Chapter 2

Priors, Predictions and Model Choice

2.1 Prior Distributions

2.1.1 Conjugate Priors

Suppose that we have a hierarchical model $f(\mathbf{x}|\theta)$: the likelihood; $\pi(\theta|\eta)$ the prior. If $\pi(\theta|\mathbf{x},\eta)$ belongs to the same parametric family as $\pi(\theta|\eta)$, then we say that $\pi(\theta|\eta)$ is a conjugate prior for θ. In these cases if we assume that η is known, the analysis becomes much easier. Natural conjugacies:

Likelihood	Prior
Binomial	Beta
Poisson	Gamma
Normal	Normal
Exponential	Gamma

2.1.2 Locally Uniform Priors

What if one has no prior information with which to choose $\pi(\theta)$? Although this is rare in practice, this type situations can be overcome by the use of what are called non-informative (vague, diffuse, flat) priors.

A basic property of a pdf is that it integrates to 1, i.e. $\int_{-\infty}^{\infty} \pi(\theta)d\theta = 1$. Sometimes we assume

prior distributions which are constant over the whole real line. For example,

$$\pi(\theta) = k,\ k > 0,\ -\infty < \theta < \infty.$$

This pdf violates the above condition. This would be called an **improper** prior distribution. It is alright to assume improper prior distributions only if the resulting posterior distribution is proper, i.e. $\int_{-\infty}^{\infty} \pi(\theta|\mathbf{x})d\theta < \infty$. Further, suppose that $\pi(\theta) = k$ only for values of θ where the likelihood function has appreciable value, and $\pi(\theta) = 0$ otherwise. This $\pi(\theta)$ will then define a proper density and no theoretical problem arises. Prior distributions like the above are called locally uniform priors.

2.1.3 Non-informative priors

If a prior distribution $\pi(\theta)$ does not contain any information for θ, it is called a *non-informative prior*. Most widely used non-informative priors are Jeffreys (1961) priors:

$$\pi(\theta) = \sqrt{I(\theta)} \tag{2.1}$$

where $I(\boldsymbol{\theta})$ is the Fisher information

$$I(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{x}|\theta) \right].$$

Note that we obtain improper priors in most situation. We have to guarantee that the resulting posterior is **proper**.

Why does Jeffreys prior (2.1) give a non-informative prior? The answer in brief is the following. The above prior induces a one-to-one function $\phi = \phi(\theta)$ for which the prior pdf of ϕ, $\pi(\phi) \propto 1$. That is, for ϕ the induced prior is locally uniform or non-informative, hence the prior (2.1) for θ (which is a one-to-one function of ϕ) is also non-informative. It does not always suffice to take $\pi(\theta) \propto 1$ as the non-informative prior since 'information' is relative to the sampling experiment, i.e. the likelihood function $f(\mathbf{x}\ \theta)$.

There is a huge literature on prior selection. Box and Tiao (Section 1.3) would be a good start. In our course we will assume (loosely) vague or flat priors, i.e., priors which are locally uniform.

♡ **Example** 2.1. For the binomial example show that

$$\pi(\theta) = \{\theta(1-\theta)\}^{-\frac{1}{2}}.$$

□

♡ **Example** 2.2. **Normal** For $N(0, \sigma^2)$ problem show that

$$\pi(\sigma^2) = \frac{1}{\sigma^2}.$$

□

2.2 Predictive Distributions

"What is the probability that the sun will rise tomorrow, given that it has risen without fail for the last n days?" In order to answer questions like these we need to learn what are called predictive distributions.

2.2.1 Posterior Predictive Distribution

Let $X_1, \ldots, X_n$ be an i.i.d. sample from the distribution $f(x|\theta)$. Let $\pi(\theta)$ be the prior distribution and $\pi(\theta|\mathbf{x})$ be the posterior distribution. We want the distribution (pdf or pmf) of $X_{n+1}|X_1, \ldots, X_n$. The given notation is to denote that $X_1, \ldots, X_n$ have already been observed, like the sun has risen for the last n days. We define the **posterior predictive distribution** to be:

$$f(x_{n+1}|x_1, \ldots, x_n) = \int_{-\infty}^{\infty} f(x_{n+1}|\theta)\ \pi(\theta|x_1, \ldots, x_n)\ d\theta. \tag{2.2}$$

It is the density of a future observation given everything else, i.e., the 'model' and the observations. (The model is really the function $f(x|\theta)$.) Intuitively, if θ is known then x_{n+1} will follow $f(x_{n+1}|\theta)$ since it is from the same population as $x_1, \ldots, x_n$ are. We do not know θ but the posterior $\pi(\theta|\mathbf{x})$ contains all that we know about θ. Therefore, the predictive distribution is obtained as an average over $\pi(\theta|\mathbf{x})$. Hence the definition. We now derive some predictive distributions.

♡ **Example** 2.3. We return to the sun example. Let

$$\begin{aligned} X_i &= 1 \quad \text{if its sunny on the } i\text{th day,} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Note that X_{n+1} will be binary as well. We want $P[X_{n+1} = 1|\mathbf{x} = (1, 1, \ldots, 1)]$. Assume $f(x_i|\theta) =$

$\theta^{x_i}(1-\theta)^{1-x_i}$, and X_i are independent. Therefore, the likelihood is

$$\begin{aligned} f(\mathbf{x}|\theta) &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}, \\ &= \theta^n \text{ if } \mathbf{x} = (1,1,\ldots,1). \end{aligned}$$

Let us assume a uniform prior for θ, i.e. $\pi(\theta) = 1$ if $0 < \theta < 1$. Now the posterior is:

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{\theta^n}{\int_0^1 \theta^n d\theta} \\ &= (n+1)\theta^n. \end{aligned}$$

Here $f(X_{n+1} = 1|\theta) = \theta$. Finally we can evaluate the posterior predictive distribution using (2.2).

$$\begin{aligned} P(X_{n+1} = 1|\mathbf{x}) &= \int_0^1 \theta(n+1)\theta^n d\theta \\ &= (n+1)\int_0^1 \theta^{n+1} d\theta \\ &= \frac{n+1}{n+2}. \end{aligned}$$

Intuitively, this probability goes to 1 as $n \to \infty$. □

Exercises: The above example assumes that $x_1, \ldots, x_n$ are all 1. Re-do this without assuming the specific values.

♡ **Example** 2.4. We return to the normal example. Suppose $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, $\pi(\theta) \sim N(\mu, \tau^2)$ for known μ and τ^2. We had

$$\pi(\theta|\mathbf{x}) = N\left(\frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right).$$

Satisfy yourself that $X_{n+1}|\mathbf{x}$ follows the normal distribution with

$$\text{mean } = \frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \text{ and variance } \sigma^2 + \frac{1}{n/\sigma^2 + 1/\tau^2}.$$

□

2.2.2 Prior Predictive Distribution

We sometimes need to define what is called the **prior predictive distribution** defined as

$$f(x) = \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)\, d\theta. \tag{2.3}$$

Note that it is simply the normalising constant in $\pi(\theta|x)$. It is also called the marginal distribution of the data. And it is of the same form as the posterior predictive distribution (2.2). The prior

predictive distribution is obtained by replacing the posterior $\pi(\theta|x_1,\ldots,x_n)$ by the prior $\pi(\theta)$ in (2.2).

With n samples, we define the (**joint**) prior predictive distribution of $x_1,\ldots,x_n$ as

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)\ \ \pi(\theta)\, d\theta. \tag{2.4}$$

♡ **Example** 2.5. For the normal-normal example, the prior predictive distribution is,

$$f(\mathbf{x}) = \int \prod_{i=1}^{n} N(x_i|\theta,\sigma^2) N(\theta|\mu,\tau^2) d\theta.$$

For this distribution show that $E(X_i) = \mu$ and $V(X_i) = \tau^2 + \sigma^2$. Are X_i & X_j marginally independent? No, they have covariance τ^2. (Derive it!)

You may find the following useful:

Result For any two random variable with finite variances:

$$E(X) = EE(X|Y), \quad \mathrm{Var}(X) = E\mathrm{Var}(X|Y) + \mathrm{Var}(E(X|Y)).$$

□

2.3 Model Choice

2.3.1 Bayes Factors

Suppose that we have to choose between two hypotheses H_0 and H_1 corresponding to assumptions of alternative models M_0 and M_1 for data $\mathbf{x}$. The likelihoods are denoted by $f_i(\mathbf{x}|\theta_i)$ and the priors by $\pi_i(\cdot), i = 0, 1$ in the following discussion. In many cases, the competing models have a common set of parameters, but this is not necessary; hence the notations f_i, π_i and θ_i. Recall that the prior predictive distribution (2.4) for model i is,

$$f(\mathbf{x}|M_i) = \int f_i(\mathbf{x}|\theta_i)\pi_i(\theta_i) d\theta_i.$$

Bayes factor is defined as:

$$B_{01}(\mathbf{x}) = \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)}. \tag{2.5}$$

Note that the Bayes factor is the ratio of the marginal likelihoods under two different models. Hence, intuitively $B_{01}(\mathbf{x}) > 1$ implies that M_0 is more relatively plausible in the light of $\mathbf{x}$. (Some authors use 3 as some sort of cut-off point.)

♡ **Example** 2.6. **Geometric versus Poisson** Suppose that:

$$M_0 : X_1, X_2, \ldots, X_n|\theta_0 \sim f_0(x|\theta_0) = \theta_0(1-\theta_0)^x, \quad x = 0, 1, \ldots .$$

$$M_1 : X_1, X_2, \ldots, X_n|\theta_1 \sim f_1(x|\theta_1) = e^{-\theta_1}\theta_1^x/x!, \quad x = 0, 1, \ldots .$$

Further, assume that θ_0 and θ_1 are known. How should we decide between the two models based on $x_1, x_2, \ldots, x_n$?

Since the parameters are known under the models, we do not need to assume any prior distributions for them. Consequently,

$$f(\mathbf{x}|M_0) = \theta_0^n(1-\theta_0)^{n\bar{x}}.$$

and

$$f(\mathbf{x}|M_1) = e^{-n\theta_1}\theta_1^{n\bar{x}}/\prod_{i=1}^{n} x_i!.$$

Now the Bayes factor is just the ratio of the above two. To illustrate, let $\theta_0 = 1/3$ and $\theta_1 = 2$ (then the two distributions have same mean). Now if $n = 2$ and $x_1 = x_2 = 0$ then $B_{01}(\mathbf{x}) = 6.1$, however if $n = 2$ and $x_1 = x_2 = 2$ then $B_{01}(\mathbf{x}) = 0.3$. □

Why it is called a factor? Let $P(M_i)$ denote the prior probability for model i. Let us now calculate the posterior probability of M_i given the data using the Bayes theorem.

$$P(M_i|\mathbf{x}) = \frac{P(M_i)f(\mathbf{x}|M_i)}{\sum_{j=0}^{1} P(M_j)f(\mathbf{x}|M_j)}.$$

So the posterior odds ratio of the two models is given by

$$\frac{P(M_0|\mathbf{x})}{P(M_1|\mathbf{x})} = \frac{P(M_0)}{P(M_1)} \times \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)}.$$

Now in words,

$$\textit{posterior odds ratio} = \textit{prior odds ratio} \times \textit{the Bayes factor}$$

That is why it is called a factor! Seen in this light we can define

$$\text{Bayes factor} = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}}.$$

Intuitively, the Bayes factor provides a measure of whether the data $\mathbf{x}$ have increased or decreased the odds on M_0 relative to M_1. Thus $B_{01}(\mathbf{x}) > 1$ signifies that M_0 is more relatively plausible in the light of $\mathbf{x}$.

Remark: We do not need to know the prior probabilities $P(M_i), i = 0, 1$ to calculate the Bayes factor. Those are needed if we wished to calculate the posterior probability $P(M_i|\mathbf{x})$. If two models are equally likely a-priori, then Bayes factor is equal to the posterior odds ratio.

2.3.2 Hypothesis Testing

Suppose that we wish to test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_A : \theta \in \Theta_1.$$

Let $f(\mathbf{x}|\theta)$ denote the likelihood of $\mathbf{x}$ given θ. Special forms:

$$\begin{aligned}
B_{01}(\mathbf{x}) &= \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} && \text{(simple versus simple test)} \\
B_{01}(\mathbf{x}) &= \frac{f(\mathbf{x}|\theta_0)}{\int_{\Theta_1} f(\mathbf{x}|\theta)\pi_1(\theta)d\theta} && \text{(simple versus composite test)} \\
B_{01}(\mathbf{x}) &= \frac{\int_{\Theta_0} f(\mathbf{x}|\theta)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(\mathbf{x}|\theta)\pi_1(\theta)d\theta} && \text{(composite versus composite test)}
\end{aligned}$$

♡ **Example** 2.7. **Taste-test** In an experiment to determine whether an individual possesses discriminating powers, she has to identify correctly which of the two brands she is provided with, over a series of trials.

Let θ denote the probability of her choosing the correct brand in any trial and X_i be the Bernoulli r.v. taking the value 1 for correct guess in the ith trial. Suppose that in first 6 trials the results are 1, 1, 1, 1, 1, 0.

We wish to test that the tester does not have any discriminatory power against the alternative that she does. So our problem is:

$$H_0 : \theta = \frac{1}{2} \text{ versus } H_1 : \theta > \frac{1}{2}.$$

It is a simple versus composite case and we have $\Theta_0 = \frac{1}{2}$ and $\Theta_1 = (\frac{1}{2}, 1)$. Let us assume uniform prior on θ under the alternative. So the prior $\pi_1(\theta) = 2$ if $\frac{1}{2} < \theta < 1$. Recall that we had 6 Bernoulli trials with the results, 1, 1, 1, 1, 1, 0. Now the Bayes factor is

$$B_{01}(\mathbf{x}) = \frac{\frac{1}{2}^6}{\int_{\frac{1}{2}}^1 \theta^5(1-\theta)2\,d\theta} = \frac{1}{2.86}.$$

This suggests that she does appear to have some discriminatory power but not a lot. □

2.3.3 P-value

This is often incorrectly interpreted as the *probability* that H_0 is true is smaller than the p-value. We have seen how to find such probability under the Bayesian setup.

♡ **Example** 2.8. **Return to the taste test** Suppose the problem is to test $H_0 : \theta = \frac{1}{2}$ against $H : \theta > \frac{1}{2}$. Here two cases arise.

Case 1: Suppose that n, the number of trials, is fixed in advance, that is binomial sampling distribution.

$$\begin{aligned} \text{p-value} &= P(X = 5 \text{ or something more extreme } |\theta = \tfrac{1}{2}) \\ &= P(X = 5 \text{ or } X = 6 | \theta = \tfrac{1}{2}) \\ &= 7 \times \left(\tfrac{1}{2}\right)^6 = 0.109. \end{aligned}$$

Case 2: Suppose that the sampling design is to continue the trials until first zero (geometric sampling).

$$\begin{aligned} \text{p-value} &= P(X = 5 \text{ or something more extreme } |\theta = \tfrac{1}{2}) \\ &= P(X = 5, 6, 7, \ldots | \theta = \tfrac{1}{2}) \\ &= \left(\tfrac{1}{2}\right)^6 + \left(\tfrac{1}{2}\right)^7 + \ldots \\ &= 0.031. \end{aligned}$$

Despite exactly the same sequence of events being observed, different inferences are made! □

2.3.4 Likelihood Principle

Consider two experiments yielding, respectively data $\mathbf{x}$ and $\mathbf{y}$ with model representation involving the same parameter $\boldsymbol{\theta} \in \Theta$ and proportional likelihoods:

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

The *likelihood principle* says that the experiments produce same conclusion about θ. It is a trivial consequence of the Bayes theorem if we assume the same prior for $\boldsymbol{\theta}$. However, the frequentist procedure typically violates the principle, since long run behavior under hypothetical repetitions depends on the entire distribution $\{f(\mathbf{x}|\boldsymbol{\theta}), \mathbf{x} \in \mathcal{X})\}$ where $\mathcal{X}$ is the sample space and not only on the likelihood. The pure likelihood approach, i.e., the attempt to produce inferences solely based on the likelihood function breaks down immediately when there are nuisance parameters.

2.4 Multi-parameter Situation

2.4.1 Basic Methods

In most realistic applications of statistical models, there are more than one unknown parameters. In principle, everything proceeds as before, except

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix},$$

where p is the number of parameters. We still start with

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}).$$

How do we summarise $\pi(\boldsymbol{\theta}|\mathbf{x})$? We use multi-variate statistical tools you are already familiar with. For example we can obtain the marginal posterior distribution of θ_1 as

$$\pi(\theta_1|\mathbf{x}) = \int \cdots \int \pi(\boldsymbol{\theta}|\mathbf{x})\, d\theta_2 d\theta_3 \ldots d\theta_p.$$

So we can calculate features of the above distribution, for example $E(\theta_1|\mathbf{x})$ and $\text{Var}(\theta_1|\mathbf{x})$. Also for example we can study correlations between θ_1 and θ_2.

♡ **Example** 2.9. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\theta, \sigma^2)$ and $\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$. Obtain the joint and the marginal posterior distributions of μ and σ^2.

♡ **Example** 2.10. **Pump Failure Data** The data set given below relates to 10 power plant pumps. The number of failures, y_i, follows a Poisson distribution with mean $\lambda_i = \theta_i t_i$ where θ_i is the failure rate for pump $i, i = 1, \ldots, 10$ and t_i is the length of operation time of the pump (in 1000s of hours). A conjugate gamma prior distribution with density $\beta^\alpha \theta_i^{\alpha-1} e^{-\beta\theta_i}/\Gamma(\alpha)$ is adopted for each θ_i, where $\alpha = 1.802$ and $\beta = 0.1$.

Obtain the marginal posterior distribution of θ_1.

□

Pump	1	2	3	4	5	6	7	8	9	10
t_i	94.3	15.7	62.9	126	5.24	31.4	1.05	1.05	2.1	10.5
y_i	5	1	5	14	3	19	1	1	4	22

□

2.4.2 Nuisance Parameters

Suppose we partition $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\eta})$ and we are interested in $\boldsymbol{\gamma}$. How do we proceed? Review classical methods. We are given $L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \mathbf{x})$. If we are lucky, there exists a sufficient statistics $\mathbf{T}$ such that

$$\mathbf{X}|\mathbf{T} \sim f(\mathbf{x}|\boldsymbol{\gamma}, \mathbf{T})$$

then base inference on the distribution of $\mathbf{T}$. This is called conditional inference. Otherwise, plug in maximum likelihood estimate $\hat{\boldsymbol{\eta}}(\boldsymbol{\gamma})$ of $\boldsymbol{\eta}$ in $L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \mathbf{x})$. This is the 'profile likelihood' technique. In a third scenario, if we are able to work out the integral,

$$\int L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \mathbf{x}) d\boldsymbol{\eta} = L(\boldsymbol{\gamma}; \mathbf{x}), \text{ say}$$

then use $L(\boldsymbol{\gamma}; \mathbf{x})$ to make inference. This is the 'marginal likelihood' technique.

From a Bayesian viewpoint, we have $\pi(\boldsymbol{\gamma}, \boldsymbol{\eta}|\mathbf{x})$. We use,

$$\pi(\boldsymbol{\gamma}|\mathbf{x}) = \int \pi(\boldsymbol{\gamma}, \boldsymbol{\eta}|\mathbf{x}) d\boldsymbol{\eta}.$$

This will be routinely done by numerical methods to be developed in the next chapter.