5 *Bayesian statistics*

5.1 Introduction

We have now seen a variety of different probability models, and we have discussed how to fit them to data, i.e., we have discussed how to compute MAP parameter estimates $\hat{\boldsymbol{\theta}} = \operatorname{argmax} p(\boldsymbol{\theta}|\mathcal{D})$, using a variety of different priors. We have also discussed how to compute the full posterior $p(\boldsymbol{\theta}|\mathcal{D})$, as well as the posterior predictive density, $p(\mathbf{x}|\mathcal{D})$, for certain special cases (and in later chapters, we will discuss algorithms for the general case).

Using the posterior distribution to summarize everything we know about a set of unknown variables is at the core of **Bayesian statistics**. In this chapter, we discuss this approach to statistics in more detail. In Chapter 6, we discuss an alternative approach to statistics known as frequentist or classical statistics.

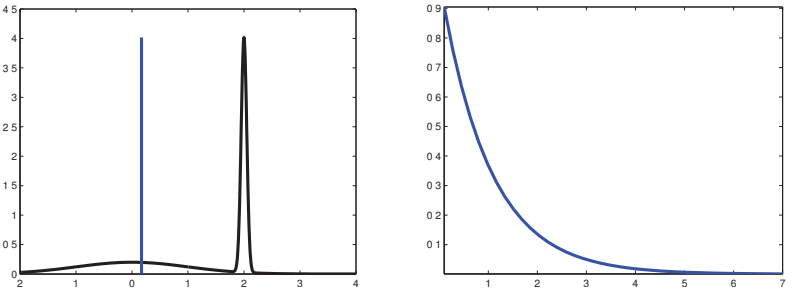
5.2 Summarizing posterior distributions

The posterior $p(\boldsymbol{\theta}|\mathcal{D})$ summarizes everything we know about the unknown quantities $\boldsymbol{\theta}$. In this section, we discuss some simple quantities that can be derived from a probability distribution, such as a posterior. These summary statistics are often easier to understand and visualize than the full joint.

5.2.1 MAP estimation

We can easily compute a **point estimate** of an unknown quantity by computing the posterior mean, median or mode. In Section 5.7, we discuss how to use decision theory to choose between these methods. Typically the posterior mean or median is the most appropriate choice for a real-valued quantity, and the vector of posterior marginals is the best choice for a discrete quantity. However, the posterior mode, aka the MAP estimate, is the most popular choice because it reduces to an optimization problem, for which efficient algorithms often exist. Futhermore, MAP estimation can be interpreted in non-Bayesian terms, by thinking of the log prior as a regularizer (see Section 6.5 for more details).

Although this approach is computationally appealing, it is important to point out that there are various drawbacks to MAP estimation, which we briefly discuss below. This will provide motivation for the more thoroughly Bayesian approach which we will study later in this chapter (and elsewhere in this book).



(a) (b)

Figure 5.1 (a) A bimodal distribution in which the mode is very untypical of the distribution. The thin blue vertical line is the mean, which is arguably a better summary of the distribution, since it is near the majority of the probability mass. Figure generated by `bimodalDemo`. (b) A skewed distribution in which the mode is quite different from the mean. Figure generated by `gammaPlotDemo`.

5.2.1.1 No measure of uncertainty

The most obvious drawback of MAP estimation, and indeed of any other **point estimate** such as the posterior mean or median, is that it does not provide any measure of uncertainty. In many applications, it is important to know how much one can trust a given estimate. We can derive such confidence measures from the posterior, as we discuss in Section 5.2.2.

5.2.1.2 Plugging in the MAP estimate can result in overfitting

In machine learning, we often care more about predictive accuracy than in interpreting the parameters of our models. However, if we don't model the uncertainty in our parameters, then our predictive distribution will be overconfident. We saw several examples of this in Chapter 3, and we will see more examples later. Overconfidence in predictions is particularly problematic in situations where we may be risk averse; see Section 5.7 for details.

5.2.1.3 The mode is an untypical point

Choosing the mode as a summary of a posterior distribution is often a very poor choice, since the mode is usually quite untypical of the distribution, unlike the mean or median. This is illustrated in Figure 5.1(a) for a 1d continuous space. The basic problem is that the mode is a point of measure zero, whereas the mean and median take the volume of the space into account. Another example is shown in Figure 5.1(b): here the mode is 0, but the mean is non-zero. Such skewed distributions often arise when inferring variance parameters, especially in hierarchical models. In such cases the MAP estimate (and hence the MLE) is obviously a very bad estimate.

How should we summarize a posterior if the mode is not a good choice? The answer is to use decision theory, which we discuss in Section 5.7. The basic idea is to specify a loss function, where $L(\theta, \hat{\theta})$ is the loss you incur if the truth is θ and your estimate is $\hat{\theta}$. If we use 0-1 loss, $L(\theta, \hat{\theta}) = \mathbb{I}(\theta \neq \hat{\theta})$, then the optimal estimate is the posterior mode. 0-1 loss means you only get "points" if you make no errors, otherwise you get nothing: there is no "partial credit" under

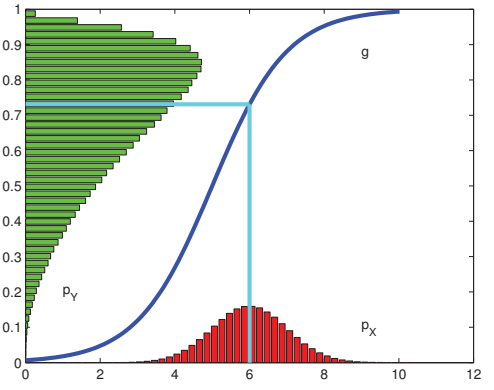


Figure 5.2 Example of the transformation of a density under a nonlinear transform. Note how the mode of the transformed distribution is not the transform of the original mode. Based on Exercise 1.4 of (Bishop 2006b). Figure generated by `bayesChangeOfVar`.

this loss function! For continuous-valued quantities, we often prefer to use squared error loss, $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$; the corresponding optimal estimator is then the posterior mean, as we show in Section 5.7. Or we can use a more robust loss function, $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, which gives rise to the posterior median.

5.2.1.4 MAP estimation is not invariant to reparameterization *

A more subtle problem with MAP estimation is that the result we get depends on how we parameterize the probability distribution. Changing from one representation to another equivalent representation changes the result, which is not very desirable, since the units of measurement are arbitrary (e.g., when measuring distance, we can use centimetres or inches).

To understand the problem, suppose we compute the posterior for x. If we define $y = f(x)$, the distribution for y is given by Equation 2.87, which we repeat here for convenience:

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| \tag{5.1}$$

The $|\frac{dx}{dy}|$ term is called the Jacobian, and it measures the change in size of a unit volume passed through f. Let $\hat{x} = \text{argmax}_x\, p_x(x)$ be the MAP estimate for x. In general it is not the case that $\hat{y} = \text{argmax}_y\, p_y(y)$ is given by $f(\hat{x})$. For example, let $x \sim \mathcal{N}(6, 1)$ and $y = f(x)$, where

$$f(x) = \frac{1}{1 + \exp(-x + 5)} \tag{5.2}$$

We can derive the distribution of y using Monte Carlo simulation (see Section 2.7.1). The result is shown in Figure 5.2. We see that the original Gaussian has become "squashed" by the sigmoid nonlinearity. In particular, we see that the mode of the transformed distribution is not equal to the transform of the original mode.

To see how this problem arises in the context of MAP estimation, consider the following example, due to Michael Jordan. The Bernoulli distribution is typically parameterized by its mean μ, so $p(y = 1|\mu) = \mu$, where $y \in \{0, 1\}$. Suppose we have a uniform prior on the unit interval: $p_\mu(\mu) = 1\ \mathbb{I}(0 \leq \mu \leq 1)$. If there is no data, the MAP estimate is just the mode of the prior, which can be anywhere between 0 and 1. We will now show that different parameterizations can pick different points in this interval arbitrarily.

First let $\theta = \sqrt{\mu}$ so $\mu = \theta^2$. The new prior is

$$p_\theta(\theta) = p_\mu(\mu)\left|\frac{d\mu}{d\theta}\right| = 2\theta \tag{5.3}$$

for $\theta \in [0, 1]$ so the new mode is

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in [0,1]} 2\theta = 1 \tag{5.4}$$

Now let $\phi = 1 - \sqrt{1 - \mu}$. The new prior is

$$p_\phi(\phi) = p_\mu(\mu)\left|\frac{d\mu}{d\phi}\right| = 2(1 - \phi) \tag{5.5}$$

for $\phi \in [0, 1]$, so the new mode is

$$\hat{\phi}_{MAP} = \arg\max_{\phi \in [0,1]} 2 - 2\phi = 0 \tag{5.6}$$

Thus the MAP estimate depends on the parameterization. The MLE does not suffer from this since the likelihood is a function, not a probability density. Bayesian inference does not suffer from this problem either, since the change of measure is taken into account when integrating over the parameter space.

One solution to the problem is to optimize the following objective function:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})|\mathbf{I}(\boldsymbol{\theta})|^{-\frac{1}{2}} \tag{5.7}$$

Here $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix associated with $p(\mathbf{x}|\boldsymbol{\theta})$ (see Section 6.2.2). This estimate is parameterization independent, for reasons explained in (Jermyn 2005; Druilhet and Marin 2007). Unfortunately, optimizing Equation 5.7 is often difficult, which minimizes the appeal of the whole approach.

5.2.2 Credible intervals

In addition to point estimates, we often want a measure of confidence. A standard measure of confidence in some (scalar) quantity θ is the "width" of its posterior distribution. This can be measured using a $100(1 - \alpha)$% **credible interval**, which is a (contiguous) region $C = (\ell, u)$ (standing for lower and upper) which contains $1 - \alpha$ of the posterior probability mass, i.e.,

$$C_\alpha(\mathcal{D}) = (\ell, u) : P(\ell \leq \theta \leq u|\mathcal{D}) = 1 - \alpha \tag{5.8}$$

There may be many such intervals, so we choose one such that there is $(1 - \alpha)/2$ mass in each tail; this is called a **central interval**.

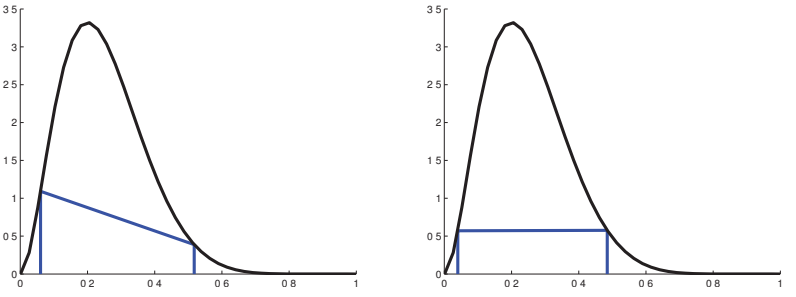


Figure 5.3 (a) Central interval and (b) HPD region for a Beta(3,9) posterior. The CI is (0.06, 0.52) and the HPD is (0.04, 0.48). Based on Figure 3.6 of (Hoff 2009). Figure generated by `betaHPD`.

If the posterior has a known functional form, we can compute the posterior central interval using $\ell = F^{-1}(\alpha/2)$ and $u = F^{-1}(1-\alpha/2)$, where F is the cdf of the posterior. For example, if the posterior is Gaussian, $p(\theta|\mathcal{D}) = \mathcal{N}(0, 1)$, and $\alpha = 0.05$, then we have $\ell = \Phi(\alpha/2) = -1.96$, and $u = \Phi(1 - \alpha/2) = 1.96$, where Φ denotes the cdf of the Gaussian. This is illustrated in Figure 2.3(c). This justifies the common practice of quoting a credible interval in the form of $\mu \pm 2\sigma$, where μ represents the posterior mean, σ represents the posterior standard deviation, and 2 is a good approximation to 1.96.

Of course, the posterior is not always Gaussian. For example, in our coin example, if we use a uniform prior and we observe $N_1 = 47$ heads out of $N = 100$ trials, then the posterior is a beta distribution, $p(\theta|\mathcal{D}) = \text{Beta}(48, 54)$. We find the 95% posterior credible interval is $(0.3749, 0.5673)$ (see `betaCredibleInt` for the one line of Matlab code we used to compute this).

If we don't know the functional form, but we can draw samples from the posterior, then we can use a Monte Carlo approximation to the posterior quantiles: we simply sort the S samples, and find the one that occurs at location α/S along the sorted list. As $S \to \infty$, this converges to the true quantile. See `mcQuantileDemo` for a demo.

People often confuse Bayesian credible intervals with frequentist confidence intervals. However, they are not the same thing, as we discuss in Section 6.6.1. In general, credible intervals are usually what people want to compute, but confidence intervals are usually what they actually compute, because most people are taught frequentist statistics but not Bayesian statistics. Fortunately, the mechanics of computing a credible interval is just as easy as computing a confidence interval (see e.g., `betaCredibleInt` for how to do it in Matlab).

5.2.2.1 Highest posterior density regions *

A problem with central intervals is that there might be points outside the CI which have higher probability density. This is illustrated in Figure 5.3(a), where we see that points outside the left-most CI boundary have higher density than those just inside the right-most CI boundary.

This motivates an alternative quantity known as the **highest posterior density** or **HPD** region. This is defined as the (set of) most probable points that in total constitute $100(1 - \alpha)\%$ of the

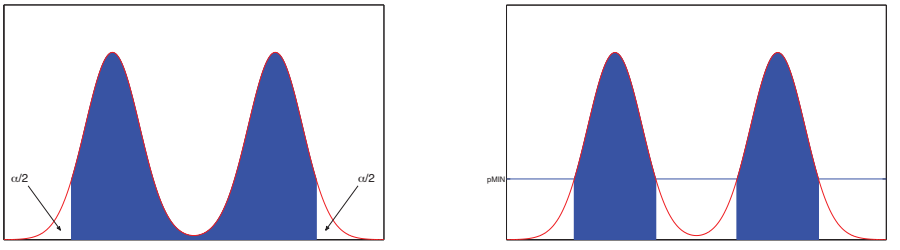


Figure 5.4 (a) Central interval and (b) HPD region for a hypothetical multimodal posterior. Based on Figure 2.2 of (Gelman et al. 2004). Figure generated by `postDensityIntervals`.

probability mass. More formally, we find the threshold p^* on the pdf such that

$$1 - \alpha = \int_{\theta: p(\theta|\mathcal{D}) > p^*} p(\theta|\mathcal{D}) d\theta \tag{5.9}$$

and then define the HPD as

$$C_\alpha(\mathcal{D}) = \{\theta : p(\theta|\mathcal{D}) \geq p^*\} \tag{5.10}$$

In 1d, the HPD region is sometimes called a **highest density interval** or **HDI**. For example, Figure 5.3(b) shows the 95% HDI of a $\text{Beta}(3, 9)$ distribution, which is $(0.04, 0.48)$. We see that this is narrower than the CI, even though it still contains 95% of the mass; furthermore, every point inside of it has higher density than every point outside of it.

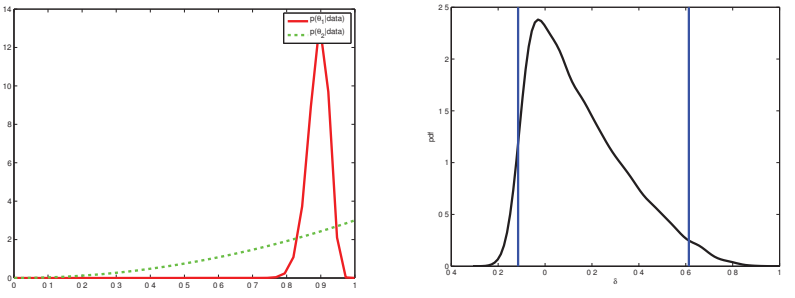
For a unimodal distribution, the HDI will be the narrowest interval around the mode containing 95% of the mass. To see this, imagine "water filling" in reverse, where we lower the level until 95% of the mass is revealed, and only 5% is submerged. This gives a simple algorithm for computing HDIs in the 1d case: simply search over points such that the interval contains 95% of the mass and has minimal width. This can be done by 1d numerical optimization if we know the inverse CDF of the distribution, or by search over the sorted data points if we have a bag of samples (see `betaHPD` for a demo).

If the posterior is multimodal, the HDI may not even be a connected region: see Figure 5.4(b) for an example. However, summarizing multimodal posteriors is always difficult.

5.2.3 Inference for a difference in proportions

Sometimes we have multiple parameters, and we are interested in computing the posterior distribution of some function of these parameters. For example, suppose you are about to buy something from Amazon.com, and there are two sellers offering it for the same price. Seller 1 has 90 positive reviews and 10 negative reviews. Seller 2 has 2 positive reviews and 0 negative reviews. Who should you buy from?[1]

1. This example is from `www.johndcook.com/blog/2011/09/27/bayesian-amazon`. See also `lingpipe-blog.com/2009/10/13/bayesian-counterpart-to-fisher-exact-test-on-contingency-tables`.



(a) (b)

Figure 5.5 (a) Exact posteriors $p(\theta_i|\mathcal{D}_i)$. (b) Monte Carlo approximation to $p(\delta|\mathcal{D})$. We use kernel density estimation to get a smooth plot. The vertical lines enclose the 95% central interval. Figure generated by `amazonSellerDemo`,

On the face of it, you should pick seller 2, but we cannot be very confident that seller 2 is better since it has had so few reviews. In this section, we sketch a Bayesian analysis of this problem. Similar methodology can be used to compare rates or proportions across groups for a variety of other settings.

Let θ_1 and θ_2 be the unknown reliabilities of the two sellers. Since we don't know much about them, we'll endow them both with uniform priors, $\theta_i \sim \text{Beta}(1,1)$. The posteriors are $p(\theta_1|\mathcal{D}_1) = \text{Beta}(91,11)$ and $p(\theta_2|\mathcal{D}_2) = \text{Beta}(3,1)$.

We want to compute $p(\theta_1 > \theta_2|\mathcal{D})$. For convenience, let us define $\delta = \theta_1 - \theta_2$ as the difference in the rates. (Alternatively we might want to work in terms of the log-odds ratio.) We can compute the desired quantity using numerical integration:

$$p(\delta > 0|\mathcal{D}) = \int_0^1 \int_0^1 \mathbb{I}(\theta_1 > \theta_2)\text{Beta}(\theta_1|y_1+1, N_1-y_1+1) \text{Beta}(\theta_2|y_2+1, N_2-y_2+1)d\theta_1 d\theta_2 \quad (5.11)$$

We find $p(\delta > 0|\mathcal{D}) = 0.710$, which means you are better off buying from seller 1! See `amazonSellerDemo` for the code. (It is also possible to solve the integral analytically (Cook 2005).)

A simpler way to solve the problem is to approximate the posterior $p(\delta|\mathcal{D})$ by Monte Carlo sampling. This is easy, since θ_1 and θ_2 are independent in the posterior, and both have beta distributions, which can be sampled from using standard methods. The distributions $p(\theta_i|\mathcal{D}_i)$ are shown in Figure 5.5(a), and a MC approximation to $p(\delta|\mathcal{D})$, together with a 95% HPD, is shown Figure 5.5(b). An MC approximation to $p(\delta > 0|\mathcal{D})$ is obtained by counting the fraction of samples where $\theta_1 > \theta_2$; this turns out to be 0.718, which is very close to the exact value. (See `amazonSellerDemo` for the code.)

5.3 Bayesian model selection

In Figure 1.18, we saw that using too high a degree polynomial results in overfitting, and using too low a degree results in underfitting. Similarly, in Figure 7.8(a), we saw that using too small

a regularization parameter results in overfitting, and too large a value results in underfitting. In general, when faced with a set of models (i.e., families of parametric distributions) of different complexity, how should we choose the best one? This is called the **model selection** problem.

One approach is to use cross-validation to estimate the generalization error of all the candiate models, and then to pick the model that seems the best. However, this requires fitting each model K times, where K is the number of CV folds. A more efficient approach is to compute the posterior over models,

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(m, \mathcal{D})} \tag{5.12}$$

From this, we can easily compute the MAP model, $\hat{m} = \operatorname{argmax} p(m|\mathcal{D})$. This is called **Bayesian model selection**.

If we use a uniform prior over models, $p(m) \propto 1$, this amounts to picking the model which maximizes

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta} \tag{5.13}$$

This quantity is called the **marginal likelihood**, the **integrated likelihood**, or the **evidence** for model m. The details on how to perform this integral will be discussed in Section 5.3.2. But first we give an intuitive interpretation of what this quantity means.

5.3.1 Bayesian Occam's razor

One might think that using $p(\mathcal{D}|m)$ to select models would always favor the model with the most parameters. This is true if we use $p(\mathcal{D}|\hat{\boldsymbol{\theta}}_m)$ to select models, where $\hat{\boldsymbol{\theta}}_m$ is the MLE or MAP estimate of the parameters for model m, because models with more parameters will fit the data better, and hence achieve higher likelihood. However, if we integrate out the parameters, rather than maximizing them, we are automatically protected from overfitting: models with more parameters do not necessarily have higher *marginal* likelihood. This is called the **Bayesian Occam's razor** effect (MacKay 1995b; Murray and Ghahramani 2005), named after the principle known as **Occam's razor**, which says one should pick the simplest model that adequately explains the data.

One way to understand the Bayesian Occam's razor is to notice that the marginal likelihood can be rewritten as follows, based on the chain rule of probability (Equation 2.5):

$$p(\mathcal{D}) = p(y_1)p(y_2|y_1)p(y_3|y_{1:2}) \ldots p(y_N|y_{1:N-1}) \tag{5.14}$$

where we have dropped the conditioning on $\mathbf{x}$ for brevity. This is similar to a leave-one-out cross-validation estimate (Section 1.4.8) of the likelihood, since we predict each future point given all the previous ones. (Of course, the order of the data does not matter in the above expression.) If a model is too complex, it will overfit the "early" examples and will then predict the remaining ones poorly.

Another way to understand the Bayesian Occam's razor effect is to note that probabilities must sum to one. Hence $\sum_{\mathcal{D}'} p(\mathcal{D}'|m) = 1$, where the sum is over all possible data sets. Complex models, which can predict many things, must spread their probability mass thinly, and hence will not obtain as large a probability for any given data set as simpler models. This is sometimes

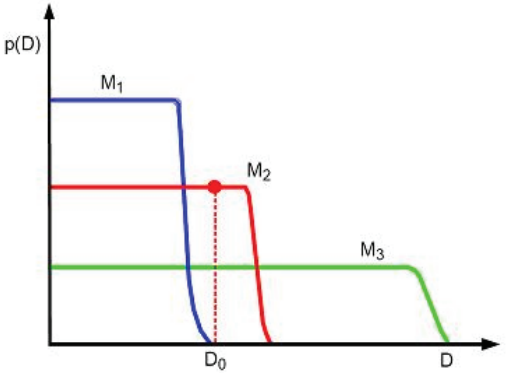


Figure 5.6 A schematic illustration of the Bayesian Occam's razor. The broad (green) curve corresponds to a complex model, the narrow (blue) curve to a simple model, and the middle (red) curve is just right. Based on Figure 3.13 of (Bishop 2006a). See also (Murray and Ghahramani 2005, Figure 2) for a similar plot produced on real data.

called the **conservation of probability mass** principle, and is illustrated in Figure 5.6. On the horizontal axis we plot all possible data sets in order of increasing complexity (measured in some abstract sense). On the vertical axis we plot the predictions of 3 possible models: a simple one, M_1; a medium one, M_2; and a complex one, M_3. We also indicate the actually observed data $\mathcal{D}_0$ by a vertical line. Model 1 is too simple and assigns low probability to $\mathcal{D}_0$. Model 3 also assigns $\mathcal{D}_0$ relatively low probability, because it can predict many data sets, and hence it spreads its probability quite widely and thinly. Model 2 is "just right": it predicts the observed data with a reasonable degree of confidence, but does not predict too many other things. Hence model 2 is the most probable model.

As a concrete example of the Bayesian Occam's razor, consider the data in Figure 5.7. We plot polynomials of degrees 1, 2 and 3 fit to $N = 5$ data points. It also shows the posterior over models, where we use a Gaussian prior (see Section 7.6 for details). There is not enough data to justify a complex model, so the MAP model is $d = 1$. Figure 5.8 shows what happens when $N = 30$. Now it is clear that $d = 2$ is the right model (the data was in fact generated from a quadratic).

As another example, Figure 7.8(c) plots $\log p(\mathcal{D}|\lambda)$ vs $\log(\lambda)$, for the polynomial ridge regression model, where λ ranges over the same set of values used in the CV experiment. We see that the maximum evidence occurs at roughly the same point as the minimum of the test MSE, which also corresponds to the point chosen by CV.

When using the Bayesian approach, we are not restricted to evaluating the evidence at a finite grid of values. Instead, we can use numerical optimization to find $\lambda^* = \mathrm{argmax}_\lambda \, p(\mathcal{D}|\lambda)$. This technique is called **empirical Bayes** or **type II maximum likelihood** (see Section 5.6 for details). An example is shown in Figure 7.8(b): we see that the curve has a similar shape to the CV estimate, but it can be computed more efficiently.

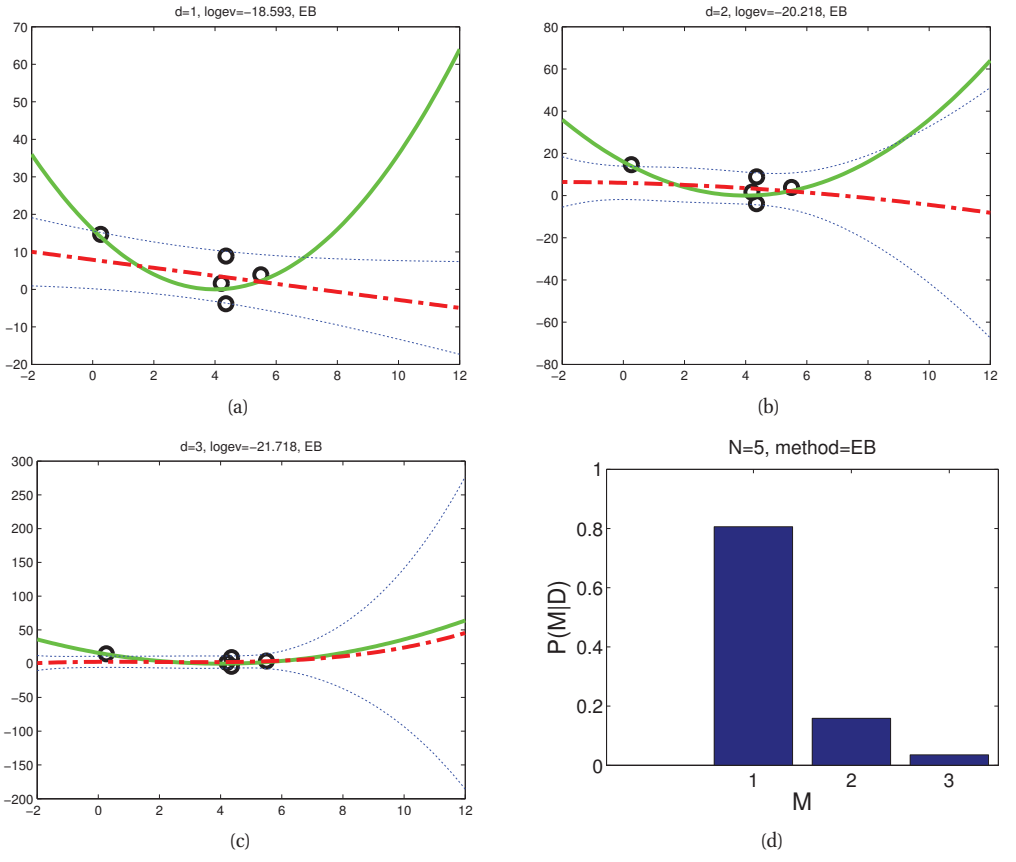


Figure 5.7 (a-c) We plot polynomials of degrees 1, 2 and 3 fit to $N = 5$ data points using empirical Bayes. The solid green curve is the true function, the dashed red curve is the prediction (dotted blue lines represent $\pm\sigma$ around the mean). (d) We plot the posterior over models, $p(d|\mathcal{D})$, assuming a uniform prior $p(d) \propto 1$. Based on a figure by Zoubin Ghahramani. Figure generated by `linregEbModelSelVsN`.

5.3.2 Computing the marginal likelihood (evidence)

When discussing parameter inference for a fixed model, we often wrote

$$p(\boldsymbol{\theta}|\mathcal{D}, m) \propto p(\boldsymbol{\theta}|m)p(\mathcal{D}|\boldsymbol{\theta}, m) \tag{5.15}$$

thus ignoring the normalization constant $p(\mathcal{D}|m)$. This is valid since $p(\mathcal{D}|m)$ is constant wrt $\boldsymbol{\theta}$. However, when comparing models, we need to know how to compute the marginal likelihood, $p(\mathcal{D}|m)$. In general, this can be quite hard, since we have to integrate over all possible parameter values, but when we have a conjugate prior, it is easy to compute, as we now show.

Let $p(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/Z_0$ be our prior, where $q(\boldsymbol{\theta})$ is an unnormalized distribution, and Z_0 is the normalization constant of the prior. Let $p(\mathcal{D}|\boldsymbol{\theta}) = q(\mathcal{D}|\boldsymbol{\theta})/Z_\ell$ be the likelihood, where Z_ℓ contains any constant factors in the likelihood. Finally let $p(\boldsymbol{\theta}|\mathcal{D}) = q(\boldsymbol{\theta}|\mathcal{D})/Z_N$ be our poste-

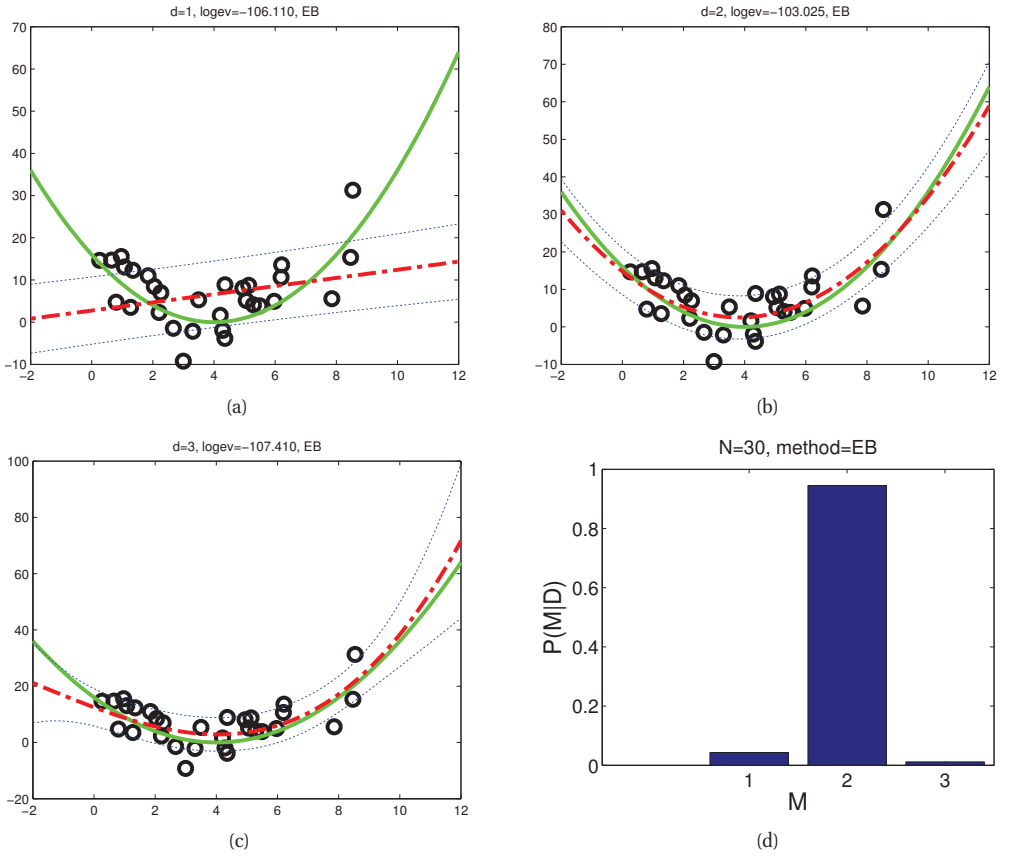


Figure 5.8 Same as Figure 5.7 except now $N = 30$. Figure generated by `linregEbModelSelVsN`.

rior, where $q(\boldsymbol{\theta}|\mathcal{D}) = q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})$ is the unnormalized posterior, and Z_N is the normalization constant of the posterior. We have

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \tag{5.16}$$

$$\frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N} = \frac{q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{Z_\ell Z_0 p(\mathcal{D})} \tag{5.17}$$

$$p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_\ell} \tag{5.18}$$

So assuming the relevant normalization constants are tractable, we have an easy way to compute the marginal likelihood. We give some examples below.

5.3.2.1 Beta-binomial model

Let us apply the above result to the Beta-binomial model. Since we know $p(\theta|\mathcal{D}) = \text{Beta}(\theta|a', b')$, where $a' = a + N_1$ and $b' = b + N_0$, we know the normalization constant of the posterior is $B(a', b')$. Hence

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{5.19}$$

$$= \frac{1}{p(\mathcal{D})}\left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right]\left[\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}\right] \tag{5.20}$$

$$= \binom{N}{N_1}\frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)}\left[\theta^{a+N_1-1}(1-\theta)^{b+N_0-1}\right] \tag{5.21}$$

So

$$\frac{1}{B(a+N_1, b+N_0)} = \binom{N}{N_1}\frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)} \tag{5.22}$$

$$p(\mathcal{D}) = \binom{N}{N_1}\frac{B(a+N_1, b+N_0)}{B(a,b)} \tag{5.23}$$

The marginal likelihood for the Beta-Bernoulli model is the same as above, except it is missing the $\binom{N}{N_1}$ term.

5.3.2.2 Dirichlet-multinoulli model

By the same reasoning as the Beta-Bernoulli case, one can show that the marginal likelihood for the Dirichlet-multinoulli model is given by

$$p(\mathcal{D}) = \frac{B(\mathbf{N}+\boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \tag{5.24}$$

where

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} \tag{5.25}$$

Hence we can rewrite the above result in the following form, which is what is usually presented in the literature:

$$p(\mathcal{D}) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)}\prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{5.26}$$

We will see many applications of this equation later.

5.3.2.3 Gaussian-Gaussian-Wishart model

Consider the case of an MVN with a conjugate NIW prior. Let Z_0 be the normalizer for the prior, Z_N be normalizer for the posterior, and let $Z_l = (2\pi)^{ND/2}$ be the normalizer for the

likelihood. Then it is easy to see that

$$p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_l} \tag{5.27}$$

$$= \frac{1}{\pi^{ND/2}} \frac{1}{2^{ND/2}} \frac{\left(\frac{2\pi}{\kappa_N}\right)^{D/2} |\mathbf{S}_N|^{-\nu_N/2} 2^{(\nu_0+N)D/2} \Gamma_D(\nu_N/2)}{\left(\frac{2\pi}{\kappa_0}\right)^{D/2} |\mathbf{S}_0|^{-\nu_0/2} 2^{\nu_0 D/2} \Gamma_D(\nu_0/2)} \tag{5.28}$$

$$= \frac{1}{\pi^{ND/2}} \left(\frac{\kappa_0}{\kappa_N}\right)^{D/2} \frac{|\mathbf{S}_0|^{\nu_0/2}}{|\mathbf{S}_N|^{\nu_N/2}} \frac{\Gamma_D(\nu_N/2)}{\Gamma_D(\nu_0/2)} \tag{5.29}$$

This equation will prove useful later.

5.3.2.4 BIC approximation to log marginal likelihood

In general, computing the integral in Equation 5.13 can be quite difficult. One simple but popular approximation is known as the **Bayesian information criterion** or **BIC**, which has the following form (Schwarz 1978):

$$\text{BIC} \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{\text{dof}(\hat{\boldsymbol{\theta}})}{2} \log N \approx \log p(\mathcal{D}) \tag{5.30}$$

where $\text{dof}(\hat{\boldsymbol{\theta}})$ is the number of **degrees of freedom** in the model, and $\hat{\boldsymbol{\theta}}$ is the MLE for the model.[2] We see that this has the form of a **penalized log likelihood**, where the penalty term depends on the model's complexity. See Section 8.4.2 for the derivation of the BIC score.

As an example, consider linear regression. As we show in Section 7.3, the MLE is given by $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\hat{\sigma}^2 = \text{RSS}/N$, where $\text{RSS} = \sum_{i=1}^N (y_i - \hat{\mathbf{w}}_{mle}^T \mathbf{x}_i)^2$. The corresponding log likelihood is given by

$$\log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) = -\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{N}{2} \tag{5.31}$$

Hence the BIC score is as follows (dropping constant terms)

$$\text{BIC} = -\frac{N}{2} \log(\hat{\sigma}^2) - \frac{D}{2} \log(N) \tag{5.32}$$

where D is the number of variables in the model. In the statistics literature, it is common to use an alternative definition of BIC, which we call the BIC *cost* (since we want to minimize it):

$$\text{BIC-cost} \triangleq -2 \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) + \text{dof}(\hat{\boldsymbol{\theta}}) \log N \approx -2 \log p(\mathcal{D}) \tag{5.33}$$

In the context of linear regression, this becomes

$$\text{BIC-cost} = N \log(\hat{\sigma}^2) + D \log(N) \tag{5.34}$$

2. Traditionally the BIC score is defined using the ML estimate $\hat{\boldsymbol{\theta}}$, so it is independent of the prior. However, for models such as mixtures of Gaussians, the ML estimate can be poorly behaved, so it is better to evaluate the BIC score using the MAP estimate, as in (Fraley and Raftery 2007).

The BIC method is very closely related to the **minimum description length** or **MDL** principle, which characterizes the score for a model in terms of how well it fits the data, minus how complex the model is to define. See (Hansen and Yu 2001) for details.

There is a very similar expression to BIC/ MDL called the **Akaike information criterion** or **AIC**, defined as

$$\text{AIC}(m, \mathcal{D}) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}) - \text{dof}(m) \tag{5.35}$$

This is derived from a frequentist framework, and cannot be interpreted as an approximation to the marginal likelihood. Nevertheless, the form of this expression is very similar to BIC. We see that the penalty for AIC is less than for BIC. This causes AIC to pick more complex models. However, this can result in better predictive accuracy. See e.g., (Clarke et al. 2009, sec 10.2) for further discussion on such information criteria.

5.3.2.5 Effect of the prior

Sometimes it is not clear how to set the prior. When we are performing posterior inference, the details of the prior may not matter too much, since the likelihood often overwhelms the prior anyway. But when computing the marginal likelihood, the prior plays a much more important role, since we are averaging the likelihood over all possible parameter settings, as weighted by the prior.

In Figures 5.7 and 5.8, where we demonstrated model selection for linear regression, we used a prior of the form $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$. Here α is a tuning parameter that controls how strong the prior is. This parameter can have a large effect, as we discuss in Section 7.5. Intuitively, if α is large, the weights are "forced" to be small, so we need to use a complex model with many small parameters (e.g., a high degree polynomial) to fit the data. Conversely, if α is small, we will favor simpler models, since each parameter is "allowed" to vary in magnitude by a lot.

If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior. That is, we should put a prior on the hyper-parameter α as well as the parametrs $\mathbf{w}$. To compute the marginal likelihood, we should integrate out all unknowns, i.e., we should compute

$$p(\mathcal{D}|m) = \int \int p(\mathcal{D}|\mathbf{w}) p(\mathbf{w}|\alpha, m) p(\alpha|m) d\mathbf{w} d\alpha \tag{5.36}$$

Of course, this requires specifying the hyper-prior. Fortunately, the higher up we go in the Bayesian hierarchy, the less sensitive are the results to the prior settings. So we can usually make the hyper-prior uninformative.

A computational shortcut is to optimize α rather than integrating it out. That is, we use

$$p(\mathcal{D}|m) \approx \int p(\mathcal{D}|\mathbf{w}) p(\mathbf{w}|\hat{\alpha}, m) d\mathbf{w} \tag{5.37}$$

where

$$\hat{\alpha} = \operatorname*{argmax}_{\alpha} p(\mathcal{D}|\alpha, m) = \operatorname*{argmax}_{\alpha} \int p(\mathcal{D}|\mathbf{w}) p(\mathbf{w}|\alpha, m) d\mathbf{w} \tag{5.38}$$

This approach is called empirical Bayes (EB), and is discussed in more detail in Section 5.6. This is the method used in Figures 5.7 and 5.8.

Bayes factor $BF(1,0)$	Interpretation
$BF < \frac{1}{100}$	Decisive evidence for M_0
$BF < \frac{1}{10}$	Strong evidence for M_0
$\frac{1}{10} < BF < \frac{1}{3}$	Moderate evidence for M_0
$\frac{1}{3} < BF < 1$	Weak evidence for M_0
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Moderate evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

Table 5.1 Jeffreys' scale of evidence for interpreting Bayes factors.

5.3.3 Bayes factors

Suppose our prior on models is uniform, $p(m) \propto 1$. Then model selection is equivalent to picking the model with the highest marginal likelihood. Now suppose we just have two models we are considering, call them the **null hypothesis**, M_0, and the **alternative hypothesis**, M_1. Define the **Bayes factor** as the ratio of marginal likelihoods:

$$BF_{1,0} \triangleq \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})} \Big/ \frac{p(M_1)}{p(M_0)} \tag{5.39}$$

(This is like a **likelihood ratio**, except we integrate out the parameters, which allows us to compare models of different complexity.) If $BF_{1,0} > 1$ then we prefer model 1, otherwise we prefer model 0.

Of course, it might be that $BF_{1,0}$ is only slightly greater than 1. In that case, we are not very confident that model 1 is better. Jeffreys (1961) proposed a scale of evidence for interpreting the magnitude of a Bayes factor, which is shown in Table 5.1. This is a Bayesian alternative to the frequentist concept of a p-value.[3] Alternatively, we can just convert the Bayes factor to a posterior over models. If $p(M_1) = p(M_0) = 0.5$, we have

$$p(M_0|\mathcal{D}) = \frac{BF_{0,1}}{1 + BF_{0,1}} = \frac{1}{BF_{1,0} + 1} \tag{5.40}$$

5.3.3.1 Example: Testing if a coin is fair

Suppose we observe some coin tosses, and want to decide if the data was generated by a fair coin, $\theta = 0.5$, or a potentially biased coin, where θ could be any value in $[0, 1]$. Let us denote the first model by M_0 and the second model by M_1. The marginal likelihood under M_0 is simply

$$p(\mathcal{D}|M_0) = \left(\frac{1}{2}\right)^N \tag{5.41}$$

3. A **p-value**, is defined as the probability (under the null hypothesis) of observing some **test statistic** $f(\mathcal{D})$ (such as the **chi-squared statistic**) that is as large *or larger* than that actually observed, i.e., $\text{pvalue}(\mathcal{D}) \triangleq P(f(\tilde{\mathcal{D}}) \geq f(\mathcal{D})|\tilde{\mathcal{D}} \sim H_0)$. Note that has almost nothing to do with what we really want to know, which is $p(H_0|\mathcal{D})$.

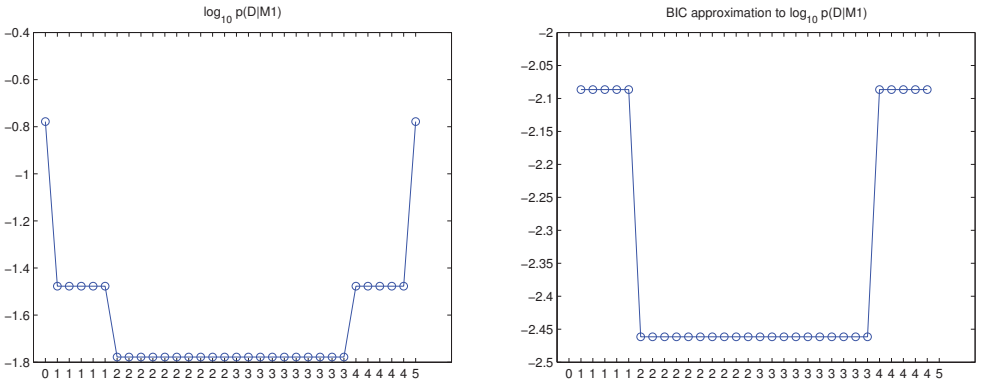


Figure 5.9 (a) Log marginal likelihood for the coins example. (b) BIC approximation. Figure generated by `coinsModelSelDemo`.

where N is the number of coin tosses. The marginal likelihood under M_1, using a Beta prior, is

$$p(\mathcal{D}|M_1) = \int p(\mathcal{D}|\theta)p(\theta)d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)} \tag{5.42}$$

We plot $\log p(\mathcal{D}|M_1)$ vs the number of heads N_1 in Figure 5.9(a), assuming $N = 5$ and $\alpha_1 = \alpha_0 = 1$. (The shape of the curve is not very sensitive to α_1 and α_0, as long as $\alpha_0 = \alpha_1$.) If we observe 2 or 3 heads, the unbiased coin hypothesis M_0 is more likely than M_1, since M_0 is a simpler model (it has no free parameters) — it would be a suspicious coincidence if the coin were biased but happened to produce almost exactly 50/50 heads/tails. However, as the counts become more extreme, we favor the biased coin hypothesis. Note that, if we plot the log Bayes factor, $\log BF_{1,0}$, it will have exactly the same shape, since $\log p(\mathcal{D}|M_0)$ is a constant. See also Exercise 3.18.

In Figure 5.9(b) shows the BIC approximation to $\log p(\mathcal{D}|M_1)$ for our biased coin example from Section 5.3.3.1. We see that the curve has approximately the same shape as the exact log marginal likelihood, which is all that matters for model selection purposes, since the absolute scale is irrelevant. In particular, it favors the simpler model unless the data is overwhelmingly in support of the more complex model.

5.3.4 Jeffreys-Lindley paradox *

Problems can arise when we use improper priors (i.e., priors that do not integrate to 1) for model selection/ hypothesis testing, even though such priors may be acceptable for other purposes. For example, consider testing the hypotheses $M_0 : \theta \in \Theta_0$ vs $M_1 : \theta \in \Theta_1$. To define the marginal density on θ, we use the following mixture model

$$p(\theta) = p(\theta|M_0)p(M_0) + p(\theta|M_1)p(M_1) \tag{5.43}$$

This is only meaningful if $p(\theta|M_0)$ and $p(\theta|M_1)$ are proper (normalized) density functions. In this case, the posterior is given by

$$p(M_0|\mathcal{D}) = \frac{p(M_0)p(\mathcal{D}|M_0)}{p(M_0)p(\mathcal{D}|M_0) + p(M_1)p(\mathcal{D}|M_1)} \tag{5.44}$$

$$= \frac{p(M_0)\int_{\Theta_0} p(\mathcal{D}|\theta)p(\theta|M_0)d\theta}{p(M_0)\int_{\Theta_0} p(\mathcal{D}|\theta)p(\theta|M_0)d\theta + p(M_1)\int_{\Theta_1} p(\mathcal{D}|\theta)p(\theta|M_1)d\theta} \tag{5.45}$$

Now suppose we use improper priors, $p(\theta|M_0) \propto c_0$ and $p(\theta|M_1) \propto c_1$. Then

$$p(M_0|\mathcal{D}) = \frac{p(M_0)c_0\int_{\Theta_0} p(\mathcal{D}|\theta)d\theta}{p(M_0)c_0\int_{\Theta_0} p(\mathcal{D}|\theta)d\theta + p(M_1)c_1\int_{\Theta_1} p(\mathcal{D}|\theta)d\theta} \tag{5.46}$$

$$= \frac{p(M_0)c_0\ell_0}{p(M_0)c_0\ell_0 + p(M_1)c_1\ell_1} \tag{5.47}$$

where $\ell_i = \int_{\Theta_i} p(\mathcal{D}|\theta)d\theta$ is the integrated or marginal likelihood for model i. Now let $p(M_0) = p(M_1) = \frac{1}{2}$. Hence

$$p(M_0|\mathcal{D}) = \frac{c_0\ell_0}{c_0\ell_0 + c_1\ell_1} = \frac{\ell_0}{\ell_0 + (c_1/c_0)\ell_1} \tag{5.48}$$

Thus we can change the posterior arbitrarily by choosing c_1 and c_0 as we please. Note that using proper, but very vague, priors can cause similar problems. In particular, the Bayes factor will always favor the simpler model, since the probability of the observed data under a complex model with a very diffuse prior will be very small. This is called the **Jeffreys-Lindley paradox**.

Thus it is important to use proper priors when performing model selection. Note, however, that, if M_0 and M_1 share the same prior over a subset of the parameters, this part of the prior can be improper, since the corresponding normalization constant will cancel out.

5.4 Priors

The most controversial aspect of Bayesian statistics is its reliance on priors. Bayesians argue this is unavoidable, since nobody is a **tabula rasa** or **blank slate**: all inference must be done conditional on certain assumptions about the world. Nevertheless, one might be interested in minimizing the impact of one's prior assumptions. We briefly discuss some ways to do this below.

5.4.1 Uninformative priors

If we don't have strong beliefs about what θ should be, it is common to use an **uninformative** or **non-informative** prior, and to "let the data speak for itself".

The issue of designing uninformative priors is actually somewhat tricky. As an example of the difficulty, consider a Bernoulli parameter, $\theta \in [0, 1]$. One might think that the most uninformative prior would be the uniform distribution, $\mathrm{Beta}(1, 1)$. But the posterior mean in this case is $\mathbb{E}[\theta|\mathcal{D}] = \frac{N_1+1}{N_1+N_0+2}$, whereas the MLE is $\frac{N_1}{N_1+N_0}$. Hence one could argue that the prior wasn't completely uninformative after all.

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

$$\lim_{c\to 0} \mathrm{Beta}(c,c) = \mathrm{Beta}(0,0) \tag{5.49}$$

which is a mixture of two equal point masses at 0 and 1 (see (Zhu and Lu 2004)). This is also called the **Haldane prior**. Note that the Haldane prior is an improper prior, meaning it does not integrate to 1. However, as long as we see at least one head and at least one tail, the posterior will be proper.

In Section 5.4.2.1 we will argue that the "right" uninformative prior is in fact $\mathrm{Beta}(\frac{1}{2},\frac{1}{2})$. Clearly the difference in practice between these three priors is very likely negligible. In general, it is advisable to perform some kind of **sensitivity analysis**, in which one checks how much one's conclusions or predictions change in response to change in the modeling assumptions, which includes the choice of prior, but also the choice of likelihood and any kind of data pre-processing. If the conclusions are relatively insensitive to the modeling assumptions, one can have more confidence in the results.

5.4.2 Jeffreys priors *

Harold Jeffreys[4] designed a general purpose technique for creating non-informative priors. The result is known as the **Jeffreys prior**. The key observation is that if $p(\phi)$ is non-informative, then any re-parameterization of the prior, such as $\theta = h(\phi)$ for some function h, should also be non-informative. Now, by the change of variables formula,

$$p_\theta(\theta) = p_\phi(\phi)\left|\frac{d\phi}{d\theta}\right| \tag{5.50}$$

so the prior will in general change. However, let us pick

$$p_\phi(\phi) \propto (I(\phi))^{\frac{1}{2}} \tag{5.51}$$

where $I(\phi)$ is the **Fisher information**:

$$I(\phi) \triangleq -\mathbb{E}\left[\left(\frac{d\log p(X|\phi)}{d\phi}\right)2\right] \tag{5.52}$$

This is a measure of curvature of the expected negative log likelihood and hence a measure of stability of the MLE (see Section 6.2.2). Now

$$\frac{d\log p(x|\theta)}{d\theta} = \frac{d\log p(x|\phi)}{d\phi}\frac{d\phi}{d\theta} \tag{5.53}$$

Squaring and taking expectations over x, we have

$$I(\theta) = -\mathbb{E}\left[\left(\frac{d\log p(X|\theta)}{d\theta}\right)^2\right] = I(\phi)\left(\frac{d\phi}{d\theta}\right)^2 \tag{5.54}$$

$$I(\theta)^{\frac{1}{2}} = I(\phi)^{\frac{1}{2}}\left|\frac{d\phi}{d\theta}\right| \tag{5.55}$$

4. Harold Jeffreys, 1891 – 1989, was an English mathematician, statistician, geophysicist, and astronomer.

so we find the transformed prior is

$$p_\theta(\theta) \;=\; p_\phi(\phi)\left|\frac{d\phi}{d\theta}\right| \propto (I(\phi))^{\frac{1}{2}}\left|\frac{d\phi}{d\theta}\right| = I(\theta)^{\frac{1}{2}} \tag{5.56}$$

So $p_\theta(\theta)$ and $p_\phi(\phi)$ are the same.

Some examples will make this clearer.

5.4.2.1 Example: Jeffreys prior for the Bernoulli and multinoulli

Suppose $X \sim \mathrm{Ber}(\theta)$. The log likelihood for a single sample is

$$\log p(X|\theta) = X \log \theta + (1 - X) \log(1 - \theta) \tag{5.57}$$

The **score function** is just the gradient of the log-likelihood:

$$s(\theta) \;\triangleq\; \frac{d}{d\theta} \log p(X|\theta) = \frac{X}{\theta} - \frac{1 - X}{1 - \theta} \tag{5.58}$$

The **observed information** is the second derivative of the log-likelihood:

$$J(\theta) \;=\; -\frac{d^2}{d\theta^2} \log p(X|\theta) = -s'(\theta|X) = \frac{X}{\theta^2} + \frac{1 - X}{(1 - \theta)^2} \tag{5.59}$$

The Fisher information is the expected information:

$$I(\theta) \;=\; E[J(\theta|X)|X \sim \theta] = \frac{\theta}{\theta^2} + \frac{1 - \theta}{(1 - \theta)^2} = \frac{1}{\theta(1 - \theta)} \tag{5.60}$$

Hence Jeffreys' prior is

$$p(\theta) \propto \theta^{-\frac{1}{2}}(1 - \theta)^{-\frac{1}{2}} = \frac{1}{\sqrt{\theta(1 - \theta)}} \propto \mathrm{Beta}(\frac{1}{2}, \frac{1}{2}) \tag{5.61}$$

Now consider a multinoulli random variable with K states. One can show that the Jeffreys' prior is given by

$$p(\boldsymbol{\theta}) \propto \mathrm{Dir}(\frac{1}{2}, \ldots, \frac{1}{2}) \tag{5.62}$$

Note that this is different from the more obvious choices of $\mathrm{Dir}(\frac{1}{K}, \ldots, \frac{1}{K})$ or $\mathrm{Dir}(1, \ldots, 1)$.

5.4.2.2 Example: Jeffreys prior for location and scale parameters

One can show that the Jeffreys prior for a location parameter, such as the Gaussian mean, is $p(\mu) \propto 1$. Thus is an example of a **translation invariant prior**, which satisfies the property that the probability mass assigned to any interval, $[A, B]$ is the same as that assigned to any other shifted interval of the same width, such as $[A - c, B - c]$. That is,

$$\int_{A-c}^{B-c} p(\mu)d\mu = (A - c) - (B - c) = (A - B) = \int_A^B p(\mu)d\mu \tag{5.63}$$

This can be achieved using $p(\mu) \propto 1$, which we can approximate by using a Gaussian with infinite variance, $p(\mu) = \mathcal{N}(\mu|0, \infty)$. Note that this is an **improper prior**, since it does not integrate to 1. Using improper priors is fine as long as the posterior is proper, which will be the case provided we have seen $N \geq 1$ data points, since we can "nail down" the location as soon as we have seen a single data point.

Similarly, one can show that the Jeffreys prior for a scale parameter, such as the Gaussian variance, is $p(\sigma^2) \propto 1/\sigma^2$. This is an example of a **scale invariant prior**, which satisfies the property that the probability mass assigned to any interval $[A, B]$ is the same as that assigned to any other interval $[A/c, B/c]$ which is scaled in size by some constant factor $c > 0$. (For example, if we change units from meters to feet we do not want that to affect our inferences.) This can be achieved by using

$$p(s) \propto 1/s \tag{5.64}$$

To see this, note that

$$\int_{A/c}^{B/c} p(s)ds = [\log s]_{A/c}^{B/c} = \log(B/c) - \log(A/c) \tag{5.65}$$

$$= \log(B) - \log(A) = \int_A^B p(s)ds \tag{5.66}$$

We can approximate this using a degenerate Gamma distribution (Section 2.4.4), $p(s) = \mathrm{Ga}(s|0, 0)$. The prior $p(s) \propto 1/s$ is also improper, but the posterior is proper as soon as we have seen $N \geq 2$ data points (since we need at least two data points to estimate a variance).

5.4.3 Robust priors

In many cases, we are not very confident in our prior, so we want to make sure it does not have an undue influence on the result. This can be done by using **robust priors** (Insua and Ruggeri 2000), which typically have heavy tails, which avoids forcing things to be too close to the prior mean.

Let us consider an example from (Berger 1985, p7). Suppose $x \sim \mathcal{N}(\theta, 1)$. We observe that $x = 5$ and we want to estimate θ. The MLE is of course $\hat{\theta} = 5$, which seems reasonable. The posterior mean under a uniform prior is also $\overline{\theta} = 5$. But now suppose we know that the prior median is 0, and the prior quantiles are at -1 and 1, so $p(\theta \leq -1) = p(-1 < \theta \leq 0) = p(0 < \theta \leq 1) = p(1 < \theta) = 0.25$. Let us also assume the prior is smooth and unimodal.

It is easy to show that a Gaussian prior of the form $\mathcal{N}(\theta|0, 2.19^2)$ satisfies these prior constraints. But in this case the posterior mean is given by 3.43, which doesn't seem very satisfactory.

Now suppose we use as a Cauchy prior $\mathcal{T}(\theta|0, 1, 1)$. This also satisfies the prior constraints of our example. But this time we find (using numerical method integration: see `robustPriorDemo` for the code) that the posterior mean is about 4.6, which seems much more reasonable.

5.4.4 Mixtures of conjugate priors

Robust priors are useful, but can be computationally expensive to use. Conjugate priors simplify the computation, but are often not robust, and not flexible enough to encode our prior knowl-

edge. However, it turns out that a **mixture of conjugate priors** is also conjugate (Exercise 5.1), and can approximate any kind of prior (Dallal and Hall 1983; Diaconis and Ylvisaker 1985). Thus such priors provide a good compromise between computational convenience and flexibility.

For example, suppose we are modeling coin tosses, and we think the coin is either fair, or is biased towards heads. This cannot be represented by a beta distribution. However, we can model it using a mixture of two beta distributions. For example, we might use

$$p(\theta) = 0.5\ \mathrm{Beta}(\theta|20, 20) + 0.5\ \mathrm{Beta}(\theta|30, 10) \tag{5.67}$$

If θ comes from the first distribution, the coin is fair, but if it comes from the second, it is biased towards heads.

We can represent a mixture by introducing a latent indicator variable z, where $z = k$ means that θ comes from mixture component k. The prior has the form

$$p(\theta) = \sum_k p(z = k)p(\theta|z = k) \tag{5.68}$$

where each $p(\theta|z = k)$ is conjugate, and $p(z = k)$ are called the (prior) mixing weights. One can show (Exercise 5.1) that the posterior can also be written as a mixture of conjugate distributions as follows:

$$p(\theta|\mathcal{D}) \quad = \quad \sum_k p(z = k|\mathcal{D})p(\theta|\mathcal{D}, z = k) \tag{5.69}$$

where $p(Z = k|\mathcal{D})$ are the posterior mixing weights given by

$$p(Z = k|\mathcal{D}) \quad = \quad \frac{p(Z = k)p(\mathcal{D}|Z = k)}{\sum_{k'} p(Z = k')p(\mathcal{D}|Z = k')} \tag{5.70}$$

Here the quantity $p(\mathcal{D}|Z = k)$ is the marginal likelihood for mixture component k (see Section 5.3.2.1).

5.4.4.1 Example

Suppose we use the mixture prior

$$p(\theta) = 0.5\mathrm{Beta}(\theta|a_1, b_1) + 0.5\mathrm{Beta}(\theta|a_2, b_2) \tag{5.71}$$

where $a_1 = b_1 = 20$ and $a_2 = b_2 = 10$. and we observe N_1 heads and N_0 tails. The posterior becomes

$$p(\theta|\mathcal{D}) = p(Z = 1|\mathcal{D})\mathrm{Beta}(\theta|a_1 + N_1, b_1 + N_0) + p(Z = 2|\mathcal{D})\mathrm{Beta}(\theta|a_2 + N_1, b_2 + N_0) \tag{5.72}$$

If $N_1 = 20$ heads and $N_0 = 10$ tails, then, using Equation 5.23, the posterior becomes

$$p(\theta|\mathcal{D}) = 0.346\ \mathrm{Beta}(\theta|40, 30) + 0.654\ \mathrm{Beta}(\theta|50, 20) \tag{5.73}$$

See Figure 5.10 for an illustration.

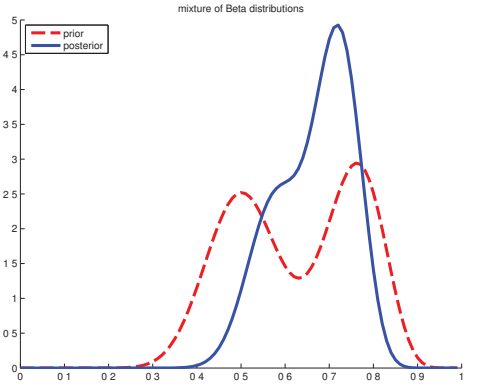


Figure 5.10 A mixture of two Beta distributions. Figure generated by `mixBetaDemo`.

5.4.4.2 Application: Finding conserved regions in DNA and protein sequences

We mentioned that Dirichlet-multinomial models are widely used in biosequence analysis. Let us give a simple example to illustrate some of the machinery that has developed. Specifically, consider the sequence logo discussed in Section 2.3.2.1. Now suppose we want to find locations which represent coding regions of the genome. Such locations often have the same letter across all sequences, because of evolutionary pressure. So we need to find columns which are "pure", or nearly so, in the sense that they are mostly all As, mostly all Ts, mostly all Cs, or mostly all Gs. One approach is to look for low-entropy columns; these will be ones whose distribution is nearly deterministic (pure).

But suppose we want to associate a confidence measure with our estimates of purity. This can be useful if we believe adjacent locations are conserved together. In this case, we can let $Z_1 = 1$ if location t is conserved, and let $Z_t = 0$ otherwise. We can then add a dependence between adjacent Z_t variables using a Markov chain; see Chapter 17 for details.

In any case, we need to define a likelihood model, $p(\mathbf{N}_t|Z_t)$, where $\mathbf{N}_t$ is the vector of (A,C,G,T) counts for column t. It is natural to make this be a multinomial distribution with parameter $\boldsymbol{\theta}_t$. Since each column has a different distribution, we will want to integrate out $\boldsymbol{\theta}_t$ and thus compute the marginal likelihood

$$p(\mathbf{N}_t|Z_t) = \int p(\mathbf{N}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|Z_t)d\boldsymbol{\theta}_t \tag{5.74}$$

But what prior should we use for $\boldsymbol{\theta}_t$? When $Z_t = 0$ we can use a uniform prior, $p(\boldsymbol{\theta}|Z_t = 0) = \mathrm{Dir}(1, 1, 1, 1)$, but what should we use if $Z_t = 1$? After all, if the column is conserved, it could be a (nearly) pure column of As, Cs, Gs, or Ts. A natural approach is to use a mixture of Dirichlet priors, each one of which is "tilted" towards the appropriate corner of the 4-dimensional simplex, e.g.,

$$p(\boldsymbol{\theta}|Z_t = 1) = \frac{1}{4}\mathrm{Dir}(\boldsymbol{\theta}|(10, 1, 1, 1)) + \cdots + \frac{1}{4}\mathrm{Dir}(\boldsymbol{\theta}|(1, 1, 1, 10)) \tag{5.75}$$

Since this is conjugate, we can easily compute $p(\mathbf{N}_t|Z_t)$. See (Brown et al. 1993) for an

application of these ideas to a real bio-sequence problem.

5.5 Hierarchical Bayes

A key requirement for computing the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is the specification of a prior $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ are the hyper-parameters. What if we don't know how to set $\boldsymbol{\eta}$? In some cases, we can use uninformative priors, we we discussed above. A more Bayesian approach is to put a prior on our priors! In terms of graphical models (Chapter 10), we can represent the situation as follows:

$$\boldsymbol{\eta} \rightarrow \boldsymbol{\theta} \rightarrow \mathcal{D} \tag{5.76}$$

This is an example of a **hierarchical Bayesian model**, also called a **multi-level model**, since there are multiple levels of unknown quantities. We give a simple example below, and we will see many others later in the book.

5.5.1 Example: modeling related cancer rates

Consider the problem of predicting cancer rates in various cities (this example is from (Johnson and Albert 1999, p24)). In particular, suppose we measure the number of people in various cities, N_i, and the number of people who died of cancer in these cities, x_i. We assume $x_i \sim \text{Bin}(N_i, \theta_i)$, and we want to estimate the cancer rates θ_i. One approach is to estimate them all separately, but this will suffer from the sparse data problem (underestimation of the rate of cancer due to small N_i). Another approach is to assume all the θ_i are the same; this is called **parameter tying**. The resulting pooled MLE is just $\hat{\theta} = \frac{\sum_i x_i}{\sum_i N_i}$. But the assumption that all the cities have the same rate is a rather strong one. A compromise approach is to assume that the θ_i are similar, but that there may be city-specific variations. This can be modeled by assuming the θ_i are drawn from some common distribution, say $\theta_i \sim \text{Beta}(a, b)$. The full joint distribution can be written as

$$p(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{N}) \;=\; p(\boldsymbol{\eta}) \prod_{i=1}^{N} \text{Bin}(x_i|N_i, \theta_i)\text{Beta}(\theta_i|\boldsymbol{\eta}) \tag{5.77}$$

where $\boldsymbol{\eta} = (a, b)$.

Note that it is crucial that we infer $\boldsymbol{\eta} = (a, b)$ from the data; if we just clamp it to a constant, the θ_i will be conditionally independent, and there will be no information flow between them. By contrast, by treating $\boldsymbol{\eta}$ as an unknown (hidden variable), we allow the data-poor cities to **borrow statistical strength** from data-rich ones.

Suppose we compute the joint posterior $p(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathcal{D})$. From this we can get the posterior marginals $p(\theta_i|\mathcal{D})$. In Figure 5.11(a), we plot the posterior means, $\mathbb{E}[\theta_i|\mathcal{D}]$, as blue bars, as well as the population level mean, $\mathbb{E}[a/(a+b)|\mathcal{D}]$, shown as a red line (this represents the average of the θ_i's). We see that the posterior mean is shrunk towards the pooled estimate more strongly for cities with small sample sizes N_i. For example, city 1 and city 20 both have a 0 observed cancer incidence rate, but city 20 has a smaller population, so its rate is shrunk more towards the population-level estimate (i.e., it is closer to the horizontal red line) than city 1.

Figure 5.11(b) shows the 95% posterior credible intervals for θ_i. We see that city 15, which has a very large population (53,637 people), has small posterior uncertainty. Consequently this city

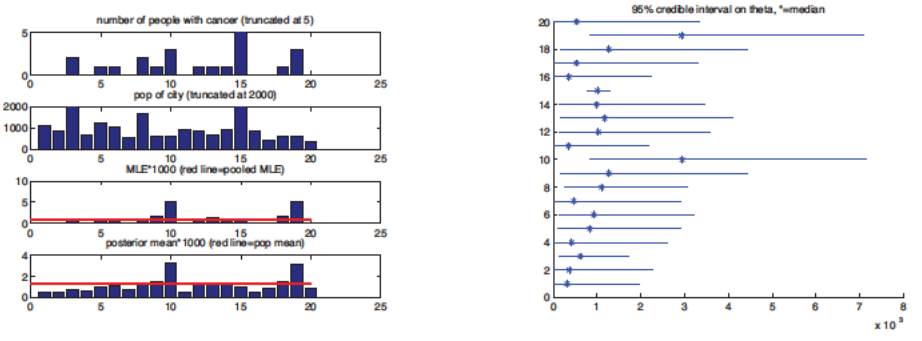


Figure 5.11 (a) Results of fitting the model using the data from (Johnson and Albert 1999, p24). First row: Number of cancer incidents x_i in 20 cities in Missouri. Second row: population size N_i. The largest city (number 15) has a population of $N_{15} = 53637$ and $x_{15} = 54$ incidents, but we truncate the vertical axes of the first two rows so that the differences between the other cities are visible. Third row: MLE $\hat{\theta}_i$. The red line is the pooled MLE. Fourth row: posterior mean $\mathbb{E}[\theta_i|\mathcal{D}]$. The red line is $\mathbb{E}[a/(a+b)|\mathcal{D}]$, the population-level mean. (b) Posterior 95% credible intervals on the cancer rates. Figure generated by `cancerRatesEb`

has the largest impact on the posterior estimate of η, which in turn will impact the estimate of the cancer rates for other cities. Cities 10 and 19, which have the highest MLE, also have the highest posterior uncertainty, reflecting the fact that such a high estimate is in conflict with the prior (which is estimated from all the other cities).

In the above example, we have one parameter per city, modeling the probability the response is on. By making the Bernoulli rate parameter be a function of covariates, $\theta_i = \text{sigm}(\mathbf{w}_i^T \mathbf{x})$, we can model multiple correlated logistic regression tasks. This is called **multi-task learning**, and will be discussed in more detail in Section 9.5.

5.6 Empirical Bayes

In hierarchical Bayesian models, we need to compute the posterior on multiple levels of latent variables. For example, in a two-level model, we need to compute

$$p(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta}) \tag{5.78}$$

In some cases, we can analytically marginalize out $\boldsymbol{\theta}$; this leaves is with the simpler problem of just computing $p(\boldsymbol{\eta}|\mathcal{D})$.

As a computational shortcut, we can approximate the posterior on the hyper-parameters with a point-estimate, $p(\boldsymbol{\eta}|\mathcal{D}) \approx \delta_{\hat{\boldsymbol{\eta}}}(\boldsymbol{\eta})$, where $\hat{\boldsymbol{\eta}} = \text{argmax}\, p(\boldsymbol{\eta}|\mathcal{D})$. Since $\boldsymbol{\eta}$ is typically much smaller than $\boldsymbol{\theta}$ in dimensionality, it is less prone to overfitting, so we can safely use a uniform prior on $\boldsymbol{\eta}$. Then the estimate becomes

$$\hat{\boldsymbol{\eta}} = \text{argmax}\, p(\mathcal{D}|\boldsymbol{\eta}) = \text{argmax} \left[\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta} \right] \tag{5.79}$$

where the quantity inside the brackets is the marginal or integrated likelihood, sometimes called the evidence. This overall approach is called **empirical Bayes** (**EB**) or **type-II maximum likelihood**. In machine learning, it is sometimes called the **evidence procedure**.

Empirical Bayes violates the principle that the prior should be chosen independently of the data. However, we can just view it as a computationally cheap approximation to inference in a hierarchical Bayesian model, just as we viewed MAP estimation as an approximation to inference in the one level model $\boldsymbol{\theta} \rightarrow \mathcal{D}$. In fact, we can construct a hierarchy in which the more integrals one performs, the "more Bayesian" one becomes:

Method	Definition
Maximum likelihood	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}\|\boldsymbol{\theta})$
MAP estimation	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})$
ML-II (Empirical Bayes)	$\hat{\boldsymbol{\eta}} = \text{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})d\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}\|\boldsymbol{\eta})$
MAP-II	$\hat{\boldsymbol{\eta}} = \text{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})p(\boldsymbol{\eta})d\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}\|\boldsymbol{\eta})p(\boldsymbol{\eta})$
Full Bayes	$p(\boldsymbol{\theta}, \boldsymbol{\eta}\|\mathcal{D}) \propto p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})p(\boldsymbol{\eta})$

Note that EB can be shown to have good frequentist properties (see e.g., (Carlin and Louis 1996; Efron 2010)), so it is widely used by non-Bayesians. For example, the popular James-Stein estimator, discussed in Section 6.3.3.2, can be derived using EB.

5.6.1 Example: beta-binomial model

Let us return to the cancer rates model. We can analytically integrate out the θ_i's, and write down the marginal likelihood directly, as follows:

$$p(\mathcal{D}|a,b) = \prod_i \int \text{Bin}(x_i|N_i,\theta_i)\text{Beta}(\theta_i|a,b)d\theta_i \tag{5.80}$$

$$= \prod_i \frac{B(a+x_i, b+N_i-x_i)}{B(a,b)} \tag{5.81}$$

Various ways of maximizing this wrt a and b are discussed in (Minka 2000e).

Having estimated a and b, we can plug in the hyper-parameters to compute the posterior $p(\theta_i|\hat{a},\hat{b},\mathcal{D})$ in the usual way, using conjugate analysis. The net result is that the posterior mean of each θ_i is a weighted average of its local MLE and the prior means, which depends on $\boldsymbol{\eta} = (a,b)$; but since $\boldsymbol{\eta}$ is estimated based on all the data, each θ_i is influenced by all the data.

5.6.2 Example: Gaussian-Gaussian model

We now study another example that is analogous to the cancer rates example, except the data is real-valued. We will use a Gaussian likelihood and a Gaussian prior. This will allow us to write down the solution analytically.

In particular, suppose we have data from multiple related groups. For example, x_{ij} could be the test score for student i in school j, for $j = 1:D$ and $i = 1:N_j$. We want to estimate the mean score for each school, θ_j. However, since the sample size, N_j, may be small for

some schools, we can regularize the problem by using a hierarchical Bayesian model, where we assume θ_j come from a common prior, $\mathcal{N}(\mu, \tau^2)$.

The joint distribution has the following form:

$$p(\boldsymbol{\theta}, \mathcal{D}|\boldsymbol{\eta}, \sigma^2) = \prod_{j=1}^{D} \mathcal{N}(\theta_j|\mu, \tau^2) \prod_{i=1}^{N_j} \mathcal{N}(x_{ij}|\theta_j, \sigma^2) \tag{5.82}$$

where we assume σ^2 is known for simplicity. (We relax this assumption in Exercise 24.4.) We explain how to estimate $\boldsymbol{\eta}$ below. Once we have estimated $\boldsymbol{\eta} = (\mu, \tau)$, we can compute the posteriors over the θ_j's. To do that, it simplifies matters to rewrite the joint distribution in the following form, exploiting the fact that N_j Gaussian measurements with values x_{ij} and variance σ^2 are equivalent to one measurement of value $\overline{x}_j \triangleq \frac{1}{N_j} \sum_{i=1}^{N_j} x_{ij}$ with variance $\sigma_j^2 \triangleq \sigma^2/N_j$. This yields

$$p(\boldsymbol{\theta}, \mathcal{D}|\hat{\boldsymbol{\eta}}, \sigma^2) = \prod_{j=1}^{D} \mathcal{N}(\theta_j|\hat{\mu}, \hat{\tau}^2) \mathcal{N}(\overline{x}_j|\theta_j, \sigma_j^2) \tag{5.83}$$

From this, it follows from the results of Section 4.4.1 that the posteriors are given by

$$p(\theta_j|\mathcal{D}, \hat{\mu}, \hat{\tau}^2) = \mathcal{N}(\theta_j|\hat{B}_j\hat{\mu} + (1 - \hat{B}_j)\overline{x}_j, (1 - \hat{B}_j)\sigma_j^2) \tag{5.84}$$

$$\hat{B}_j \triangleq \frac{\sigma_j^2}{\sigma_j^2 + \hat{\tau}^2} \tag{5.85}$$

where $\hat{\mu} = \overline{x}$ and $\hat{\tau}^2$ will be defined below.

The quantity $0 \leq \hat{B}_j \leq 1$ controls the degree of **shrinkage** towards the overall mean, μ. If the data is reliable for group j (e.g., because the sample size N_j is large), then σ_j^2 will be small relative to τ^2; hence $\hat{B}_j$ will be small, and we will put more weight on $\overline{x}_j$ when we estimate θ_j. However, groups with small sample sizes will get regularized (shrunk towards the overall mean μ) more heavily. We will see an example of this below.

If $\sigma_j = \sigma$ for all groups j, the posterior mean becomes

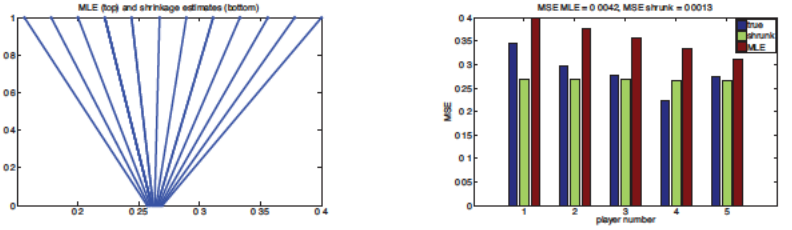
$$\hat{\theta}_j = \hat{B}\overline{x} + (1 - \hat{B})\overline{x}_j = \overline{x} + (1 - \hat{B})(\overline{x}_j - \overline{x}) \tag{5.86}$$

This has exactly the same form as the James Stein estimator discussed in Section 6.3.3.2.

5.6.2.1 Example: predicting baseball scores

We now give an example of shrinkage applied to baseball batting averages, from (Efron and Morris 1975). We observe the number of hits for $D = 18$ players during the first $T = 45$ games. Call the number of hits b_i. We assume $b_j \sim \text{Bin}(T, \theta_j)$, where θ_j is the "true" batting average for player j. The goal is to estimate the θ_j. The MLE is of course $\hat{\theta}_j = x_j$, where $x_j = b_j/T$ is the empirical batting average. However, we can use an EB approach to do better.

To apply the Gaussian shrinkage approach described above, we require that the likelihood be Gaussian, $x_j \sim \mathcal{N}(\theta_j, \sigma^2)$ for known σ^2. (We drop the i subscript since we assume $N_j = 1$,



(a) (b)

Figure 5.12 (a) MLE parameters (top) and corresponding shrunken estimates (bottom). (b) We plot the true parameters (blue), the posterior mean estimate (green), and the MLEs (red) for 5 of the players. Figure generated by `shrinkageDemoBaseball`.

since x_j already represents the average for player j.) However, in this example we have a binomial likelihood. While this has the right mean, $\mathbb{E}[x_j] = \theta_j$, the variance is not constant:

$$\text{var}[x_j] = \frac{1}{T^2}\,\text{var}[b_j] = \frac{T\theta_j(1-\theta_j)}{T^2} \tag{5.87}$$

So let us apply a **variance stabilizing transform**[5] to x_j to better match the Gaussian assumption:

$$y_j = f(y_j) = \sqrt{T}\arcsin(2y_j - 1) \tag{5.88}$$

Now we have approximately $y_j \sim \mathcal{N}(f(\theta_j), 1) = \mathcal{N}(\mu_j, 1)$. We use Gaussian shrinkage to estimate the μ_j using Equation 5.86 with $\sigma^2 = 1$, and we then transform back to get

$$\hat{\theta}_j = 0.5(\sin(\hat{\mu}_j/\sqrt{T}) + 1) \tag{5.89}$$

The results are shown in Figure 5.12(a-b). In (a), we plot the MLE $\hat{\theta}_j$ and the posterior mean $\overline{\theta}_j$. We see that all the estimates have shrunk towards the global mean, 0.265. In (b), we plot the true value θ_j, the MLE $\hat{\theta}_j$ and the posterior mean $\overline{\theta}_j$. (The "true" values of θ_j are estimated from a large number of independent games.) We see that, on average, the shrunken estimate is much closer to the true parameters than the MLE is. Specifically, the mean squared error, defined by $\text{MSE} = \frac{1}{N}\sum_{j=1}^{D}(\theta_j - \overline{\theta}_j)^2$, is over three times smaller using the shrinkage estimates $\overline{\theta}_j$ than using the MLEs $\hat{\theta}_j$.

5.6.2.2 Estimating the hyper-parameters

In this section, we give an algorithm for estimating $\boldsymbol{\eta}$. Suppose initially that $\sigma_j^2 = \sigma^2$ is the same for all groups. In this case, we can derive the EB estimate in closed form, as we now show. From Equation 4.126, we have

$$p(\overline{x}_j|\mu,\tau^2,\sigma^2) = \int \mathcal{N}(\overline{x}_j|\theta_j,\sigma^2)\mathcal{N}(\theta_j|\mu,\tau^2)d\theta_j = \mathcal{N}(\overline{x}_j|\mu,\tau^2+\sigma^2) \tag{5.90}$$

5. Suppose $\mathbb{E}[X] = \mu$ and $\text{var}[X] = \sigma^2(\mu)$. Let $Y = f(X)$. Then a Taylor series expansions gives $Y \approx f(\mu) + (X-\mu)f'(\mu)$. Hence $\text{var}[Y] \approx f'(\mu)^2\text{var}[X-\mu] = f'(\mu)^2\sigma^2(\mu)$. A variance stabilizing transformation is a function f such that $f'(\mu)^2\sigma^2(\mu)$ is independent of μ.

Hence the marginal likelihood is

$$p(\mathcal{D}|\mu, \tau^2, \sigma^2) = \prod_{j=1}^{D} \mathcal{N}(\overline{x}_j|\mu, \tau^2 + \sigma^2) \tag{5.91}$$

Thus we can estimate the hyper-parameters using the usual MLEs for a Gaussian. For μ, we have

$$\hat{\mu} = \frac{1}{D}\sum_{j=1}^{D} \overline{x}_j = \overline{x} \tag{5.92}$$

which is the overall mean.

For the variance, we can use moment matching (which is equivalent to the MLE for a Gaussian): we simply equate the model variance to the empirical variance:

$$\hat{\tau}^2 + \sigma^2 = \frac{1}{D}\sum_{j=1}^{D}(\overline{x}_j - \overline{x})^2 \triangleq s^2 \tag{5.93}$$

so $\hat{\tau}^2 = s^2 - \sigma^2$. Since we know τ^2 must be positive, it is common to use the following revised estimate:

$$\hat{\tau}^2 = \max\{0, s^2 - \sigma^2\} = (s^2 - \sigma^2)_+ \tag{5.94}$$

Hence the shrinkage factor is

$$\hat{B} = \frac{\sigma^2}{\sigma^2 + \hat{\tau}^2} = \frac{\sigma^2}{\sigma^2 + (s^2 - \sigma^2)_+} \tag{5.95}$$

In the case where the σ_j^2's are different, we can no longer derive a solution in closed form. Exercise 11.13 discusses how to use the EM algorithm to derive an EB estimate, and Exercise 24.4 discusses how to perform full Bayesian inference in this hierarchical model.

5.7 Bayesian decision theory

We have seen how probability theory can be used to represent and updates our beliefs about the state of the world. However, ultimately our goal is to convert our beliefs into actions. In this section, we discuss the optimal way to do this.

We can formalize any given statistical decision problem as a game against nature (as opposed to a game against other strategic players, which is the topic of game theory, see e.g., (Shoham and Leyton-Brown 2009) for details). In this game, nature picks a state or parameter or label, $y \in \mathcal{Y}$, unknown to us, and then generates an observation, $\mathbf{x} \in \mathcal{X}$, which we get to see. We then have to make a decision, that is, we have to choose an action a from some **action space** $\mathcal{A}$. Finally we incur some **loss**, $L(y, a)$, which measures how compatible our action a is with nature's hidden state y. For example, we might use misclassification loss, $L(y, a) = \mathbb{I}(y \neq a)$, or squared loss, $L(y, a) = (y - a)^2$. We will see some other examples below.

Our goal is to devise a **decision procedure** or **policy**, $\delta : \mathcal{X} \to \mathcal{A}$, which specifies the optimal action for each possible input. By optimal, we mean the action that minimizes the expected loss:

$$\delta(\mathbf{x}) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \, \mathbb{E}\left[L(y, a)\right] \tag{5.96}$$

In economics, it is more common to talk of a **utility function**; this is just negative loss, $U(y, a) = -L(y, a)$. Thus the above rule becomes

$$\delta(\mathbf{x}) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathbb{E}\left[U(y, a)\right] \tag{5.97}$$

This is called the **maximum expected utility principle**, and is the essence of what we mean by **rational behavior**.

Note that there are two different interpretations of what we mean by "expected". In the Bayesian version, which we discuss below, we mean the expected value of y given the data we have seen so far. In the frequentist version, which we discuss in Section 6.3, we mean the expected value of y and $\mathbf{x}$ that we expect to see in the future.

In the Bayesian approach to decision theory, the optimal action, having observed $\mathbf{x}$, is defined as the action a that minimizes the **posterior expected loss**:

$$\rho(a|\mathbf{x}) \triangleq \mathbb{E}_{p(y|\mathbf{x})}\left[L(y, a)\right] = \sum_y L(y, a) p(y|\mathbf{x}) \tag{5.98}$$

(If y is continuous (e.g., when we want to estimate a parameter vector), we should replace the sum with an integral.) Hence the **Bayes estimator**, also called the **Bayes decision rule**, is given by

$$\delta(\mathbf{x}) = \arg\min_{a \in \mathcal{A}} \rho(\mathbf{a}|\mathbf{x}) \tag{5.99}$$

5.7.1 Bayes estimators for common loss functions

In this section we show how to construct Bayes estimators for the loss functions most commonly arising in machine learning.

5.7.1.1 MAP estimate minimizes 0-1 loss

The **0-1 loss** is defined by

$$L(y, a) = \mathbb{I}(y \neq a) = \begin{cases} 0 & \text{if } a = y \\ 1 & \text{if } a \neq y \end{cases} \tag{5.100}$$

This is commonly used in classification problems where y is the true class label and $a = \hat{y}$ is the estimate.

For example, in the two class case, we can write the loss matrix as follows:

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	1
$y = 0$	1	0

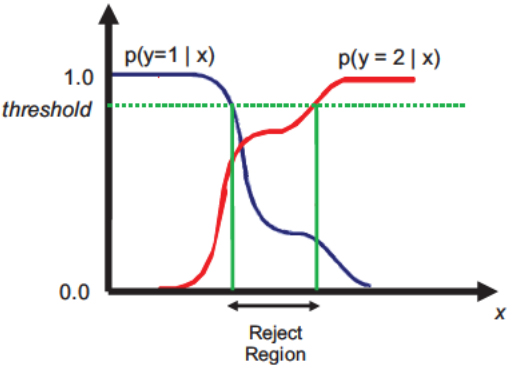


Figure 5.13 For some regions of input space, where the class posteriors are uncertain, we may prefer not to choose class 1 or 2; instead we may prefer the reject option. Based on Figure 1.26 of (Bishop 2006a).

(In Section 5.7.2, we generalize this loss function so it penalizes the two kinds of errors on the off-diagonal differently.)

The posterior expected loss is

$$\rho(a|\mathbf{x}) = p(a \neq y|\mathbf{x}) = 1 - p(y|\mathbf{x}) \tag{5.101}$$

Hence the action that minimizes the expected loss is the posterior mode or MAP estimate

$$y^*(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}) \tag{5.102}$$

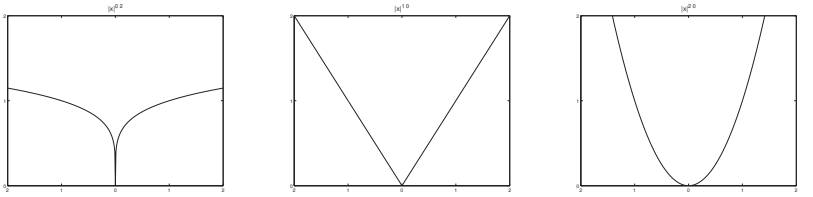
5.7.1.2 Reject option

In classification problems where $p(y|\mathbf{x})$ is very uncertain, we may prefer to choose a **reject action**, in which we refuse to classify the example as any of the specified classes, and instead say "don't know". Such ambiguous cases can be handled by e.g., a human expert. See Figure 5.13 for an illustration. This is useful in **risk averse** domains such as medicine and finance.

We can formalize the reject option as follows. Let choosing $a = C + 1$ correspond to picking the reject action, and choosing $a \in \{1, \ldots, C\}$ correspond to picking one of the classes. Suppose we define the loss function as

$$L(y = j, a = i) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases} \tag{5.103}$$

where λ_r is the cost of the reject action, and λ_s is the cost of a substitution error. In Exercise 5.3, you will show that the optimal action is to pick the reject action if the most probable class has a probability below $1 - \frac{\lambda_r}{\lambda_s}$; otherwise you should just pick the most probable class.



(a) (b) (c)

Figure 5.14 (a-c). Plots of the $L(y,a) = |y-a|^q$ vs $|y-a|$ for $q = 0.2$, $q = 1$ and $q = 2$. Figure generated by `lossFunctionFig`.

5.7.1.3 Posterior mean minimizes ℓ_2 (quadratic) loss

For continuous parameters, a more appropriate loss function is **squared error**, ℓ_2 **loss**, or **quadratic loss**, defined as

$$L(y,a) = (y-a)^2 \tag{5.104}$$

The posterior expected loss is given by

$$\rho(a|\mathbf{x}) \quad = \quad \mathbb{E}\left[(y-a)^2|\mathbf{x}\right] = \mathbb{E}\left[y^2|\mathbf{x}\right] - 2a\mathbb{E}\left[y|\mathbf{x}\right] + a^2 \tag{5.105}$$

Hence the optimal estimate is the posterior mean:

$$\frac{\partial}{\partial a}\rho(a|\mathbf{x}) = -2\mathbb{E}\left[y|\mathbf{x}\right] + 2a = 0 \;\Rightarrow\; \hat{y} = \mathbb{E}\left[y|\mathbf{x}\right] = \int yp(y|\mathbf{x})dy \tag{5.106}$$

This is often called the **minimum mean squared error** estimate or **MMSE** estimate.

In a linear regression problem, we have

$$p(y|\mathbf{x},\boldsymbol{\theta}) = \mathcal{N}(y|\mathbf{x}^T\mathbf{w},\sigma^2) \tag{5.107}$$

In this case, the optimal estimate given some training data $\mathcal{D}$ is given by

$$\mathbb{E}\left[y|\mathbf{x},\mathcal{D}\right] = \mathbf{x}^T\mathbb{E}\left[\mathbf{w}|\mathcal{D}\right] \tag{5.108}$$

That is, we just plug-in the posterior mean parameter estimate. Note that this is the optimal thing to do no matter what prior we use for $\mathbf{w}$.

5.7.1.4 Posterior median minimizes ℓ_1 (absolute) loss

The ℓ_2 loss penalizes deviations from the truth quadratically, and thus is sensitive to outliers. A more robust alternative is the absolute or ℓ_1 **loss**, $L(y,a) = |y-a|$ (see Figure 5.14). The optimal estimate is the posterior median, i.e., a value a such that $P(y < a|\mathbf{x}) = P(y \geq a|\mathbf{x}) = 0.5$. See Exercise 5.9 for a proof.

5.7.1.5 Supervised learning

Consider a prediction function $\delta : \mathcal{X} \to \mathcal{Y}$, and suppose we have some cost function $\ell(y,y')$ which gives the cost of predicting y' when the truth is y. We can define the loss incurred by

taking action δ (i.e., using this predictor) when the unknown state of nature is $\boldsymbol{\theta}$ (the parameters of the data generating mechanism) as follows:

$$L(\boldsymbol{\theta}, \delta) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|\boldsymbol{\theta})}\left[\ell(y, \delta(\mathbf{x})\right] = \sum_{\mathbf{x}} \sum_{y} L(y, \delta(\mathbf{x}))p(\mathbf{x}, y|\boldsymbol{\theta}) \tag{5.109}$$

This is known as the **generalization error**. Our goal is to minimize the posterior expected loss, given by

$$\rho(\delta|\mathcal{D}) = \int p(\boldsymbol{\theta}|\mathcal{D})L(\boldsymbol{\theta}, \delta)d\boldsymbol{\theta} \tag{5.110}$$

This should be contrasted with the frequentist risk which is defined in Equation 6.47.

5.7.2 The false positive vs false negative tradeoff

In this section, we focus on binary decision problems, such as hypothesis testing, two-class classification, object/ event detection, etc. There are two types of error we can make: a **false positive** (aka **false alarm**), which arises when we estimate $\hat{y} = 1$ but the truth is $y = 0$; or a **false negative** (aka **missed detection**), which arises when we estimate $\hat{y} = 0$ but the truth is $y = 1$. The 0-1 loss treats these two kinds of errors equivalently. However, we can consider the following more general loss matrix:

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	L_{FN}
$y = 0$	L_{FP}	0

where L_{FN} is the cost of a false negative, and L_{FP} is the cost of a false positive. The posterior expected loss for the two possible actions is given by

$$\rho(\hat{y} = 0|\mathbf{x}) = L_{FN}\ p(y = 1|\mathbf{x}) \tag{5.111}$$

$$\rho(\hat{y} = 1|\mathbf{x}) = L_{FP}\ p(y = 0|\mathbf{x}) \tag{5.112}$$

Hence we should pick class $\hat{y} = 1$ iff

$$\rho(\hat{y} = 0|\mathbf{x}) > \rho(\hat{y} = 1|\mathbf{x}) \tag{5.113}$$

$$\frac{p(y = 1|\mathbf{x})}{p(y = 0|\mathbf{x})} > \frac{L_{FP}}{L_{FN}} \tag{5.114}$$

If $L_{FN} = cL_{FP}$, it is easy to show (Exercise 5.10) that we should pick $\hat{y} = 1$ iff $p(y = 1|\mathbf{x})/p(y = 0|\mathbf{x}) > \tau$, where $\tau = c/(1 + c)$ (see also (Muller et al. 2004)). For example, if a false negative costs twice as much as false positive, so $c = 2$, then we use a decision threshold of $2/3$ before declaring a positive.

Below we discuss ROC curves, which provide a way to study the FP-FN tradeoff without having to choose a specific threshold.

5.7.2.1 ROC curves and all that

Suppose we are solving a binary decision problem, such as classification, hypothesis testing, object detection, etc. Also, assume we have a labeled data set, $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$. Let $\delta(\mathbf{x}) =$

		Truth		
		1	0	Σ
Estimate	1	TP	FP	$\hat{N}_+ = TP + FP$
	0	FN	TN	$\hat{N}_- = FN + TN$
	Σ	$N_+ = TP + FN$	$N_- = FP + TN$	$N = TP + FP + FN + TN$

Table 5.2 Quantities derivable from a confusion matrix. N_+ is the true number of positives, $\hat{N}_+$ is the "called" number of positives, N_- is the true number of negatives, $\hat{N}_-$ is the "called" number of negatives.

	$y = 1$	$y = 0$
$\hat{y} = 1$	TP/N_+=TPR=sensitivity=recall	FP/N_-=FPR=type I
$\hat{y} = 0$	FN/N_+=FNR=miss rate=type II	TN/N_-=TNR=specifity

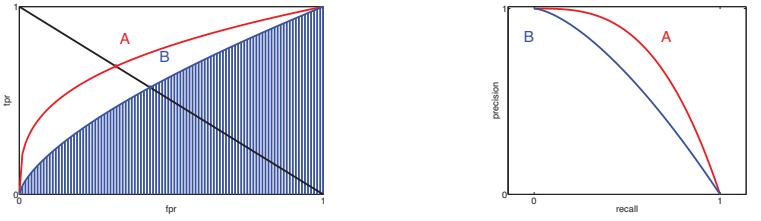
Table 5.3 Estimating $p(\hat{y}|y)$ from a confusion matrix. Abbreviations: FNR = false negative rate, FPR = false positive rate, TNR = true negative rate, TPR = true positive rate.

$\mathbb{I}(f(\mathbf{x}) > \tau)$ be our decision rule, where $f(\mathbf{x})$ is a measure of confidence that $y = 1$ (this should be monotonically related to $p(y = 1|\mathbf{x})$, but does not need to be a probability), and τ is some threshold parameter. For each given value of τ, we can apply our decision rule and count the number of true positives, false positives, true negatives, and false negatives that occur, as shown in Table 5.2. This table of errors is called a **confusion matrix**.

From this table, we can compute the **true positive rate** (TPR), also known as the **sensitivity**, **recall** or **hit rate**, by using $TPR = TP/N_+ \approx p(\hat{y} = 1|y = 1)$. We can also compute the **false positive rate** (FPR), also called the **false alarm rate**, or the **type I error rate**, by using $FPR = FP/N_- \approx p(\hat{y} = 1|y = 0)$. These and other definitions are summarized in Tables 5.3 and 5.4. We can combine these errors in any way we choose to compute a loss function.

However, rather than than computing the TPR and FPR for a fixed threshold τ, we can run our detector for a set of thresholds, and then plot the TPR vs FPR as an implicit function of τ. This is called a **receiver operating characteristic** or **ROC** curve. See Figure 5.15(a) for an example. Any system can achieve the point on the bottom left, $(FPR = 0, TPR = 0)$, by setting $\tau = 1$ and thus classifying everything as negative; similarly any system can achieve the point on the top right, $(FPR = 1, TPR = 1)$, by setting $\tau = 0$ and thus classifying everything as positive. If a system is performing at chance level, then we can achieve any point on the diagonal line $TPR = FPR$ by choosing an appropriate threshold. A system that perfectly separates the positives from negatives has a threshold that can achieve the top left corner, $(FPR = 0, TPR = 1)$; by varying the threshold such a system will "hug" the left axis and then the top axis, as shown in Figure 5.15(a).

The quality of a ROC curve is often summarized as a single number using the **area under the curve** or **AUC**. Higher AUC scores are better; the maximum is obviously 1. Another summary statistic that is used is the **equal error rate** or **EER**, also called the **cross over rate**, defined as the value which satisfies $FPR = FNR$. Since $FNR = 1 - TPR$, we can compute the EER by drawing a line from the top left to the bottom right and seeing where it intersects the ROC curve (see points A and B in Figure 5.15(a)). Lower EER scores are better; the minimum is obviously 0.



(a) (b)

Figure 5.15 (a) ROC curves for two hypothetical classification systems. A is better than B. We plot the true positive rate (TPR) vs the false positive rate (FPR) as we vary the threshold τ. We also indicate the equal error rate (EER) with the red and blue dots, and the area under the curve (AUC) for classifier B. (b) A precision-recall curve for two hypothetical classification systems. A is better than B. Figure generated by `PRhand`.

	$y=1$	$y=0$
$\hat{y}=1$	$TP/\hat{N}_+$=precision=PPV	$FP/\hat{N}_+$=FDP
$\hat{y}=0$	$FN/\hat{N}_-$	$TN/\hat{N}_-$=NPV

Table 5.4 Estimating $p(y|\hat{y})$ from a confusion matrix. Abbreviations: FDP = false discovery probability, NPV = negative predictive value, PPV = positive predictive value,

5.7.2.2 Precision recall curves

When trying to detect a rare event (such as retrieving a relevant document or finding a face in an image), the number of negatives is very large. Hence comparing $TPR = TP/N_+$ to $FPR = FP/N_-$ is not very informative, since the FPR will be very small. Hence all the "action" in the ROC curve will occur on the extreme left. In such cases, it is common to plot the TPR versus the number of false positives, rather than vs the false positive rate.

However, in some cases, the very notion of "negative" is not well-defined. For example, when detecting objects in images (see Section 1.2.1.3), if the detector works by classifying patches, then the number of patches examined — and hence the number of true negatives — is a parameter of the algorithm, not part of the problem definition. So we would like to use a measure that only talks about positives.

The **precision** is defined as $TP/\hat{N}_+ = p(y=1|\hat{y}=1)$ and the **recall** is defined as $TP/N_+ = p(\hat{y}=1|y=1)$. Precision measures what fraction of our detections are actually positive, and recall measures what fraction of the positives we actually detected. If $\hat{y}_i \in \{0,1\}$ is the predicted label, and $y_i \in \{0,1\}$ is the true label, we can estimate precision and recall using

$$P = \frac{\sum_i y_i \hat{y}_i}{\sum_i \hat{y}_i}, \; R = \frac{\sum_i y_i \hat{y}_i}{\sum_i y_i} \tag{5.115}$$

A **precision recall curve** is a plot of precision vs recall as we vary the threshold τ. See Figure 5.15(b). Hugging the top right is the best one can do.

This curve can be summarized as a single number using the mean precision (averaging over

	Class 1	
	$y=1$	$y=0$
$\hat{y}=1$	10	10
$\hat{y}=0$	10	970

	Class 2	
	$y=1$	$y=0$
$\hat{y}=1$	90	10
$\hat{y}=0$	10	890

	Pooled	
	$y=1$	$y=0$
$\hat{y}=1$	100	20
$\hat{y}=0$	20	1860

Table 5.5 Illustration of the difference between macro- and micro-averaging. y is the true label, and $\hat{y}$ is the called label. In this example, the macro-averaged precision is $[10/(10+10)+90/(10+90)]/2 = (0.5+0.9)/2 = 0.7$. The micro-averaged precision is $100/(100+20) \approx 0.83$. Based on Table 13.7 of (Manning et al. 2008).

recall values), which approximates the area under the curve. Alternatively, one can quote the precision for a fixed recall level, such as the precision of the first $K = 10$ entities recalled. This is called the **average precision at K** score. This measure is widely used when evaluating information retrieval systems.

5.7.2.3 F-scores *

For a fixed threshold, one can compute a single precision and recall value. These are often combined into a single statistic called the **F score**, or **F1 score**, which is the harmonic mean of precision and recall:

$$F_1 \triangleq \frac{2}{1/P + 1/R} = \frac{2PR}{R+P} \tag{5.116}$$

Using Equation 5.115, we can write this as

$$F_1 = \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i} \tag{5.117}$$

This is a widely used measure in information retrieval systems.

To understand why we use the harmonic mean instead of the arithmetic mean, $(P+R)/2$, consider the following scenario. Suppose we recall all entries, so $R = 1$. The precision will be given by the **prevalence**, $p(y=1)$. Suppose the prevalence is low, say $p(y=1) = 10^{-4}$. The arithmetic mean of P and R is given by $(P+R)/2 = (10^{-4}+1)/2 \approx 50\%$. By contrast, the harmonic mean of this strategy is only $\frac{2\times 10^{-4}\times 1}{1+10^{-4}} \approx 0.2\%$.

In the multi-class case (e.g., for document classification problems), there are two ways to generalize F_1 scores. The first is called **macro-averaged F1**, and is defined as $\sum_{c=1}^{C} F_1(c)/C$, where $F_1(c)$ is the F_1 score obtained on the task of distinguishing class c from all the others. The other is called **micro-averaged F1**, and is defined as the F_1 score where we pool all the counts from each class's contingency table.

Table 5.5 gives a worked example that illustrates the difference. We see that the precision of class 1 is 0.5, and of class 2 is 0.9. The macro-averaged precision is therefore 0.7, whereas the micro-averaged precision is 0.83. The latter is much closer to the precision of class 2 than to the precision of class 1, since class 2 is five times larger than class 1. To give equal weight to each class, use macro-averaging.

5.7.2.4 False discovery rates *

Suppose we are trying to discover a rare phenomenon using some kind of high throughput measurement device, such as a gene expression micro array, or a radio telescope. We will need to make many binary decisions of the form $p(y_i = 1|\mathcal{D}) > \tau$, where $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ and N may be large. This is called **multiple hypothesis testing**. Note that the difference from standard binary classification is that we are classifying y_i based on all the data, not just based on $\mathbf{x}_i$. So this is a simultaneous classification problem, where we might hope to do better than a series of individual classification problems.

How should we set the threshold τ? A natural approach is to try to minimize the expected number of false positives. In the Bayesian approach, this can be computed as follows:

$$FD(\tau, \mathcal{D}) \triangleq \sum_i \underbrace{(1 - p_i)}_{\text{pr. error}} \underbrace{\mathbb{I}(p_i > \tau)}_{\text{discovery}} \tag{5.118}$$

where $p_i \triangleq p(y_i = 1|\mathcal{D})$ is your belief that this object exhibits the phenomenon in question. We then define the posterior expected **false discovery rate** as follows:

$$FDR(\tau, \mathcal{D}) \triangleq FD(\tau, \mathcal{D})/N(\tau, \mathcal{D}) \tag{5.119}$$

where $N(\tau, \mathcal{D}) = \sum_i \mathbb{I}(p_i > \tau)$ is the number of discovered items. Given a desired FDR tolerance, say $\alpha = 0.05$, one can then adapt τ to achieve this; this is called the **direct posterior probability approach** to controlling the FDR (Newton et al. 2004; Muller et al. 2004).

In order to control the FDR it is very helpful to estimate the p_i's jointly (e.g., using a hierarchical Bayesian model, as in Section 5.5), rather than independently. This allows the pooling of statistical strength, and thus lower FDR. See e.g., (Berry and Hochberg 1999) for more information.

5.7.3 Other topics *

In this section, we briefly mention a few other topics related to Bayesian decision theory. We do not have space to go into detail, but we include pointers to the relevant literature.

5.7.3.1 Contextual bandits

A **one-armed bandit** is a colloquial term for a slot machine, found in casinos around the world. The game is this: you insert some money, pull an arm, and wait for the machine to stop; if you're lucky, you win some money. Now imagine there is a bank of K such machines to choose from. Which one should you use? This is called a **multi-armed bandit**, and can be modeled using Bayesian decision theory: there are K possible actions, and each action has an unknown reward (payoff function) r_k. By maintaining a belief state, $p(r_{1:K}|\mathcal{D}) = \prod_k p(r_k|\mathcal{D})$, one can devise an optimal policy; this can be compiled into a series of **Gittins Indices** (Gittins 1989). This optimally solves the **exploration-exploitation** tradeoff, which specifies how many times one should try each action before deciding to go with the winner.

Now consider an extension where each arm, and the player, has an associated feature vector; call all these features $\mathbf{x}$. This is called a **contextual bandit** (see e.g., (Sarkar 1991; Scott 2010; Li et al. 2011)). For example, the "arms" could represent ads or news articles which we want to show to the user, and the features could represent properties of these ads or articles, such

as a bag of words, as well as properties of the user, such as demographics. If we assume a linear model for reward, $r_k = \boldsymbol{\theta}_k^T \mathbf{x}$, we can maintain a distribution over the parameters of each arm, $p(\boldsymbol{\theta}_k|\mathcal{D})$, where $\mathcal{D}$ is a series of tuples of the form $(a, \mathbf{x}, r)$, which specifies which arm was pulled, what its features were, and what the resulting outcome was (e.g., $r = 1$ if the user clicked on the ad, and $r = 0$ otherwise). We discuss ways to compute $p(\boldsymbol{\theta}_k|\mathcal{D})$ from linear and logistic regression models in later chapters.

Given the posterior, we must decide what action to take. One common heuristic, known as **UCB** (which stands for "upper confidence bound") is to take the action which maximizes

$$k^* = \underset{k=1}{\overset{K}{\operatorname{argmax}}} \mu_k + \lambda \sigma_k \tag{5.120}$$

where $\mu_k = \mathbb{E}[r_k|\mathcal{D}]$, $\sigma_k^2 = \operatorname{var}[r_k|\mathcal{D}]$ and λ is a tuning parameter that trades off exploration and exploitation. The intuition is that we should pick actions about which we believe are good (μ_k is large), and/ or actions about which we are uncertain (σ_k is large).

An even simpler method, known as **Thompson sampling**, is as follows. At each step, we pick action k with a probability that is equal to its probability of being the optimal action:

$$p_k = \int \mathbb{I}(\mathbb{E}[r|a, \mathbf{x}, \boldsymbol{\theta}] = \max_{a'} \mathbb{E}[r|a', \mathbf{x}, \boldsymbol{\theta}]) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{5.121}$$

We can approximate this by drawing a single sample from the posterior, $\boldsymbol{\theta}^t \sim p(\boldsymbol{\theta}|\mathcal{D})$, and then choosing $k^* = \operatorname{argmax}_k \mathbb{E}\left[r|\mathbf{x}, k, \boldsymbol{\theta}^t\right]$. Despite its simplicity, this has been shown to work quite well (Chapelle and Li 2011).

5.7.3.2 Utility theory

Suppose we are a doctor trying to decide whether to operate on a patient or not. We imagine there are 3 states of nature: the patient has no cancer, the patient has lung cancer, or the patient has breast cancer. Since the action and state space is discrete, we can represent the loss function $L(\theta, a)$ as a **loss matrix**, such as the following:

	Surgery	No surgery
No cancer	20	0
Lung cancer	10	50
Breast cancer	10	60

These numbers reflects the fact that not performing surgery when the patient has cancer is very bad (loss of 50 or 60, depending on the type of cancer), since the patient might die; not performing surgery when the patient does not have cancer incurs no loss (0); performing surgery when the patient does not have cancer is wasteful (loss of 20); and performing surgery when the patient does have cancer is painful but necessary (10).

It is natural to ask where these numbers come from. Ultimately they represent the personal **preferences** or values of a fictitious doctor, and are somewhat arbitrary: just as some people prefer chocolate ice cream and others prefer vanilla, there is no such thing as the "right" loss/ utility function. However, it can be shown (see e.g., (DeGroot 1970)) that any set of consistent preferences can be converted to a scalar loss/ utility function. Note that utility can be measured on an arbitrary scale, such as dollars, since it is only relative values that matter.[6]

6. People are often squeamish about talking about human lives in monetary terms, but all decision making requires

5.7.3.3 Sequential decision theory

So far, we have concentrated on **one-shot decision problem**s, where we only have to make one decision and then the game ends. In Setion 10.6, we will generalize this to multi-stage or sequential decision problems. Such problems frequently arise in many business and engineering settings. This is closely related to the problem of reinforcement learning. However, further discussion of this point is beyond the scope of this book.

Exercises

Exercise 5.1 Proof that a mixture of conjugate priors is indeed conjugate

Derive Equation 5.69.

Exercise 5.2 Optimal threshold on classification probability

Consider a case where we have learned a conditional probability distribution $P(y|\mathbf{x})$. Suppose there are only two classes, and let $p_0 = P(Y = 0|\mathbf{x})$ and $p_1 = P(Y = 1|\mathbf{x})$. Consider the loss matrix below:

predicted label $\hat{y}$	true label y	
	0	1
0	0	λ_{01}
1	λ_{10}	0

a. Show that the decision $\hat{y}$ that minimizes the expected loss is equivalent to setting a probability threshold θ and predicting $\hat{y} = 0$ if $p_1 < \theta$ and $\hat{y} = 1$ if $p_1 \geq \theta$. What is θ as a function of λ_{01} and λ_{10}? (Show your work.)

b. Show a loss matrix where the threshold is 0.1. (Show your work.)

Exercise 5.3 Reject option in classifiers

(Source: (Duda et al. 2001, Q2.13).)

In many classification problems one has the option either of assigning $\mathbf{x}$ to class j or, if you are too uncertain, of choosing the **reject option**. If the cost for rejects is less than the cost of falsely classifying the object, it may be the optimal action. Let α_i mean you choose action i, for $i = 1 : C + 1$, where C is the number of classes and $C + 1$ is the reject action. Let $Y = j$ be the true (but unknown) **state of nature**. Define the loss function as follows

$$\lambda(\alpha_i|Y = j) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases} \tag{5.122}$$

In otherwords, you incur 0 loss if you correctly classify, you incur λ_r loss (cost) if you choose the reject option, and you incur λ_s loss (cost) if you make a substitution error (misclassification).

---

tradeoffs, and one needs to use some kind of "currency" to compare different courses of action. Insurance companies do this all the time. Ross Schachter, a decision theorist at Stanford University, likes to tell a story of a school board who rejected a study on absestos removal from schools because it performed a **cost-benefit analysis**, which was considered "inhumane" because they put a dollar value on children's health; the result of rejecting the report was that the absestos was not removed, which is surely more "inhumane". In medical domains, one often measures utility in terms of **QALY**, or quality-adjusted life-years, instead of dollars, but it's the same idea. Of course, even if you do not explicitly specify how much you value different people's lives, your *behavior* will reveal your implicit values/ preferences, and these preferences can then be converted to a real-valued scale, such as dollars or QALY. Inferring a utility function from behavior is called **inverse reinforcement learning**.

Decision	true label y	
$\hat{y}$	0	1
predict 0	0	10
predict 1	10	0
reject	3	3

a. Show that the minimum risk is obtained if we decide $Y = j$ if $p(Y = j|\mathbf{x}) \geq p(Y = k|\mathbf{x})$ for all k (i.e., j is the most probable class) *and* if $p(Y = j|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$; otherwise we decide to reject.

b. Describe qualitatively what happens as λ_r/λ_s is increased from 0 to 1 (i.e., the relative cost of rejection increases).

Exercise 5.4 More reject options

In many applications, the classifier is allowed to "reject" a test example rather than classifying it into one of the classes. Consider, for example, a case in which the cost of a misclassification is \$10 but the cost of having a human manually make the decison is only \$3. We can formulate this as the following loss matrix:

a. Suppose $P(y = 1|\mathbf{x})$ is predicted to be 0.2. Which decision minimizes the expected loss?

b. Now suppose $P(y = 1|\mathbf{x})$=0.4. Now which decision minimizes the expected loss?

c. Show that in general, for this loss matrix, but for any posterior distribution, there will be two thresholds θ_0 and θ_1 such that the optimal decisionn is to predict 0 if $p_1 < \theta_0$, reject if $\theta_0 \leq p_1 \leq \theta_1$, and predict 1 if $p_1 > \theta_1$ (where $p_1 = p(y = 1|\mathbf{x})$). What are these thresholds?

Exercise 5.5 Newsvendor problem

Consider the following classic problem in decision theory/ economics. Suppose you are trying to decide how much quantity Q of some product (e.g., newspapers) to buy to maximize your profits. The optimal amount will depend on how much demand D you think there is for your product, as well as its cost to you C and its selling price P. Suppose D is unknown but has pdf $f(D)$ and cdf $F(D)$. We can evaluate the expected profit by considering two cases: if $D > Q$, then we sell all Q items, and make profit $\pi = (P - C)Q$; but if $D < Q$, we only sell D items, at profit $(P - C)D$, but have wasted $C(Q - D)$ on the unsold items. So the expected profit if we buy quantity Q is

$$E\pi(Q) = \int_Q^\infty (P - C)Qf(D)dD + \int_0^Q (P - C)Df(D) - \int_0^Q C(Q - D)f(D)dD \tag{5.123}$$

Simplify this expression, and then take derivatives wrt Q to show that the optimal quantity Q^* (which maximizes the expected profit) satisfies

$$F(Q^*) = \frac{P - C}{P} \tag{5.124}$$

Exercise 5.6 Bayes factors and ROC curves

Let $B = p(D|H_1)/p(D|H_0)$ be the bayes factor in favor of model 1. Suppose we plot two ROC curves, one computed by thresholding B, and the other computed by thresholding $p(H_1|D)$. Will they be the same or different? Explain why.

Exercise 5.7 Bayes model averaging helps predictive accuracy

Let Δ be a quantity that we want to predict, let $\mathcal{D}$ be the observed data and $\mathcal{M}$ be a finite set of models. Suppose our action is to provide a probabilistic prediction $p()$, and the loss function is $L(\Delta, p()) =$

$-\log p(\Delta)$. We can either perform Bayes model averaging and predict using

$$p^{BMA}(\Delta) = \sum_{m \in \mathcal{M}} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \tag{5.125}$$

or we could predict using any single model (a plugin approximation)

$$p^m(\Delta) = p(\Delta|m, \mathcal{D}) \tag{5.126}$$

Show that, for all models $m \in \mathcal{M}$, the posterior expected loss using BMA is lower, i.e.,

$$\mathbb{E}\left[L(\Delta, p^{BMA})\right] \leq \mathbb{E}\left[L(\Delta, p^m)\right] \tag{5.127}$$

where the expectation over Δ is with respect to

$$p(\Delta|\mathcal{D}) = \sum_{m \in \mathcal{M}} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \tag{5.128}$$

Hint: use the non-negativity of the KL divergence.

Exercise 5.8 MLE and model selection for a 2d discrete distribution

(Source: Jaakkola.)

Let $x \in \{0, 1\}$ denote the result of a coin toss ($x = 0$ for tails, $x = 1$ for heads). The coin is potentially biased, so that heads occurs with probability θ_1. Suppose that someone else observes the coin flip and reports to you the outcome, y. But this person is unreliable and only reports the result correctly with probability θ_2; i.e., $p(y|x, \theta_2)$ is given by

	$y = 0$	$y = 1$
$x = 0$	θ_2	$1 - \theta_2$
$x = 1$	$1 - \theta_2$	θ_2

Assume that θ_2 is independent of x and θ_1.

a. Write down the joint probability distribution $p(x, y|\boldsymbol{\theta})$ as a 2×2 table, in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

b. Suppose have the following dataset: $\mathbf{x} = (1, 1, 0, 1, 1, 0, 0)$, $\mathbf{y} = (1, 0, 0, 0, 1, 0, 1)$. What are the MLEs for θ_1 and θ_2? Justify your answer. Hint: note that the likelihood function factorizes,

$$p(x, y|\boldsymbol{\theta}) = p(y|x, \theta_2) p(x|\theta_1) \tag{5.129}$$

What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_2)$ where M_2 denotes this 2-parameter model? (You may leave your answer in fractional form if you wish.)

c. Now consider a model with 4 parameters, $\boldsymbol{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$, representing $p(x, y|\boldsymbol{\theta}) = \theta_{x,y}$. (Only 3 of these parameters are free to vary, since they must sum to one.) What is the MLE of $\boldsymbol{\theta}$? What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4)$ where M_4 denotes this 4-parameter model?

d. Suppose we are not sure which model is correct. We compute the leave-one-out cross validated log likelihood of the 2-parameter model and the 4-parameter model as follows:

$$L(m) = \sum_{i=1}^{n} \log p(x_i, y_i|m, \hat{\theta}(\mathcal{D}_{-i})) \tag{5.130}$$

and $\hat{\theta}(\mathcal{D}_{-i}))$ denotes the MLE computed on $\mathcal{D}$ excluding row i. Which model will CV pick and why? Hint: notice how the table of counts changes when you omit each training case one at a time.

e. Recall that an alternative to CV is to use the BIC score, defined as

$$\text{BIC}(M, \mathcal{D}) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}) - \frac{\text{dof}(M)}{2} \log N \tag{5.131}$$

where $\text{dof}(M)$ is the number of free parameters in the model, Compute the BIC scores for both models (use log base e). Which model does BIC prefer?

Exercise 5.9 Posterior median is optimal estimate under L1 loss

Prove that the posterior median is optimal estimate under L1 loss.

Exercise 5.10 Decision rule for trading off FPs and FNs

If $L_{FN} = cL_{FP}$, show that we should pick $\hat{y} = 1$ iff $p(y = 1|\mathbf{x})/p(y = 0|\mathbf{x}) > \tau$, where $\tau = c/(1 + c)$