta|\vec{\alpha}) = Dir(\theta|\vec{\alpha}) = \frac{1}{Z(\vec{\alpha})} \cdot \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1} I(\sum_k \theta_k = 1) \tag{44}$$

$$Z_{Dir}(\vec{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \tag{45}$$

Here the term $I(\sum_k \theta_k = 1)$ just ensures the probabilities sum to one, i.e., they lie on the **simplex**. See Figure 3 for some examples of Dirichlet distributions.

If $x \sim Dir(x|\alpha_1, \ldots, \alpha_K)$, then we have these properties

$$E[x_k] = \frac{\alpha_k}{\alpha} \tag{46}$$

$$\text{mode}[x_k] = \frac{\alpha_k - 1}{\alpha - K} \tag{47}$$

$$\text{Var}\ [x_k] = \frac{\alpha_k(\alpha - \alpha_k)}{\alpha^2(\alpha + 1)} \tag{48}$$

where $\alpha = \sum_k \alpha_k$.

2.2.3 Posterior

$$p(\theta|D, \vec{\alpha}) = Dir(\alpha_1 + N_1, \ldots, \alpha_K + N_K) = Dir(\alpha'_1, \ldots, \alpha'_K) \tag{49}$$

2.2.4 Posterior predictive

For a single new sample of a K-ary variable, we have

$$p(X = j|\mathcal{D}) = \frac{\alpha_j + N_j}{N + \sum_k \alpha_k} \tag{50}$$

2.2.5 Marginal likelihood

$$p(D) = \frac{Z_{Dir}(\vec{N} + \vec{\alpha})}{Z_{Dir}(\vec{\alpha})} = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(M + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{51}$$

2.3 Normal-normal model

Now let us consider Bayesian estimation of the mean of a univariate Gaussian, whose variance is assumed to be known. (If the variance is also unknown, we can use the normal-gamma prior: see Section 2.4.)

2.3.1 Likelihood

Let $D = (x_1, \ldots, x_n)$ be the data. The likelihood is

$$p(D|\mu, \sigma^2) = \prod_{i=1}^n p(x_i|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \tag{52}$$

Let us define the empirical mean and variance

$$\overline{x} = \frac{1}{n} \sum_{i=1}^n x_i \tag{53}$$

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2 \tag{54}$$

We can rewrite the term in the exponent as follows

$$\begin{aligned} \sum_i (x_i - \mu)^2 &= \sum_i [(x_i - \overline{x}) - (\mu - \overline{x})]^2 && (55) \\ &= \sum_i (x_i - \overline{x})^2 + \sum_i (\overline{x} - \mu)^2 - 2 \sum_i (x_i - \overline{x})(\mu - \overline{x}) && (56) \\ &= ns^2 + n(\overline{x} - \mu)^2 && (57) \end{aligned}$$

since

$$\sum_i (x_i - \overline{x})(\mu - \overline{x}) = (\mu - \overline{x}) \left((\sum_i x_i) - n\overline{x} \right) = (\mu - \overline{x})(n\overline{x} - n\overline{x}) = 0 \tag{58}$$

Hence

$$\begin{aligned} p(D|\mu, \sigma^2) &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \left[ns^2 + n(\overline{x} - \mu)^2 \right] \right) && (59) \\ &\propto \left(\frac{1}{\sigma^2} \right)^{n/2} \exp\left(-\frac{n}{2\sigma^2} (\overline{x} - \mu)^2 \right) \exp\left(-\frac{ns^2}{2\sigma^2} \right) && (60) \end{aligned}$$

If σ^2 is a constant, we can write this as

$$p(D|\mu) \propto \exp\left(-\frac{n}{2\sigma^2} (\overline{x} - \mu)^2 \right) \propto \mathcal{N}(\overline{x}|\mu, \frac{\sigma^2}{n}) \tag{61}$$

since we are free to drop constant factors in the definition of the likelihood. Thus n observations with variance σ^2 and mean $\overline{x}$ is equivalent to 1 observation $x_1 = \overline{x}$ with variance σ^2/n.

2.3.2 Prior

Since the likelihood has the form

$$p(D|\mu) \propto \exp\left(-\frac{n}{2\sigma^2} (\overline{x} - \mu)^2 \right) \propto \mathcal{N}(\overline{x}|\mu, \frac{\sigma^2}{n}) \tag{62}$$

the **natural conjugate prior** has the form

$$p(\mu) \propto \exp\left(-\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right) \propto \mathcal{N}(\mu|\mu_0, \sigma_0^2) \tag{63}$$

(Do not confuse σ_0^2, which is the variance of the prior, with σ^2, which is the variance of the observation noise.) (A natural conjugate prior is one that has the same form as the likelihood.)

2.3.3 Posterior

Hence the posterior is given by

$$
\begin{align}
p(\mu|D) &\propto p(D|\mu,\sigma)p(\mu|\mu_0,\sigma_0^2) \tag{64}\\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right] \times \exp\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right] \tag{65}\\
&= \exp\left[\frac{-1}{2\sigma^2}\sum_i(x_i^2+\mu^2-2x_i\mu)+\frac{-1}{2\sigma_0^2}(\mu^2+\mu_0^2-2\mu_0\mu)\right] \tag{66}
\end{align}
$$

Since the product of two Gaussians is a Gaussian, we will rewrite this in the form

$$
\begin{align}
p(\mu|D) &\propto \exp\left[-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)+\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i x_i}{\sigma^2}\right)-\left(\frac{\mu_0^2}{2\sigma_0^2}+\frac{\sum_i x_i^2}{2\sigma^2}\right)\right] \tag{67}\\
&\overset{\text{def}}{=} \exp\left[-\frac{1}{2\sigma_n^2}(\mu^2-2\mu\mu_n+\mu_n^2)\right] = \exp\left[-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right] \tag{68}
\end{align}
$$

Matching coefficients of μ^2, we find σ_n^2 is given by

$$
\begin{align}
\frac{-\mu^2}{2\sigma_n^2} &= \frac{-\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right) \tag{69}\\
\frac{1}{\sigma_n^2} &= \frac{1}{\sigma_0^2}+\frac{n}{\sigma^2} \tag{70}\\
\sigma_n^2 &= \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} \tag{71}
\end{align}
$$

Matching coefficients of μ we get

$$
\begin{align}
\frac{-2\mu\mu_n}{-2\sigma_n^2} &= \mu\left(\frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) \tag{72}\\
\frac{\mu_n}{\sigma_n^2} &= \frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2} \tag{73}\\
&= \frac{\sigma_0^2 n\overline{x}+\sigma^2\mu_0}{\sigma^2\sigma_0^2} \tag{74}
\end{align}
$$

Hence

$$
\mu_n = \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0+\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\overline{x} = \sigma_n^2\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{x}}{\sigma^2}\right) \tag{75}
$$

This operation of matching first and second powers of μ is called **completing the square**.

Another way to understand these results is if we work with the **precision** of a Gaussian, which is 1/variance (high precision means low variance, low precision means high variance). Let

$$
\begin{align}
\lambda &= 1/\sigma^2 \tag{76}\\
\lambda_0 &= 1/\sigma_0^2 \tag{77}\\
\lambda_n &= 1/\sigma_n^2 \tag{78}
\end{align}
$$

Then we can rewrite the posterior as

$$
\begin{align}
p(\mu|D,\lambda) &= \mathcal{N}(\mu|\mu_n,\lambda_n^{-1}) \tag{79}\\
\lambda_n &= \lambda_0+n\lambda \tag{80}\\
\mu_n &= \frac{\overline{x}n\lambda+\mu_0\lambda_0}{\lambda_n} = w\mu_{ML}+(1-w)\mu_0 \tag{81}
\end{align}
$$

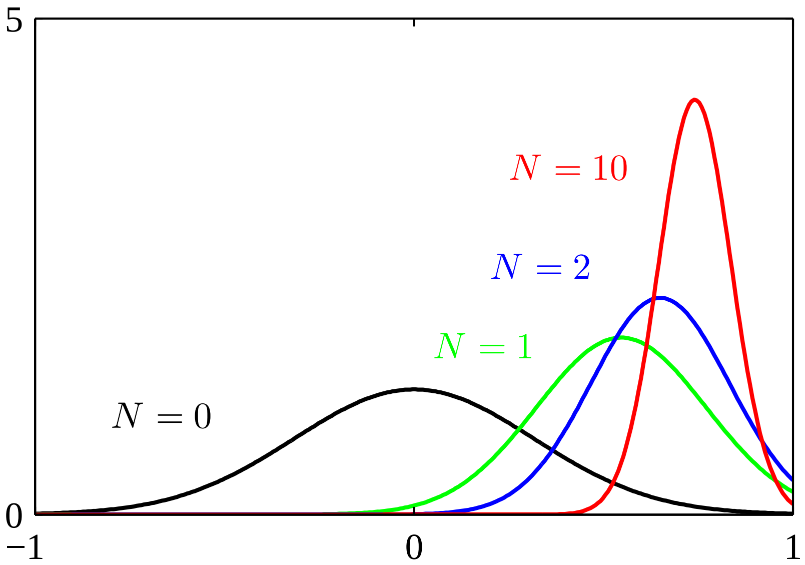


Figure 4: Sequentially updating a Gaussian mean starting with a prior centered on $\mu_0 = 0$. The true parameters are $\mu^* = 0.8$ (unknown), $(\sigma^2)^* = 0.1$ (known). Notice how the data quickly overwhelms the prior, and how the posterior becomes narrower. Source: Figure 2.12 [Bis06].

where $n\overline{x} = \sum_{i=1}^{n} x_i$ and $w = \frac{n\lambda}{\lambda_n}$. The precision of the posterior λ_n is the precision of the prior λ_0 plus one contribution of data precision λ for each observed data point. Also, we see the mean of the posterior is a convex combination of the prior and the MLE, with weights proportional to the relative precisions.

To gain further insight into these equations, consider the effect of sequentially updating our estimate of μ (see Figure 4). After observing one data point x (so $n = 1$), we have the following posterior mean

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + \sigma_0^2}\mu_0 + \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}x \tag{82}$$

$$= \mu_0 + (x - \mu_0)\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2} \tag{83}$$

$$= x - (x - \mu_0)\frac{\sigma^2}{\sigma^2 + \sigma_0^2} \tag{84}$$

The first equation is a convex combination of the prior and MLE. The second equation is the prior mean ajusted towards the data x. The third equation is the data x adjusted towads the prior mean; this is called **shrinkage**. These are all equivalent ways of expressing the tradeoff between likelihood and prior. See Figure 5 for an example.

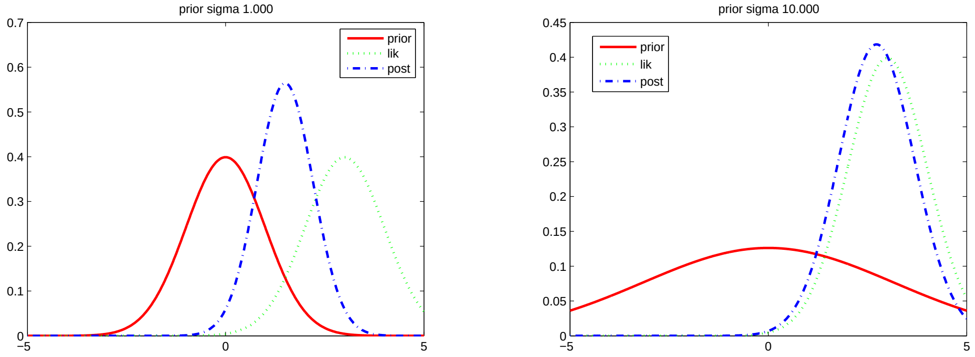
2.3.4 Posterior predictive

The posterior predictive is given by

$$p(x|D) = \int p(x|\mu)p(\mu|D)d\mu \tag{85}$$

$$= \int \mathcal{N}(x|\mu, \sigma^2)\mathcal{N}(\mu|\mu_n, \sigma_n^2)d\mu \tag{86}$$

$$= \mathcal{N}(x|\mu_n, \sigma_n^2 + \sigma^2) \tag{87}$$



(a) (b)

Figure 5: Bayesian estimation of the mean μ of a Gaussian from one sample, $\overline{x} = 3$, $n = 1$. We assume $\sigma^2 = 1$, so the likelihood is $p(x|\mu) = \mathcal{N}(3|\mu, 1)$. (Left) Strong prior $p(\mu) = \mathcal{N}(\mu|0, 1)$. The posterior is $p(\mu|D) = \mathcal{N}(\mu|1.5, 0.5)$. The posterior mean is half way between the MLE ($\overline{x} = 3$) and the prior mean ($\mu_0 = 0$), since the prior variance and observation variance are equal. Note the posterior is narrower than the prior. (Right) Weak (broad) prior $p(\mu) = \mathcal{N}(\mu|0, 10)$. Posterior is $p(\mu|D) = \mathcal{N}(\mu|2.72, 0.91)$. The posterior is very similar to the (normalized) likelihood. Figure produced by `gaussBayesDemo`.

This follows from general properties of the Gaussian distribution (see Equation 2.115 of [Bis06]). An alternative proof is to note that

$$\begin{aligned} x &= \mu + \epsilon && (88) \\ \mu &\sim \mathcal{N}(\mu_n, \sigma_n^2) && (89) \\ \epsilon &\sim \mathcal{N}(0, \sigma^2) && (90) \end{aligned}$$

where ϵ is a noise term independent of μ. Since $E[X_1 + X_2] = E[X_1] + E[X_2]$ and Var $[X_1 + X_2] =$ Var $[X_1] +$ Var $[X_2]$ if X_1, X_2 are independent, we have

$$X \sim \mathcal{N}(\mu_n, \sigma_n^2 + \sigma^2) \tag{91}$$

since we assume that the residual error is conditionally independent of the parameter. Thus the predictive variance is the uncertainty due to the observation noise σ^2 plus the uncertainty due to the parameters, σ_n^2.

2.3.5 Marginal likelihood

Writing $m = \mu_0$ and $\tau^2 = \sigma_0^2$ for the hyper-parameters, we can derive the marginal likelihood as follows:

$$\begin{aligned} \ell = p(\mathcal{D}|m, \sigma^2, \tau^2) &= \int [\prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \sigma^2)] \mathcal{N}(\mu|m, \tau^2) d\mu && (92) \\ &= \frac{\sigma}{(\sqrt{2\pi}\sigma)^n \sqrt{n\tau^2 + \sigma^2}} \exp\left(-\frac{\sum_i x_i^2}{2\sigma^2} - \frac{m^2}{2\tau^2} \right) \exp\left(\frac{\frac{\tau^2 n^2 \overline{x}^2}{\sigma^2} + \frac{\sigma^2 m^2}{\tau^2} + 2n\overline{x}m}{2(n\tau^2 + \sigma^2)} \right) && (93) \end{aligned}$$

The proof can be found in the appendix of [DMP$^+$06].

2.4 Normal-Gamma model

In this section, we consider the case where the mean and precision are both unknown. We just state the results without proofs. Derivations may be found in [Mur07]. First we introduce two useful distributions.

2.4.1 Gamma distribution

The gamma distribution is a flexible distribution for positive real valued rv's, $x > 0$. It is defined in terms of two parameters. There are two common parameterizations. This is the one used by Bishop [Bis06] (and many other

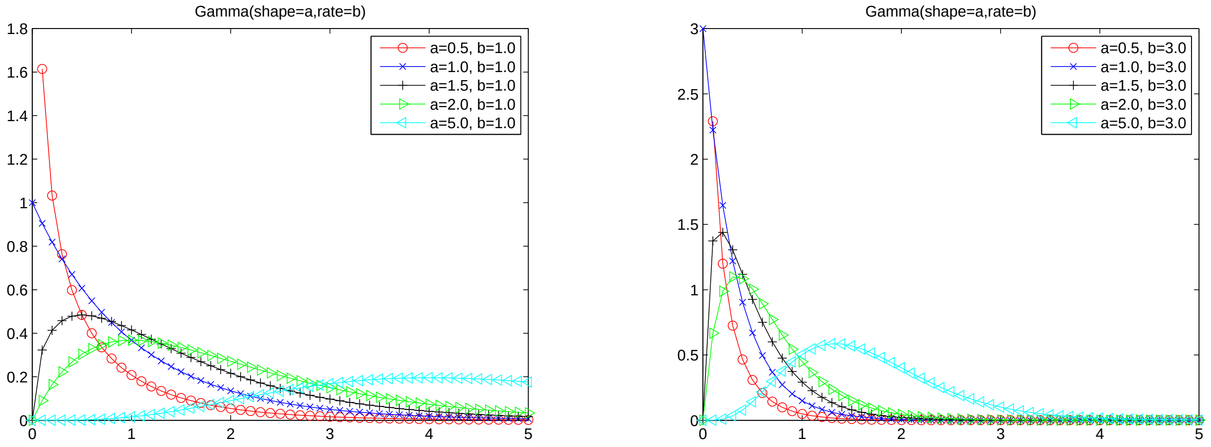


Figure 6: Some $Ga(a, b)$ distributions. If $a < 1$, the peak is at 0. As we increase b, we squeeze everything leftwards and upwards. Figures generated by `gammaDistPlot2`.

authors):

$$Ga(x|\text{shape} = a, \text{rate} = b) \quad = \quad \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}, \quad x, a, b > 0 \tag{94}$$

The second parameterization (and the one used by Matlab's `gampdf`) is

$$Ga(x|\text{shape} = \alpha, \text{scale} = \beta) \quad = \quad \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \tag{95}$$

Note that the shape parameter controls the shape; the scale parameter merely defines the measurement scale (the horizontal axis). The rate parameter is just the inverse of the scale. See Figure 6 for some examples. This distribution has the following properties (using the rate parameterization):

$$\text{mean} \quad = \quad \frac{a}{b} \tag{96}$$

$$\text{mode} \quad = \quad \frac{a-1}{b} \text{ for } a \geq 1 \tag{97}$$

$$\text{var} \quad = \quad \frac{a}{b^2} \tag{98}$$

2.4.2 Student t distribution

The generalized t-distribution is given as

$$t_\nu(x|\mu, \sigma^2) \quad = \quad c \left[1 + \frac{1}{\nu} (\frac{x-\mu}{\sigma})^2 \right]^{-(\frac{\nu+1}{2})} \tag{99}$$

$$c \quad = \quad \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}\sigma} \tag{100}$$

where c is the normalization consant. μ is the mean, $\nu > 0$ is the **degrees of freedom**, and $\sigma^2 > 0$ is the scale. (Note that the ν parameter is written as a subscript.)

The distribution has the following properties:

$$\text{mean} \quad = \quad \mu, \ \nu > 1 \tag{101}$$

$$\text{mode} \quad = \quad \mu \tag{102}$$

$$\text{var} \quad = \quad \frac{\nu\sigma^2}{(\nu-2)}, \quad \nu > 2 \tag{103}$$

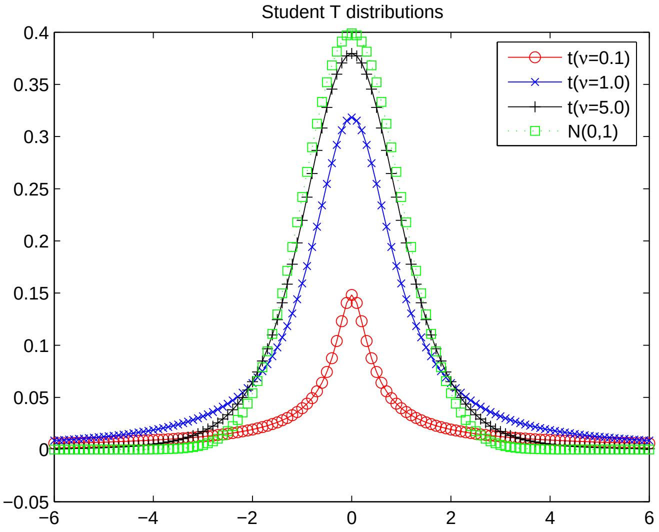


Figure 7: Student t-distributions $T_\nu(\mu, \sigma^2)$ for $\mu = 0$. The effect of σ is just to scale the horizontal axis. As $\nu\to\infty$, the distribution approaches a Gaussian. See `studentTplot`.

Note: if $x \sim t_\nu(\mu, \sigma^2)$, then

$$\frac{x-\mu}{\sigma} \sim t_\nu \tag{104}$$

which corresponds to a standard t-distribution with $\mu = 0, \sigma^2 = 1$ (Matlab's `tpdf`):

$$t_\nu(x) \quad = \quad \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + x^2/\nu)^{-(\nu+1)/2} \tag{105}$$

In Figure 7, we plot the density for different parameter values. T-distributions are like Gaussian distributions with **heavy tails**. Hence they are more robust to outliers (see Figure 8). As $\nu \to \infty$, the T approaches a Gaussian.

If $\nu = 1$, this is called a **Cauchy distribution**. This is an interesting distribution since if $X \sim Cauchy$, then $E[X]$ does not exist, since the corresponding integral diverges. Essentially this is because the tails are so heavy that samples from the distribution can get very far from the center μ.

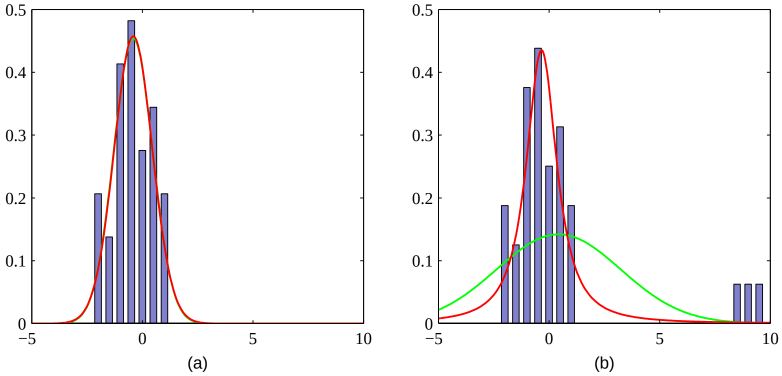


Figure 8: Fitting a Gaussian and a Student distribution to some data (left) and to some data with outliers (right). The Student distribution (red) is much less affected by outliers than the Gaussian (green). Source: [Bis06] Figure 2.16.

It can be shown that the t-distribution is like an infinite sum of Gaussians, where each Gaussian has a different variance [Arc05, p111]:

$$t_\nu(x|\mu,\lambda^{-1}) = \int_0^\infty \mathcal{N}(x|\mu,(u\lambda)^{-1}) Ga(u|\text{shape}=\frac{\nu}{2},\text{rate}=\frac{\nu}{2}) du \tag{106}$$

(See exercise 2.46 of [Bis06].)

2.5 Likelihood

The likelihood can be written in this form

$$\begin{aligned} p(D|\mu,\lambda) &= \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^n (x_i-\mu)^2\right) & (107)\\ &= \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2 + \sum_{i=1}^n (x_i-\overline{x})^2\right]\right) & (108) \end{aligned}$$

2.6 Prior

The conjugate prior is the **normal-Gamma**:

$$\begin{aligned} NG(\mu,\lambda|\mu_0,\kappa_0,\alpha_0,\beta_0) &\overset{\text{def}}{=} \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1}) Ga(\lambda|\alpha_0,\text{rate}=\beta_0) & (109)\\ &= \frac{1}{Z_{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)} \lambda^{\frac{1}{2}} \exp(-\frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2) \lambda^{\alpha_0-1} e^{-\lambda\beta_0} & (110)\\ &= \frac{1}{Z_{NG}} \lambda^{\alpha_0-\frac{1}{2}} \exp\left(-\frac{\lambda}{2}\left[\kappa_0(\mu-\mu_0)^2 + 2\beta_0\right]\right) & (111)\\ Z_{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0) &= \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}} \left(\frac{2\pi}{\kappa_0}\right)^{\frac{1}{2}} & (112) \end{aligned}$$

See Figure 9 for some plots. Here μ_0 is what we think μ is and κ_0 is how much we believe this; and β_0 is what we think $\sigma^2 = \lambda^{-1}$ is, and α_0 is how much we believe this.

2.7 Posterior

The posterior is

$$\begin{aligned} p(\mu,\lambda|D) &= NG(\mu,\lambda|\mu_n,\kappa_n,\alpha_n,\beta_n) & (113)\\ \mu_n &= \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_0+n} & (114)\\ \kappa_n &= \kappa_0 + n & (115)\\ \alpha_n &= \alpha_0 + n/2 & (116)\\ \beta_n &= \beta_0 + \tfrac{1}{2}\sum_{i=1}^n (x_i-\overline{x})^2 + \frac{\kappa_0 n(\overline{x}-\mu_0)^2}{2(\kappa_0+n)} & (117) \end{aligned}$$

We see that the posterior sum of squares, β_n, combines the prior sum of squares, β_0, the sample sum of squares, $\sum_i (x_i-\overline{x})^2$, and a term due to the discrepancy between the prior mean and sample mean. As can be seen from Figure 9, the range of probable values for μ and σ^2 can be quite large even after for moderate n. Keep this picture in mind whenever someones claims to have "fit a Gaussian" to their data.

The posterior marginals are

$$\begin{aligned} p(\lambda|D) &= Ga(\lambda|\alpha_n,\beta_n) & (118)\\ p(\mu|D) &= T_{2\alpha_n}(\mu|\mu_n,\frac{\beta_n}{\alpha_n\kappa_n}) & (119) \end{aligned}$$

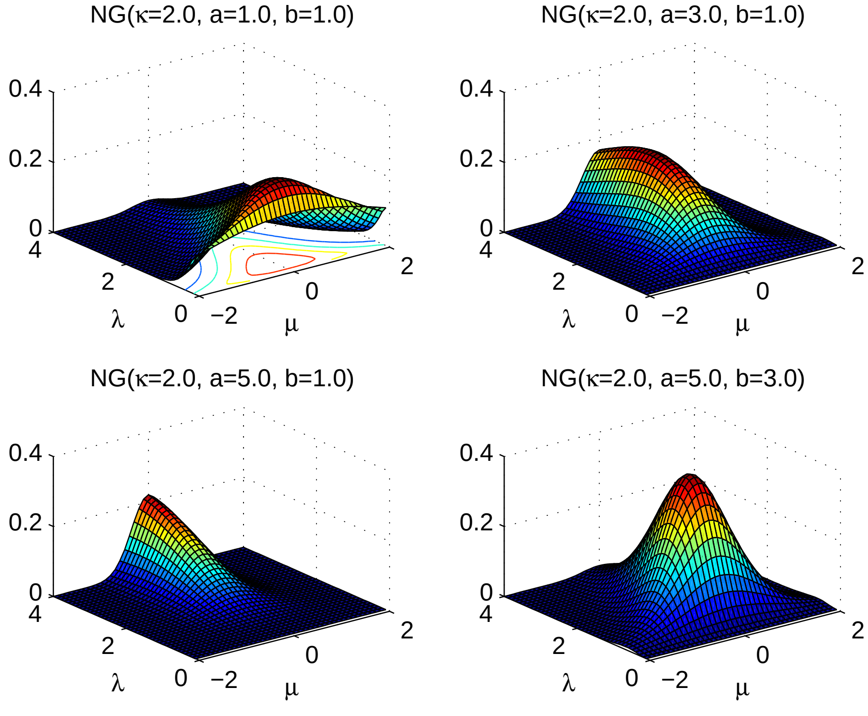


Figure 9: Some Normal-Gamma distributions. Produced by `NGplot2`.

2.8 Marginal likelihood

$$p(D) = \frac{Z_n}{Z_0}(2\pi)^{-n/2} \tag{120}$$

$$= \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)}\frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}}(\frac{\kappa_0}{\kappa_n})^{\frac{1}{2}}(2\pi)^{-n/2} \tag{121}$$

2.9 Posterior predictive

$$p(x|D) = t_{2\alpha_n}(x|\mu_n, \frac{\beta_n(\kappa_n+1)}{\alpha_n\kappa_n}) \tag{122}$$

3 Priors

Picking priors is one of the more controversial aspects of Bayesian statistics, because it is seen as “subjective”. But data analysis is never performed in a vacuum. Even babies are not born as “tabula rasa”. So we always have some kind of prior knowledge. Below we examine various issues concerning priors.

3.1 Mixture of conjugate priors

Suppose our prior beliefs about a coin are that it is biased, but either to have heads with probability near 1/3 or 2/3. We can model this using a **mixture prior**:

$$p(\theta) = \sum_{k=1}^{K}\alpha_k p_k(\theta) \tag{123}$$

which is a convex combination of K priors p_k. The terms α_k are called **mixing weights**, and satisfy $0 \leq \alpha_k \leq 1$, and $\sum_{k=1}^{K}\alpha_k = 1$. In the coin example, we may have $p_1(\theta) = Be(\theta|10, 20)$, $p_2(\theta) = Be(\theta|20, 10)$, and $\alpha_1 = \alpha_2 = 0.5$. We now show that the posterior is also a convex combination of the individual posteriors:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{124}$$

$$= \frac{p(x|\theta)\sum_k \alpha_k p_k(\theta)}{\int p(x|\theta)\sum_k \alpha_k p_k(\theta)d\theta} \tag{125}$$

$$= \frac{\sum_k \alpha_k p_k(x,\theta)}{\sum_k \alpha_k \int p_k(x,\theta)d\theta} \tag{126}$$

$$= \frac{\sum_k \alpha_k p_k(\theta|x)p_k(x)}{\sum_k \alpha_k p_k(x)} \tag{127}$$

$$= \sum_k \alpha'_k p_k(\theta|x) \tag{128}$$

where

$$\alpha'_k \propto \alpha_k p_k(x) = \alpha_k \int p_k(x|\theta)p_k(\theta)d\theta \tag{129}$$

So if the individual priors are conjugate, the posterior will be easy to analyse.

3.2 Improper priors

A distribution that does not integrate to 1 is called **improper**. If the prior is improper, the posterior will be proper, provided

$$\int p(\theta)p(D|\theta)d\theta < \infty \tag{130}$$

Typically when we have enough data (sometimes just a single point), the posterior will be proper even if the prior is not. We will see examples of improper priors below.

3.3 Jeffreys prior

Jeffreys designed a general technique for creating a certain kind of **non-informative** or **reference** prior. This can be used to perform **objective Bayesian analysis**. The key observation is that if $p(\phi)$ non-informative, then any re-parameterization of the prior, such as $\theta = h(\phi)$, should also be non-informative. Now, by the change of variables formula,

$$p_\theta(\theta) = p_\phi(\phi)|\frac{d\phi}{d\theta}| \tag{131}$$

so the prior will in general change.[2]

Let us pick

$$p_\phi(\phi) \propto (I(\phi))^{\frac{1}{2}} \tag{132}$$

where

$$I(\phi) = -E\left(\frac{d^2 \log p(X|\phi)}{d\phi^2}\right) = -E\left(\left[\frac{d \log p(X|\phi)}{d\phi}\right]^2\right) \tag{133}$$

is the **Fisher information**. Now

$$\frac{\partial \log p(x|\theta)}{\partial\theta} = \frac{\partial \log p(x|\phi)}{\partial\phi}\frac{d\phi}{d\theta} \tag{134}$$

Squareing and taking expectations over x, we have

$$\begin{aligned} I(\theta) &= -E\left(\left[\frac{d \log p(X|\theta)}{d\theta}\right]^2\right) && (135)\\ &= I(\phi)\left(\frac{d\phi}{d\theta}\right)^2 && (136)\end{aligned}$$

so we find the transformed prior is

$$\begin{aligned} p_\theta(\theta) &= p_\phi(\phi)|\frac{d\phi}{d\theta}| && (137)\\ &= (I(\phi))^{\frac{1}{2}}|\frac{d\phi}{d\theta}| && (138)\\ &\propto (I(\theta))^{\frac{1}{2}} && (139)\end{aligned}$$

In the multivariate case, we use

$$p(\theta) \propto \sqrt{\det I(\theta)} \tag{140}$$

where

$$I(\theta) = -E[\frac{\partial^2 \log p(x|\theta)}{\partial\theta_j \partial\theta_k}] \tag{141}$$

is the **Fisher information matrix**.

3.3.1 Jeffreys prior for the Binomial distribution

In the case of a Binomial distribution, the log-likelihood is

$$\log p(x|\theta) = x\log\theta + (N-x)\log(1-\theta) + \text{const} \tag{142}$$

so the Fisher information (using $E[x|\theta] = N\theta$) is

$$I(\theta) = \frac{N}{\theta(1-\theta)} \tag{143}$$

[2] Note that the fact that we need to use to the change of variables formula when we reparameterize the prior implies that MAP estimates are dependent on the parameterization. MLEs are invariant to the parameterization, since the likelihood $p(D|\theta)$ is a function, not a density, so the change of variable rule does not apply.

Hence Jeffrey's prior is

$$p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} = \frac{1}{\sqrt{\theta(1-\theta)}} \propto \text{Beta}(\tfrac{1}{2}, \tfrac{1}{2}) \tag{144}$$

One might think that, since $Beta(1,1)$ (called the **Laplace prior**) is a uniform distribution, that this would be an uniformative prior. But the posterior mean in this case is

$$E[\theta|D] = \frac{N_1 + 1}{N_1 + N_0 + 2} \tag{145}$$

whereas the MLE is $\frac{N_1}{N_1+N_0}$. Clearly by decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior. By the above argument, the most non-informative prior is

$$\lim_{c \longrightarrow 0} Beta(c, c) \tag{146}$$

which is a mixture of two equal point masses at 0 and 1 (see [ZL04] for a proof). This is also called the **Haldane prior**. (For a Gaussian, the maximum variance distribution is flattest (as we will see later), but for a Beta (because of its compact support in 0:1), the maximum variance distribution is this mixture of spikes.)

Note that the Haldane prior is an **improper prior**, in the sense that $\int Be(\theta|0,0)d\theta = \infty$. However, as long as we see at least one head and at least one tail, the posterior will be proper (integrate to 1).

So we see that there are several possible natural candidates for non-informative prior in the case of the Binomial/ Bernoulli distribution: $Be(0,0)$, $Be(\frac{1}{2}, \frac{1}{2})$ or $Be(1,1)$. Below we will see that for other kinds of parameters, such as location and scale, there is a unique definition of non-informative.

3.3.2 Jeffreys prior for a location parameter

Consider estimating the mean of a Gaussian. The log likelihood (for $N = 1$)

$$L(\mu) = \log p(x|\mu) = -\tfrac{1}{2}(x-\mu)^2/v + \text{const} \tag{147}$$

where $v = \sigma^2$ is the known variance. So

$$\begin{aligned} \frac{\partial L}{\partial \mu} &= \frac{x - \mu}{v} && (148) \\ \frac{\partial^2 L}{\partial \mu^2} &= -1/v && (149) \\ I(\mu) &= 1/v && (150) \end{aligned}$$

so the Jeffrey's prior is $p(\mu) \propto 1/\sqrt{v} = \text{const}$. We can approximate this with a conjugate prior $\mathcal{N}(\mu|\mu_0, \sigma_0^2)$ by letting $\sigma_0 \rightarrow \infty$, corresponding to a "flat" prior.

In general, if a density has the form $p(x|\mu) = f(x - \mu)$ then μ is called a **location parameter**. If the density satisfies $p(\hat{x}|\hat{\mu}) = f(\hat{x} - \hat{\mu})$, where $\hat{x} = x + c$ and $\hat{\mu} = \mu + c$, then it is called **translation invariant**. We would like our prior for the Gaussian mean to be translation invariant, so our results don't depend on the units of measurement that we use, so we require

$$\int_A^B p(\mu)d\mu = \int_{A-c}^{B-c} p(\mu)d\mu = \int_A^B p(\mu - c)d\mu \tag{151}$$

Hence $p(\mu - c) = p(\mu)$ so $p(\mu) = \text{const}$. Note that this is an improper prior.

3.3.3 Jeffreys prior for a scale parameter

The log-likelihood (using $v = \sigma^2$) is

$$L = -\tfrac{1}{2}\log v - \tfrac{1}{2}(x-\mu)^2/v + \text{const} \tag{152}$$

So

$$\frac{\partial^2 L}{\partial v^2} = \tfrac{1}{2}v^{-2} - (x-\mu)^2/v^3 \tag{153}$$

Since $E[(x-\mu)^2] = v$, we have

$$I(v) \quad = \quad -\tfrac{1}{2}v^{-2} + v/v^3 = \tfrac{1}{2}v^{-2} \tag{154}$$

so the Jeffreys prior is $p(v) \propto 1/v$.

In general, if a density has the form $p(x|\sigma) = \frac{1}{\sigma}f(x/\sigma)$ where $\sigma > 0$, then σ is called a **scale parameter**. If the density satisfies

$$p(\hat{x}|\hat{\sigma}) = \frac{1}{\hat{\sigma}}f(\frac{\hat{x}}{\hat{\sigma}}) \tag{155}$$

where $\hat{x} = cx$ and $\hat{\sigma} = c\sigma$, then it is called **scale invariant**. We would like our prior for σ^2 to be scale invariant , so we require

$$\int_A^B p(\sigma)d\sigma = \int_{A/c}^{B/c} p(\sigma)d\sigma = \int_A^B p(\frac{\sigma}{c})\frac{1}{c}d\sigma \tag{156}$$

Hence $p(\sigma) = p(\frac{\sigma}{c})\frac{1}{c}$ so $p(\sigma) \propto 1/\sigma$ will work, since then we have $\frac{1}{\sigma} = \frac{c}{\sigma}\frac{1}{c}$. Note that this is equivalent to $p(\log\sigma) \propto 1$, since

$$p(\log\sigma) = p(\sigma)\frac{d\sigma}{d\log\sigma} = (1/\sigma)\sigma = 1 \tag{157}$$

So the reference prior for the variance will be $p(\sigma^2) \propto \sigma^{-2}$. For the precision, the Jeffrey's prior is

$$p(\lambda) \propto \lambda^{-1} \tag{158}$$

which can be approximated using $Ga(\lambda|0,0)$ (but see [Gel06] for discussion).

3.3.4 Reference prior for the NG model

The reference prior is $p(\mu,\sigma^2) \propto (\sigma^2)^{-1}$ which can be modeled by $\kappa_0 = 0$, $a_0 = -1/2$, $b_0 = 0$, since then we get

$$p(\mu,\lambda) \propto \lambda^{-1} \tag{159}$$

With the reference prior, the posterior is

$$\mu_n \quad = \quad \overline{x} \tag{160}$$

$$\kappa_n \quad = \quad n \tag{161}$$

$$\alpha_n \quad = \quad (n-1)/2 \tag{162}$$

$$\beta_n \quad = \quad \tfrac{1}{2}\sum_{i=1}^n (x_i - \overline{x})^2 \tag{163}$$

The posterior marginals are

$$p(\lambda|D) \quad = \quad Ga(\lambda|\frac{n-1}{2}, \frac{\sum_i (x_i-\overline{x})^2}{2}) \tag{164}$$

$$p(\mu|D) \quad = \quad t_{n-1}(\mu|\overline{x}, \frac{\sum_i (x_i-\overline{x})^2}{n(n-1)}) \tag{165}$$

which are very closely related to the sampling distribution of the MLE. The posterior predictive is

$$p(x|D) = t_{n-1}\left(\overline{x}, \frac{(1+n)\sum_i (x_i-\overline{x})^2}{n(n-1)}\right) \tag{166}$$

4 Summaries of the posterior

The posterior $p(\theta|\mathcal{D})$ contains all the information we need for summarizing our beliefs and for making optimal decisions. However, often θ is high dimensional, so representing and computing $p(\theta|\mathcal{D})$ can be difficult. It is common to summarize the full posterior using various measures. This can be formalized using Bayesian decision theory. However, for now we just informally summarize some standard summaries.

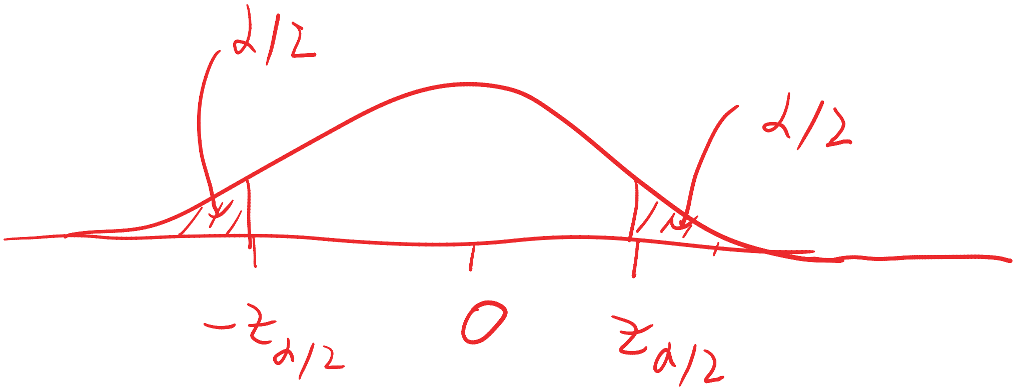


Figure 10: A $\mathcal{N}(0,1)$ distribution with the $z_{\alpha/2} = \Phi^{-1}(\alpha/2)$ cutoff points shown, where Φ is the cdf of the Gaussian. The central non shaded area contains $1-\alpha$ of the probability mass. If $\alpha = 0.05$, then $z_{\alpha/2} = 1.96 \approx 2$.

4.1 Point estimates

The most common summaries are point estimates at the mean, mode or median:

$$\begin{aligned} \hat{\theta}_{mean} &= E[\theta|\mathcal{D}] && (167) \\ \hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|\mathcal{D}) && (168) \\ \hat{\theta}_{median} &= t : p(\theta > t|\mathcal{D}) = 0.5 && (169) \end{aligned}$$

For simple distributions, these can be computed in closed form.

4.2 Bayesian credible intervals

It is common to specify a measure of uncertainty in addition to a point estimate. A α **credible interval** is a (contiguous) region C of parameter space such that $p(\theta \in C|\mathcal{D}) = \alpha$. Often we use $\alpha = 0.95$ centered on the posterior mean (see Figure 10). This is also called a **central interval**. We can find the range C using the cumulative distribution function of $p(\theta|D)$: if θ has cdf F, then $P(\theta \leq \alpha) = F^{-1}(\alpha)$ is called the α **quantile** or **critical** value of distribution F. To find interval (ℓ, u) such that $P(l \leq \theta \leq u|D) \geq \alpha$ we use $\ell = F^{-1}(\alpha/2)$ and $u = F^{-1}(1-\alpha/2)$. For example, if $p(\theta) = Be(1,1)$ and we observe $S = 47$ heads out of $N = 100$ trials, then the posterior is $p(\theta|D) = Be(a,b)$, where $a = 47+1$ and $b = 100-47+1$; a 95% posterior credible interval can be computed in Matlab as follows:

```
% betaCredibleInt
S = 47; N = 100; a = S+1; b  = (N-S)+1; alpha = 0.05;
l = betainv(alpha/2, a, b);
u = betainv(1-alpha/2, a, b);
CI = [l,u] % 0.3749    0.5673
```

An alternative summary to the central interval is to return a 95% **high posterior density** (HPD) region; this is defined as the smallest region which contains 95% of the posterior mass, which is more probable than any points outside the region. If the posterior is multimodal, the HPD may not be the same as a central posterior region: see Figure 11. However, summarizing multimodal posteriors is always difficult.

4.3 Posterior sampling

In many cases, the posterior over the quantity of interest cannot be computed in closed form. In general, we may want to compute the expected value of various features of the posterior

$$E[f(\theta)|D] = \int f(\theta)p(\theta|D) \tag{170}$$

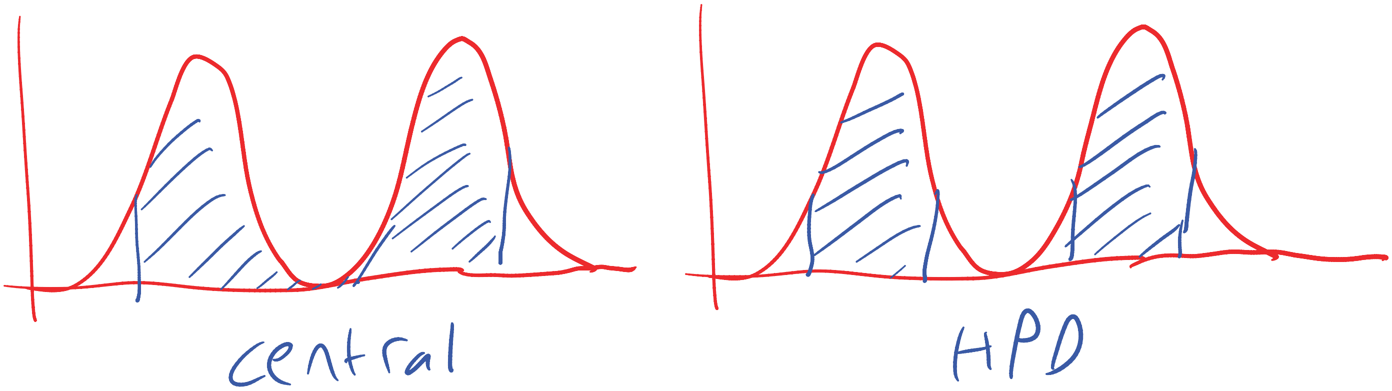


Figure 11: Central vs high posterior density intervals. Based on [GCSR04] Figure 2.2.

We can approximate such quantities using **Monte Carlo** integration:

$$E[f(\theta)|D] = \int f(\theta)p(\theta|D) \approx \frac{1}{S}\sum_{s=1}^{S} f(\theta^s) \tag{171}$$

where $\theta^s \sim p(\theta|D)$ is a sample from the posterior.

For example, suppose we toss two independent coins a bunch of times, so the posterior is

$$p(\theta_1, \theta_2|D) = Be(\theta_1|a_1, b_1)Be(\theta_2|a_2, b_2) \tag{172}$$

for some values of a_1, b_1, a_2, b_2. Suppose we want to know if coin 2 is more likely to produce heads than coin 1. This is simply

$$\begin{aligned} P(\theta_2 > \theta_1|D) &= \int_0^1 \int_0^1 I(\theta_2 > \theta_1)p(\theta_1|D)p(\theta_2|D)d\theta_1 d\theta_2 & (173)\\ &= \int_0^1 \left[\int_0^{\theta_2} p(\theta_1|D)d\theta_1\right] p(\theta_2|D)d\theta_2 & (174)\end{aligned}$$

Thus $f(\theta) = I(\theta_2 > \theta_1)$. We can approximate this as follows

$$P(\theta_2 > \theta_1) \approx \frac{1}{S}\sum_{s=1}^{S} I(\theta_2^s > \theta_1^s) \tag{175}$$

In Matlab, this becomes

```
% betaMCdemo
S = 1000;
p1 = betarnd(a1,b1,S,1);
p2 = betarnd(a2,b2,S,1);
dif = (p2-p1);
mean(dif > 0)
```

In general, **Monte Carlo integration** means approximating integrals of the form

$$E[h(X)] = I = \int h(x)p(x)dx \tag{176}$$

using

$$\hat{I} = \frac{1}{S}\sum_{s=1}^{S} h(x^s) \tag{177}$$

This can be shown to converge to the true integral as $S \rightarrow \infty$. The **standard error** of the estimate is

$$\begin{aligned} se &\overset{\text{def}}{=} \sqrt{\frac{\hat{\sigma}^2}{S}} && (178) \\ \hat{\sigma}^2 &= \frac{1}{S-1} \sum_{s=1}^{S} (h(x_s) - \hat{I})^2 && (179) \end{aligned}$$

So a $1 - \alpha$ confidence interval of I is $\hat{I} \pm z_{\alpha/2}\hat{s}$, where z_q is the q'th quantile of a standard $\mathcal{N}(0,1)$ variable. If we want to approximate the probability of a binary event, $q = P(X \in A)$, for some set A, we can use

$$\hat{q} \approx \frac{1}{S} \sum_{s=1}^{S} I(x^s \in A) \quad (180)$$

with standard error

$$\sqrt{\frac{\hat{q}(1-\hat{q})}{S}} \quad (181)$$

4.4 Posterior predictive checks

The most fundamental way to check model fit is to sample data from its posterior, $D' \sim p(x|D)$, and plot it. In cases where the data is high dimensional, and is hard to visualize, one must devise one dimensional **test statistics**, $T(D')$, and compare them to the test statistic on the actual data, $T(D)$. These statistics should measure features of interest (since it will not, in general, be possible to capture every aspect of the data). If there is a large difference between the distribution of $T(D')$ across different D' and the value of $T(D)$, it suggests the model is not a good one and/or the posterior has not been well estimated. We illustrate this below.

4.4.1 Worked example: Newcomb's speed of light data

Here we consider an example from [GCSR04]. In 1882, Newcomb measured the speed of light using a certain method and obtained the distribution in Figure 12. There are clearly two outliers in the left tails, suggesting that the distribution is not Gaussian. Let us none the less fit a Gaussian to it, using a noninformative prior. We can test our fit by sampling from the posterior:

$$x \sim p(x|D) = t_{\nu_n}(x|\mu_n, \frac{(1+\kappa_n)\sigma_n^2}{\kappa_n}) \quad (182)$$

Let D'_s be the s'th dataset of size $n = 66$ generated in this way. The histogram of D'_s for $s = 1 : 20$ is shown in Figure 13. It is clear that the model is not capable of generating the large negative examples that were seen in the real data. (We are assuming these are scientifically interesting, and not noise that we want to eliminate.)

A more formal way to test fit is to define a test statistic. Since we are interested in small values, let us use

$$T(D) = \min\{x : x \in D\} \quad (183)$$

For the real data, $T(D) = -44$, but the distribution of $T(D'_s)$ for $s = 1 : 1000$ is shown in Figure 14. It is clear that $T(D)$ is very unlikely according to our fitted model.

The code to generate these plots is shown below.

```
% newcomb.m
% Example from Gelman04 p77 - see if Newcomb's speed of light data is Gaussian

seed = 0; randn('state', seed); rand('state', seed);

% Data from http://www.stat.columbia.edu/~gelman/book/data/light.asc
D = [28 26 33 24 34 -44 27 16 40 -2  29 22 24 21 25 30 23 29 31 19 ...
     24 20 36 32 36 28 25 21 28 29  37 25 28 26 30 32 36 26 30 22 ...
     36 23 27 27 28 27 31 27 26 33  26 32 32 24 39 28 24 25 32 25 ...
      29 27 28 29 16 23];

% uninformative prior
k0 = 0; v0 = -1; s0 = 0; mu0 =0;
```

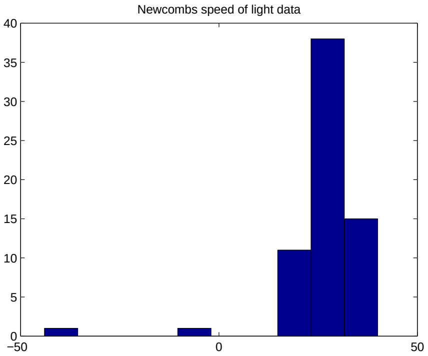


Figure 12: Histogram of Newcomb's data. We plot the measured time it takes light to travel 7442 m minus $24,800ns$.

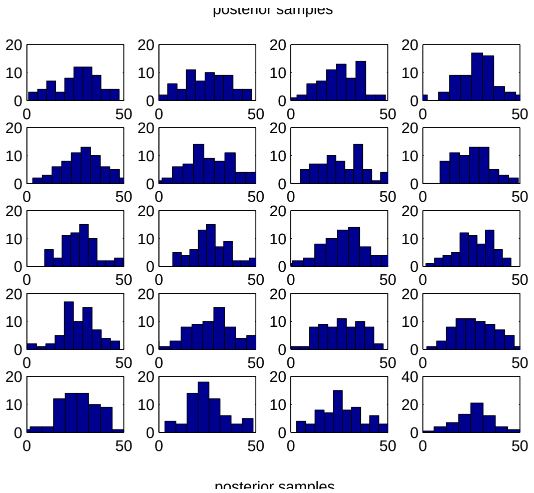


Figure 13: Histogram of data sampled from Newcomb's posterior.

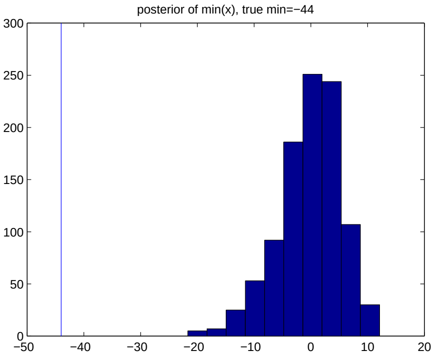


Figure 14: Histogram of test statistic on data sampled from Newcomb's posterior. The vertical line is the test statistic on the true data.

```
% suff stat
xbar = mean(D); n = length(D); s2 = mean( (D-xbar).^2);

% posterior
kn = k0+n;
mun = (k0*mu0+n*xbar)/kn;
vn = v0+n;
s2n = (v0*s0 + n*s2 + k0*mu0^2 + n*xbar^2 -kn*mun^2)/vn;

% credible interval for mu
low = mun + tinv(0.025, vn)*sqrt(s2n/kn) %23.5706
high = mun + tinv(1-0.025, vn)*sqrt(s2n/kn) %28.8537

% generate posterior samples
S = 1000;
sigma2 =  (1+kn)*s2n/kn;
rep = trnd(vn, S, n)*sqrt(sigma2) + mun;

figure(1); clf
hist(D); title('Newcombs speed of light data')

figure(1); clf
for i=1:20
  subplot(5,4,i)
  hist(rep(i,:))
  set(gca,'xlim',[0 50])
  %title(sprintf('synth %d', i))
end
suplabel('posterior samples', 't')

% compute distribution of test statistic
test=inline('min(x)','x');
for s=1:S
  testVal(s) = test(rep(s,:));
end
testValTrue = test(D);
figure(2);clf
hist(testVal);
title(sprintf('posterior of min(%s), true min=%d', 'x', testValTrue))
hold on
line([testValTrue, testValTrue], get(gca,'ylim'))
```

5 Approximate inference

Often it is difficult to compute $p(\theta|D)$ in closed form. One approach is to try to approximate the posterior using a simpler kind of parametric distribution, such as a Gaussian. Another approach is to represent the posterior implicitly, in terms of a set of samples $\theta^s \sim p(\theta|D)$. Note that generating such samples can be difficult, but once we have them, it becomes easy to compute arbitrary posterior features $E[f(\theta)]$ as we saw above. This topic is beyond the scope of this chapter. However, we introduce one very simple approach below.

5.1 Laplace approximation

The **Laplace approximation** is to approximate $p(\theta|D)$ by a multivariate Gaussian. (In physics, this is called a **saddle point approximation**.) See Figure 15 for an example.

Suppose $\theta \in \mathbb{R}^d$. Let $p(\theta|D) = \frac{1}{Z}f(\theta)$. Performing a Taylor series expansion around a mode θ_0 we get

$$\ln f(\theta) \approx \ln f(\theta_0) + \frac{1}{2}(\theta - \theta_0)^T H(\theta - \theta_0) \tag{184}$$

where

$$H \overset{\text{def}}{=} \frac{\partial^2 \log f(\theta)}{\partial\theta\partial\theta^T} \, |_{\theta=\theta_0} \tag{185}$$

is the Hessian of $\log p(\theta)$. Hence

$$\hat{f}(\theta) = f(\theta_0) \exp\left[-\frac{1}{2}(\theta - \theta_0)^T C^{-1}(\theta - \theta_0)\right] \tag{186}$$

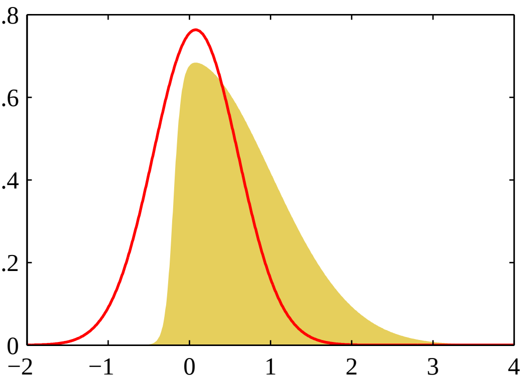


Figure 15: Laplace approximation (red) to the function $\exp(-x^2/2)\sigma(20x+4)$. From [Bis06] Figure 4.14.

where $C = -H^{-1}$. So

$$\hat{p}(\theta) = \frac{1}{\hat{Z}}\hat{f}(\theta) = \mathcal{N}(\theta|\theta_0, C) \tag{187}$$

$$\hat{Z} = \int \hat{f}(\theta)d\theta = f(\theta_0)(2\pi)^{d/2}|C|^{\frac{1}{2}} \tag{188}$$

We will use the term $\hat{Z}$ when we derive the BIC score below.

Since the Laplace approximation assumes the posterior is approximately Gaussian, it is often necessary to transform the parameters so that this is a reasonable assumption. For example, when estimating a positive term, we can take logs. We will see an example of this below.

5.1.1 Worked example

Let us consider the following example from [GCSR04, p102]. Consider estimating the mean and variance of a 1D Gaussian using a non-informative prior $p(\mu, \log\sigma) \propto 1$. Define

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{189}$$

The log posterior is given by

$$\log p(\mu, \log\sigma|D) = \text{const} - n\log\sigma - \frac{1}{2\sigma^2}[ns^2 + n(\overline{y}-\mu)^2] \tag{190}$$

For brevity, let $\lambda = \log\sigma$. The first derivatives are

$$\frac{\partial}{\partial\mu}\log p(\mu,\lambda|D) = \frac{n(\overline{y}-\mu)}{\sigma^2} \tag{191}$$

$$\frac{\partial}{\partial\lambda}\log p(\mu,\lambda|D) = -n + \frac{ns^2 + n(\overline{y}-\mu)^2}{\sigma^2} \tag{192}$$

from which the posterior mode is easily seen to be

$$\hat{\mu} = \overline{y} \tag{193}$$

$$\log\hat{\sigma} = \tfrac{1}{2}\log\left(\frac{n}{n}s^2\right) \tag{194}$$

The Hessian matrix is given by

$$H = \begin{pmatrix} \frac{\partial^2}{\partial\mu^2}\log p(\mu,\lambda|D) & \frac{\partial^2}{\partial\mu\partial\lambda}\log p(\mu,\lambda|D) \\ \frac{\partial^2}{\partial\lambda^2}\log p(\mu,\lambda|D) & \frac{\partial^2}{\partial\mu\partial\lambda}\log p(\mu,\lambda|D) \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & -2n\frac{\overline{y}-\mu}{\sigma^2} \\ -2n\frac{\overline{y}-\mu}{\sigma^2} & -\frac{2}{\sigma^2}(ns^2 + n(\overline{y}-\mu)^2) \end{pmatrix} \tag{195}$$

Evaluating this at the mode we have

$$H|_{\hat{\theta}} \quad = \quad \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -2n \end{pmatrix} \tag{196}$$

Hence the approximate posterior is

$$p(\mu, \log \sigma | D) \approx \mathcal{N}\left(\begin{pmatrix} \overline{y} \\ \log \hat{\sigma} \end{pmatrix}, \begin{pmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{1}{2n} \end{pmatrix} \right) \tag{197}$$

6 Bayesian model selection

Suppose we have K possible models for our data. Let us write $p(D|M_i)$ to represent the data generated from model i, for $i = 1 : K$. We can express our belief in which model is correct using

$$p(M = i|D) = p(M_i|D) = \frac{p(D|M_i)p(M_i)}{p(D)} \tag{198}$$

where

$$p(D|M_i) = \int p(D|\theta, M_i)p(\theta|M_i)d\theta \tag{199}$$

is the marginal likelihood, also called the **evidence** for model M_i. We cannot use the “standard” likelihood $p(D|M_i, \theta)$ for model selection, because θ is unknown; and we cannot use $p(D|M_i, \hat{\theta}_{ML})$ since the maximum likelihood model is always the most complex one (since the most complex model can always fit the training data the best).

Notice that the normalizing constant used for parameter estimation becomes the likelihood for the next level up the modeling hierarchy:

$$p(\theta|D, M_i) = \frac{p(D|\theta, M_i)p(\theta|M_i)}{p(D|M_i)} \tag{200}$$

The term $p(M_i)$ is our prior preference for model i. Sometimes we explicitly encode a preference for simpler models, by penalizing models with many parameters, although, as we will see in Section 6.2, this is not strictly necessary.

If we just want to compare two models, we can compute their **posterior odds ratio**

$$O_{ij} = \frac{p(M_i|D)}{p(M_j|D)} = \frac{p(D|M_i)p(M_i)}{p(D|M_j)p(M_j)} \tag{201}$$

where $\frac{p(M_i)}{p(M_j)}$ is called the **prior odds ratio** and

$$BF(M_i, M_j) = \frac{p(D|M_i)}{p(D|M_j)} = \frac{p(M_i|D)}{p(M_j|D)} / \frac{p(M_i)}{p(M_j)} \tag{202}$$

is called the **Bayes factor** (posterior to prior odds ratio). For two models, we write

$$BF(1, 2) \quad = \quad \frac{P(D|H_1)}{P(D|H_2)} \tag{203}$$

This is like a **likelihood ratio**, except we integrate out the parameters. If we have O_{ij} for all pairs, we can infer the distribution over models $p(M_i|D)$ using the fact that $\sum_i p(M_i|D) = 1$. For example, for 2 models, we have

$$\begin{aligned} p(M_1|D) &= O_{12}p(M_2|D) && (204) \\ &= O_{12}(1 - p(M_1|D)) && (205) \\ &= \frac{O_{12}}{1 + O_{12}} && (206) \end{aligned}$$

6.1 Worked example: is the coin biased?

Consider the problem of determinining if a coin is biased. Let θ be the probability of heads. We want to compare two hypotheses or models, H_0 that $\theta = 0.5$, and H_1 that $\theta \neq 0.5$. In fact, since the probability that θ is exactly equal to 0.5 is zero (because $p(\theta)$ is a density function), we can let H_1 be the hypothesis that $\theta \in [0, 1]$, without worrying about excluding 0.5.

For H_0, there is no free parameter, so the marginal likelihood is

$$P(D|H_0) = 0.5^N \tag{207}$$

where N is the number of coin tosses in D. For H_1, we need to integrate out θ:

$$P(D|H_1) = \int_0^1 P(D|\theta, H_1)P(\theta|H_1)d\theta \tag{208}$$

For simplicity, let us use a $Beta(\alpha_1, \alpha_2)$ prior on θ, where $\alpha_1 = \alpha_2 = \alpha$.

Suppose, following [Mac03], that we toss a coin $N = 250$ times, and observe $N_1 = 141$ heads and $N_0 = 109$ tails. Then

$$BF(1,0) \quad = \quad \frac{P(D|H_1)}{P(D|H_0)} = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)} \frac{1}{0.5^N} \tag{209}$$

To test the **robustness** of our conclusion to our prior, we compute $BF(1, 0)$ for a range of prior strengths α. The results are shown in Figure 16. For a uniform prior, $\alpha = 1$, $\frac{P(H_1|D)}{P(H_0|D)} = 0.48$, (weakly) favoring the fair coin hypothesis H_0. At best, for $\alpha = 50$, we can make the biased hypothesis twice as likely. A Bayes factor of 2 is not evidence in favor of a hypothesis.

The code to implement Figure 16 is shown below. Note that `betaln` is the log-beta function; we must work in log domain to avoid underflow.

```
% modelSelCoinDemo
alphas = [0.37 1 2.7 7.4 20 55 148 403 1096];
Nh = 140; Nt = 110; N = Nh+Nt;
figure(1);clf
logBF = betaln(Nh+alphas, Nt+alphas) - betaln(alphas, alphas) - N*log(0.5);
plot(alphas, exp(logBF), 'o-');
```

6.2 Bayesian Occam's razor

A simple approach to model selection is to pick the one with the largest **penalized likelihood**, where the penalty is proportional to the number of parameters in the model (see Section 6.4). However, simply counting parameters is a rather blunt instrument. It turns out that, for many models, the *magnitude* of the parameters is at least as important as the number of parameters. To see why, consider linear regression, $y = \theta^T f(x) + \epsilon$, where $f(x)$ is a basis function expansion of x, such as a polynomial expansion. If $\theta_i \approx 0$ for the features i that represent higher order terms, then the function will be fairly linear, but if θ_i is large for such terms, the function will be very "wiggly". Hence the parameter prior $p(\theta|M_i)$ turns out to control model complexity as well. In the Bayesian approach, by integrating over all parameters, we are seeking a model that is good, no matter what parameters it uses. This discourages picking models that only fit the data well at a particular θ (by chance). Thus the mere act of integrating over θ will automatically pick simpler models. This is called the Bayesian **Occam's razor**. (Occam's razor says: "if two models are equally good at predicting, pick the simpler one".) In other words, even if we have no explicit penalty on complex models (so $P(M_i)$ is uniform), merely by integrating over all possible parameter values (i.e., by using $P(D|M_i) = \int P(D, \theta|M_i)d\theta$), we automatically prefer models that are not too complex (provided they fit the data well).

An overly simple model M_1 has low $P(D|M_1)$ since it has poor fit to the data. An overly complex model M_3 has lower $P(D)$ than a medium model M_2, since a complex model spreads its probability mass over more possible datasets, but this mass must sum to one (**conservation of belief**). Put another way, we trust an expert who predicts a few *specific* (and correct!) things more than an expert who predicts many things. See Figure 17.

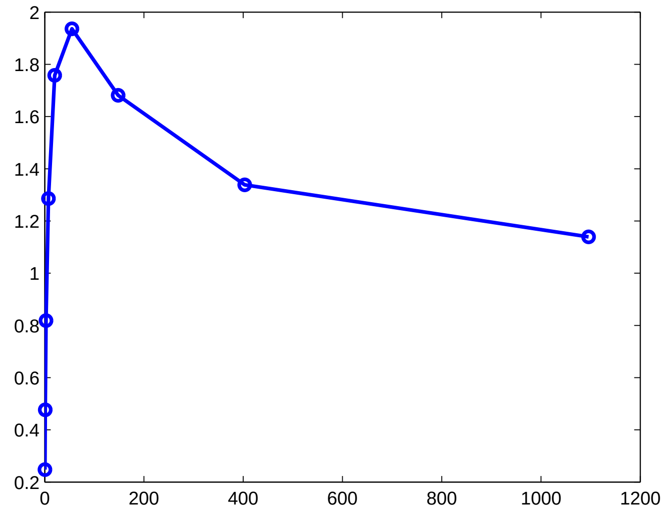


Figure 16: Bayes factor in favor of biased coin versus strenght of symmetric Beta hyperparameter. Produced by `modelSelCoinDemo.m`.

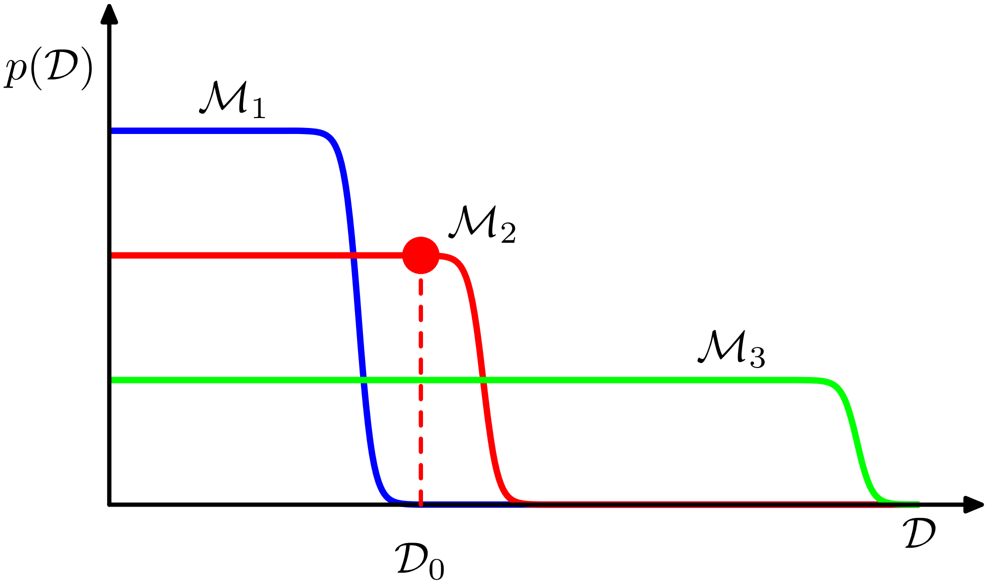


Figure 17: An illustration of the Bayesian Occam's razor. The broad (green) curve corresponds to a complex model, the narrow (blue) curve to a simple model, and the middle (red) curve is just right. Source: Figure 3.13 [Bis06].

6.2.1 Worked example: is the coin biased?

Let us consider a simple example. Consider comparing the model M_0 that a coin is unbiased, $\theta = 0.5$, to the model M_1 that says the coin may be biased, $\theta \sim Be(1,1)$. (Note that M_1 includes M_0, but only assigns infinitessimal probability mass to the event $\theta = 0.5$.) The marginal likelihood under M_0 is simply

$$p(D|M_0) = \frac{1}{2}^N \tag{210}$$

where N is the number of coin tosses. The marginal likelihood under M_1 is

$$p(D|M_1) = \int p(D|\theta)p(\theta)d\theta = \int Bin(N_1|N,\theta)Beta(\theta|1,1)d\theta = \frac{B(1+N_1, 1+N_0)}{B(1,1)} \tag{211}$$

We plot this vs the number of heads N_1 in Figure 18 (assuming $N = 5$). We see is that if we observe 2 or 3 heads, the unbiased coin hypothesis M_0 is more likely, since it is simpler (has no free parameters); but if we observe 0, 1, 4, or 5 heads, the biased coin hypothesis M_1 is more likely. It would be a **suspicious coincidence** if the coin were biased but happened to produce almost exactly 50/50 heads/tails, so we discount model M_1 for the data in the middle of the curve.

Another interesting feature of this plot is the strong probability of getting all heads or all tails under M_1. To understand this, let us use the chain rule to write

$$p(D) = p(x_{1:N}) = p(x_1)p(x_2|x_1)p(x_3|x_{1:2})\ldots \tag{212}$$

Now, the posterior predictive distribution is

$$p(X=1|D_{1:N}) = \frac{N_1 + \alpha_1}{N_1 + \alpha_1 + N_0 + \alpha_0} \overset{\text{def}}{=} \frac{N_1 + \alpha_1}{N + \alpha} \tag{213}$$

where $D_{1:N}$ is the data seen so far and $\alpha = \alpha_0 + \alpha_1$. So a sequence of all 0 heads, say, is much more likely than a sequence with 1 head or 2 heads:

$$\begin{aligned}
p(0,0,0,0,0) &= \frac{\alpha_0}{\alpha} \cdot \frac{\alpha_0+1}{\alpha+1} \cdot \frac{\alpha_0+2}{\alpha+2} \cdot \frac{\alpha_0+3}{\alpha+3} \cdot \frac{\alpha_0+4}{\alpha+4} = 0.1667 && (214)\\
p(0,0,0,0,1) &= \frac{\alpha_0}{\alpha} \cdot \frac{\alpha_0+1}{\alpha+1} \cdot \frac{\alpha_0+2}{\alpha+2} \cdot \frac{\alpha_0+3}{\alpha+3} \cdot \frac{\alpha_1}{\alpha+4} = 0.0333 && (215)\\
p(0,0,0,1,1) &= \frac{\alpha_0}{\alpha} \cdot \frac{\alpha_0+1}{\alpha+1} \cdot \frac{\alpha_0+2}{\alpha+2} \cdot \frac{\alpha_1}{\alpha+3} \cdot \frac{\alpha_1+1}{\alpha+4} = 0.0167 && (216)
\end{aligned}$$

Note that the order of the data does not matter. Also, the shape of the curve is not very sensitive to α.

The code to produce Figure 18 is shown below.

```
%joshCoins4

theta = 0.7; N = 5; alpha = 1;
alphaH = alpha; alphaT = alpha;
for i=1:(2^N)
  flips(i,:) = ind2subv(2*ones(1,N), i); % convert i to  bit vector
  Nh(i) = length(find(flips(i,:)==1));
  Nt(i) = length(find(flips(i,:)==2));
  nh = Nh(i); nt = Nt(i);
  margLik(i) = exp(betaln(alphaH+nh, alphaT+nt) - betaln(alphaH, alphaT));
end

% sort in order of number of heads
[Nh, ndx] = sort(Nh);
margLik = margLik(ndx);

figure(1); clf
hold on
p0 = (1/2)^N;
h=plot(margLik, 'o-');
h = line([0 2^N], [p0 p0]); set(h,'color','k','linewidth',3);
```

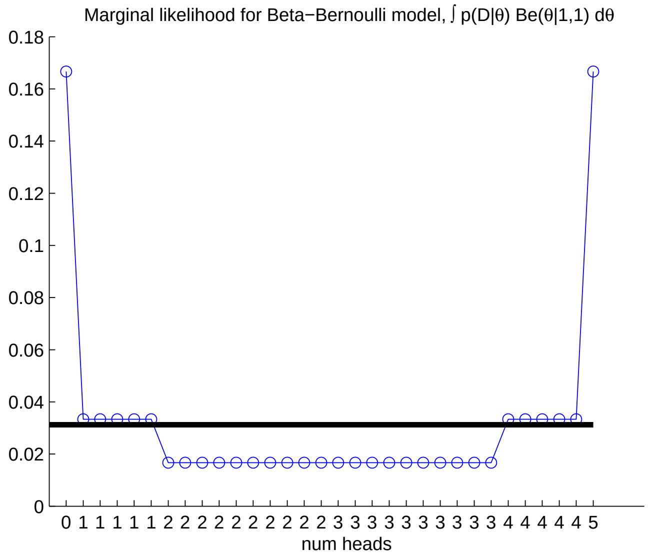


Figure 18: Marginal likelihood for different data sets under two different models: horizontal black line asserts $\theta = 0.5$; other blue line asserts only that $\theta \in [0, 1]$. Produced by `joshCoins4`.

```
set(gca,'xtick', 1:2^N)
set(gca,'xticklabel',Nh)
xlabel('num heads')
title(sprintf('Marginal likelihood for Beta-Bernoulli model, %s p(D|%s) Be(%s|1,1) d%s', ...
              '\int', '\theta', '\theta', '\theta'))
```

6.3 Lindley's paradox

Problems can arise when we use improper priors for model selection/ hypothesis testing. For example, consider testing the hypotheses $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. Let p_0 be the prior H_0, and $p_1 = 1 - p_0$ be the prior of H_1. To define the prior *density* on θ, we use the following mixture model

$$p(\theta) \quad = \quad p(\theta|H_0)p(H_0) + p(\theta|H_1)p(H_1) = \pi_0(\theta)p_1 + \pi_1(\theta)p_0 \tag{217}$$

The mixing weights p_0, p_1 are only meaningful if π_0 and π_1 are proper (normalized) density functions. In this case, the posterior is given by

$$p(H_0|x) \quad = \quad \frac{p_0 p(x|H_0)}{p_0 p(x|H_0) + p_1 p(x|H_1)} \tag{218}$$

$$= \quad \frac{p_0 \int_{\Theta_0} p(x|\theta)\pi_0(\theta)d\theta}{p_0 \int_{\Theta_0} p(x|\theta)\pi_0(\theta)d\theta + (1 - p_0) \int_{\Theta_1} p(x|\theta)\pi_1(\theta)d\theta} \tag{219}$$

Now suppose we use improper priors, $\pi_0(\theta) \propto c_0$ and $\pi_1(\theta) \propto c_1$. Then

$$p(H_0|x) \quad = \quad \frac{p_0 c_0 \int_{\Theta_0} p(x|\theta)d\theta}{p_0 c_0 \int_{\Theta_0} p(x|\theta)d\theta + (1 - p_0)c_1 \int_{\Theta_1} p(x|\theta)d\theta} \tag{220}$$

$$= \quad \frac{p_0 c_0 \ell_0}{p_0 c_0 \ell_0 + (1 - p_0)c_1 \ell_1} \tag{221}$$

where $\ell_i = \int_{\Theta_i} p(x|\theta) d\theta$ is the integrated likelihood. Now let $p_0 = p_1 = \frac{1}{2}$. Hence

$$\begin{aligned} p(H_0|x) &= \frac{c_0 \ell_0}{c_0 \ell_0 + c_1 \ell_1} && (222) \\ &= \frac{\ell_0}{\ell_0 + (c_1/c_0)\ell_1} && (223) \end{aligned}$$

Thus we can change the posterior arbitrarily by choosing c_1 and c_0. Note that using proper, but very vague, priors can cause the same problem. In particular, the Bayes factor will always favor the simpler model. This is called **Lindley's paradox**. Thus choosing the hyper-parameters of a prior is a way of controlling the complexity of the chosen model.

Note that, if H_0 and H_1 share the same prior over certain parameters, this part of the prior can be improper, since the normalization constant will cancel out.

6.4 Bayesian information criterion (BIC)

We can approximate the marginal likelihood in the large sample setting as follows. Let us apply the Laplace approximation (Section 5.1) to the posterior, so $f(\theta) = p(D|\theta)p(\theta)$, and $Z = p(D)$. From Equation 188, the Laplace approximation to the marginal likelihood is

$$p(D) \approx p(D|\theta_0)p(\theta_0)(2\pi)^{d/2}|C|^{\frac{1}{2}} \qquad (224)$$

where θ_0 is a posterior mode.

The **Bayesian information criterion** (BIC) is an approximation to the above approximation in which we assume $p(\theta) \propto 1$ and $|H| \approx n^d$, where n is the number of data points and d is the number of parameters (length of θ). Since $C = -H^{-1}$, we have

$$\log p(D) \approx \log p(D|\hat{\theta}_{MLE}) - \frac{1}{2} d \log n \qquad (225)$$

dropping additive constants.

The **Akaike Information Criterion** (AIC) is derived from a different framework, but the final answer is similar:

$$\log P(D) \approx \log P(D|\hat{\theta}_{MLE}) - d \qquad (226)$$

Note that determining the “effective number of parameters” d is a difficult problem in general, especially in latent variable models.

References

[Arc05] C. Archamebau. *Probabilistic models in noisy environments*. PhD thesis, U. Catholique de Louvain, Machine learning group, 2005.

[Ber85] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 1985.

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[DMP$^+$06] F. Demichelis, P. Magni, P. Piergiorgi, M. Rubin, and R. Bellazzi. A hierarchical Naive Bayes model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC Bioinformatics*, 7:514, 2006.

[GCSR04] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman and Hall, 2004. 2nd edition.

[Gel06] Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 2006.

[Jay03] E. T. Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

[Mac03] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Mur07] K. Murphy. Conjugate bayesian analysis of the gaussian distribution. Technical report, UBC, 2007.

[ZL04] M. Zhu and A. Lu. The counter-intuitive non-informative prior for the bernoulli family. *J. Statistics Education*, 2004.