Chapter 7

Bayesian inference

7.1 Introduction

Bayesian inference (BI) gives the theoretical basis to Bayesian (statistical) methods the same way as frequentist (statistical) inference is the basis for frequentist (statistical) analysis. There are some philosophical and technical differences between frequentist (i.e. classical) and Bayesian approaches, but actually many parts of the inference are done similarly.

The philosophical difference is in the way the unknown parameters are interpreted. In frequentist inference the parameter is an unknown but a fixed constant, while in BI the parameter itself is a random variable. In what follows we do not concentrate on the philosophical differences that much, but give guidance to the technical procedure and theory behind BI.

The one formula behind the whole Bayesian standpoint is, of course, the Bayes formula as in Eq. (1.9). Let us write it here for continuous variables using pdf's:

$$\mathrm{f}_{\boldsymbol{\Theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\mathrm{f}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,\mathrm{f}_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta})}{\mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y})} = \frac{\mathrm{f}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,\mathrm{f}_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta})}{\int_{\Omega} \mathrm{f}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,\mathrm{f}_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta})\,d\boldsymbol{\theta}} \tag{7.1}$$

We explicitly write out here the random variables the different pdf's are referring to, but in what follows we will often shorten it, e.g. $\mathrm{f}_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta}) = \mathrm{f}(\boldsymbol{y}|\boldsymbol{\theta})$.

From the way Eq. (7.1) is written, one can immediately recognize the application to parameter estimation. The left side is the pdf of the unknown parameter vector $\boldsymbol{\theta}$, given that we have observed data $\boldsymbol{y}$. The left side is called the posterior distribution of the parameters. The numerator of the right side(s) is from the chain rule, it has both the prior distribution for the parameter, $\mathrm{f}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ and the distribution of data given the parameters, $\mathrm{f}_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta})$.

An important point in BI is that the denominator of Eq. (7.1) is often unnecessary to be known. The denominator is the (unconditional) distribution of the parameters. Definition $\mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_{\Omega} \mathrm{f}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,\mathrm{f}_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta})\,d\boldsymbol{\theta}$ uses the formula of total probability and integrates over the possible parameter space Ω. However, the denominator is constant with respect to θ. In fact, the role of the denominator is only to scale the resulting formula to pdf, i.e. to ensure that the volume of $\mathrm{f}_{\boldsymbol{\Theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}) = 1$.

In many applications the knowledge of properly scaled posterior distribution is not important. If you compare to the task of maximum likelihood parameter estimation with frequentist approach, one is only interested of the maximization of $\mathrm{f}(\boldsymbol{y};\boldsymbol{\theta})$, i.e. the probability density of data with given parameter value $\boldsymbol{\theta}$. In comparable BI case it is enough to know the unscaled posterior, $\mathrm{f}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,\mathrm{f}_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta})$. There is even closer connection to classical inference — if unscaled posterior is enough, we can use the likelihood function instead of the pdf. So, the version of the Bayes formula that is usually applied in BI is

$$\mathrm{f}(\boldsymbol{\theta}|\boldsymbol{y}) \;\propto\; \mathrm{f}(\boldsymbol{\theta})\,\mathrm{f}(\boldsymbol{y};\boldsymbol{\theta}) \;\propto\; \mathrm{f}(\boldsymbol{\theta})\,\mathrm{L}(\boldsymbol{\theta};\boldsymbol{y}) \tag{7.2}$$

7.2 Prior distributions

When comparing Eq. (7.2) to traditional maximum likelihood problems, one can see that the main difference is the presence of the prior distribution. Selecting prior pdf is subjective decision, that should of course be somehow justified by the researcher. In principle any pdf can be used as a prior pdf, or the prior does not even need to be a proper pdf, but there are some common approaches to the problem.

7.2.1 Conjugate prior distributions

Especially in times before efficient computers and easy-to-use software, the concept of conjugate prior (*liitännäispriori*) was important, since it allowed analytical, closed-form formulas to be derived. In short, a conjugate prior $\mathrm{f}(\boldsymbol{\theta})$ is such a distribution that the posterior $\mathrm{f}(\boldsymbol{\theta}|\boldsymbol{y})$ has the same distribution family as the prior. The selection of a conjugate prior is always related to the probability model of the data, $\mathrm{f}(\boldsymbol{y}|\boldsymbol{\theta})$.

The attractive benefit in using conjugate prior is that the results can be easily computed and interpreted, and the influence of both the data and the choice of parameters of prior distribution, i.e. *hyperparameters*, to the posterior parameters is clear. For example, if we conduct n independent Bernoulli trials with parameter (probability of success) π, and receive k positive outcomes, the likelihood model for the data is $\mathrm{L}(\pi;k) = \pi^k(1-\pi)^{1-k}$. Now, the Beta distribution is the conjugate prior for Bernoulli data. That means that if $\pi \sim \mathcal{B}(\alpha,\beta)$, then the posterior is also Beta-distribution but with some other parameters. Without calculating anything ourselves we can check from literature that the posterior is

$$\pi \,|\, k \sim \mathcal{B}(\alpha+k, \beta+n-k) \tag{7.3}$$

The complete Bayesian analysis of the case is now done and the result is compressed into the distribution and its parameters.

The simpleness of the conjugate prior approach is at the same time its shortcoming. The subjective choice of prior distribution is the key point in BI. In this era of efficient computing tools a conjugate prior should be used only if the prior would suit the case anyway, not just because the result is easy to derive and interpret. Lists of likelihood models with their prior distributions can be found in the literature, for example in Wikipedia. For the most common model of normal likelihood the prior distribution for the expectation parameter μ is also the normal distribution, and for variance σ^2 it is the inverse gamma distribution.

7.2.2 Uninformative prior distributions

Another common approach, or rather a framework of approaches, is the use of uninformative or vague priors. This means that if the researches does not have any particular information of the parameter *a priori* the observations, the uncertainty should be described in the prior. The idea is straightforward, but the practice might not be so simple to implement.

It is easy to think that if there is no knowledge of the location parameter, the μ for normal model for example, all the values of μ should be equally probable, $\mathrm{f}(\mu) \propto c$. So, the uninformative prior for μ should be the uniform distribution.

The first immediate problem is that the uniform distribution over the real axis is not a proper distribution since it does not integrate to one, it is a so-called improper prior. If the prior distribution is improper, the posterior is often also an improper distribution. However, in many BI analysis this problem can be avoided by using the form in Eq. (7.2) and deriving computational results by Monte Carlo or Markov chain Monte Carlo sampling. The recommended uninformative prior for scale parameter (i.e. variance) is of the form $\sigma^2 \propto 1/\sigma^2$

If the improper prior is not a problem, the reparametrization of the model might arise new problems. Reparametrization means that the original parameter of the model is transformed by some function. In many physical models it is possible to change from one set of parameters to another. For example astronomical coordinates can be defined in several ways. The reparametrization will also transform the shape of the prior distribution. It can easily happen that 'uniform' distribution in one parametrization will transform into something quite non-uniform in another parametrization.

It can be thought that the prior information should be invariant under parameter transformations. The prior that implements this principle is the Jeffreys prior. It has the form

$$\mathrm{f}(\boldsymbol{\theta}) \propto \sqrt{\det(\mathbf{I}(\boldsymbol{\theta}))}, \tag{7.4}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the so-called Fischer information matrix for the parameter $\boldsymbol{\theta}$. The Fischer information is the expectancy of the Hessian matrix $\mathbf{H}$ of the models second partial derivatives mentioned in Eq. (2.6). While Jeffreys prior solves the reparametrization problem, it is not always evident if the Jeffreys prior will describe the uncertainty in a meaningful way. With normal distribution and location parameter this is not the case, since the Jeffreys prior for that model is $\mathrm{f}(\mu) \propto c$.

Other common choices for uninformative priors, or at least for priors with very small amount of information, are proper distributions with very large variances so that they are 'almost flat' but still integrate to one. For example, with normal likelihood model the normal distribution itself is a conjugate distribution for the expectancy μ. If normal distribution with hyperparameter σ_0^2 is very large, the prior is almost flat but the posterior is a proper normal distribution.

7.2.3 Informative or subjective prior distributions

A criticism towards the use of uninformative priors is that, first, sometimes it can be difficult to actually express the lack of information as seen above. Second, BI with uninformative priors will actually give more or less the same result as the traditional frequentist approach since the results will only depend on the likelihood function of the data. Third, choosing an uninformative prior is also a subjective choice. Therefore, the most rewarding case for BI is when there actually is a priori information about the parameter and when that information can be represented in the form of a (prior) distribution.

In this case of subjective choice or prior distribution, a sensitivity analysis would often be a good idea. If the variance of the prior pdf is small, a lot of observations are needed to shift the posterior estimate away from the prior. The sensitivity of the posterior to observations is weak. If the variance of the prior is large, already a few observations can overdrive the prior information in the posterior, and the sensitivity to observations is strong. Often it needs some numerical tests to assure that the sensitivity is on the right level. An example of two priors, observations and posteriors is shown in Fig. 7.1.

7.3 Parameter estimation

Derivation of the point-estimates to the (unknown) model parameters $\boldsymbol{\theta}$ within Bayesian framework is based on either Eq. (7.1) or Eq. (7.2). There are three common choices for parameter estimate $\hat{\boldsymbol{\theta}}$: the posterior median, the posterior mean, and the maximum a posteriori (MAP) estimates. The analytical derivation of posterior median and mean estimates require the knowledge of the proper posterior

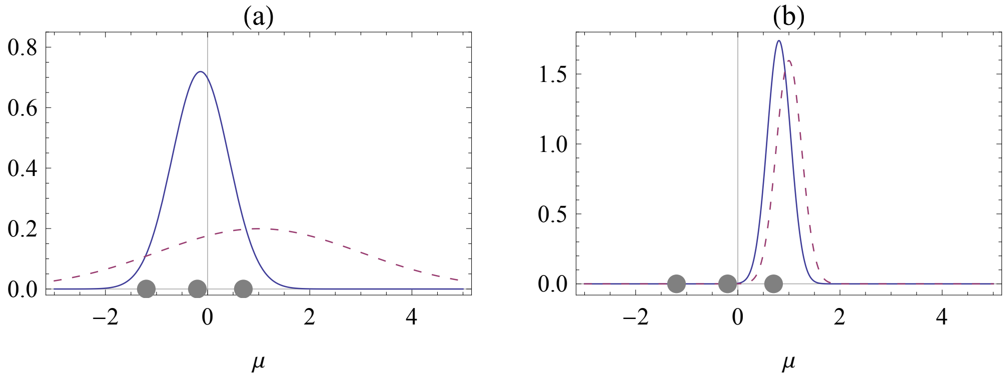

Figure 7.1: Three observations from normal distribution, normal prior (red dashed line) and posterior (blue solid line) for the parameter μ. In (a) the prior variance is large, and in (b) it is small.

distribution (Eq. (7.1)), because e.g. the posterior mean is calculated as

$$\text{The posterior mean } \hat{\boldsymbol{\theta}} = \int_{\Omega} \boldsymbol{\theta}\, \mathrm{f}(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \tag{7.5}$$

With Markov chain Monte Carlo (MCMC) methods we will see that the explicit formulation of the proper posterior distribution is not always necessary, and posterior mean or median estimates can be computed from samples.

With MAP estimate however, the proper form of posterior distribution is not needed. Maximization of Eq. (7.1) can be equally well done using only Eq. (7.2). Note the similarities with the MLE estimate which is computed in the similar manner, only without the prior distribution.

The fact that there are three equally justified and popular methods for parameter estimation in Bayesian framework is somehow typical for BI. There is a certain amount of subjectivity in every Bayesian analysis, and the best practice is to write out all the choices made, so that other researchers can reproduce the results and follow the formulations if needed.

7.3.1 Bayesian interval estimation

With frequentist ML inference the uncertainty about the ML estimate is described with confidence intervals. The similar construction in BI is the credible interval. Because in BI it is natural to speak about probability of the parameter, the credible interval is defined as

$$\int_{\theta_1}^{\theta_2} \mathrm{f}(\theta|\boldsymbol{y})\, d\theta = 1 - \alpha \tag{7.6}$$

The problem with the equation above is that it does not define the limits θ_1 and θ_2 unambiguously. There are two different extra conditions that can be used to define the interval properly. The first one is the *equal tail credible interval* where we require that the tail probabilities are the same:

$$\int_{-\infty}^{\theta_1} \mathrm{f}(\theta|\boldsymbol{y})\, d\theta = \int_{\theta_2}^{\infty} \mathrm{f}(\theta|\boldsymbol{y})\, d\theta = \frac{\alpha}{2} \tag{7.7}$$

The second possibility is that we require the posterior densities inside the credible interval to be larger than any density value outside the interval. This is called the *highest posterior density region*:

$$\begin{gathered}\theta_1 \text{ and } \theta_2 \text{ so that} \quad \mathrm{f}(\theta|\boldsymbol{y}) \geq \mathrm{f}(\theta^*|\boldsymbol{y}), \\ \text{when } \theta_1 \leq \theta \leq \theta_2 \text{ , and } \theta^* < \theta_1 \text{ or } \theta_2 < \theta^*\end{gathered} \tag{7.8}$$

For symmetric unimodal distribution these intervals will coincide.