Chapter 1

Bayesian Data Analysis

Michael S. Wheatland
School of Physics,
University of Sydney,
NSW 2006
m.wheatland@physics.usyd.edu.au

Bayesian methods provide a systematic approach to inference and data analysis in science. This chapter presents a tutorial on Bayesian analysis, with emphasis on the relationship to conventional methods. An application to solar flare prediction is then described.

Preprint of a chapter to appear in *Complex Physical, Biophysical and Econophysical Systems,* Eds. Robert L. Dewar and Frank Detering, World Scientific Publishing Company, Singapore (2010) (accepted 9 June 2009).

1.1. Scientific inference

Inference is the process of going from observed effects to underlying causes, and is the inverse process to deduction. Whereas deduction is exact, inference is imprecise, and necessarily probabilistic. Inference is the basis of science: we are always faced with observations we would like to explain in terms of underlying physical causes.

Bayesian inference is an approach to the problem based on an identity in conditional probability (Bayes's theorem). Notable Bayesians have included Pierre-Simon Laplace (who inferred the mass of Saturn from contemporary observations using Bayesian methods, and obtained a value consistent with modern estimates), the economist John Maynard Keynes, and the applied mathematician and geophysicist Harold Jeffreys. Bayesian inference has at times been controversial, because of its incorporation of subjective prior information into the process of inference. Historically the Bayesian approach was referred to as "subjective probability." In recent decades there has

been wider acceptance and application of Bayesian methods in a range of disciplines, driven by a recognition of the utility and power of the methods. Increases in computational speed and the use of Markov chain Monte Carlo methods have also played a part in this adoption.

This chapter presents an overview of Bayesian methods in Section 1.2, and then an example of their application to the problem of solar flare prediction in Section 1.3. Section 1.2.1 presents Bayes's theorem and explains its use for inference. Sections 1.2.2 and 1.2.3 describe basic approaches to parameter estimation and hypothesis testing in the Bayesian method, and Section 1.2.4 illustrates these approaches in application to a simple example: coin tossing. Section 1.2.5 gives a brief account of a relatively recent development, Markov chain Monte Carlo (MCMC) methods. Sections 1.2.6 and 1.2.7 discuss the relationship between Bayesian and classical methods of parameter estimation and hypothesis testing. Section 1.3.1 provides background on the problem of solar flare prediction, and Section 1.3.2 describes properties of flare statistics. A Bayesian approach to prediction exploiting these statistics is then presented in Section 1.3.3, and is illustrated in application to whole-Sun prediction of soft X-ray flares in Section 1.3.4.

1.2. A tutorial on Bayesian methods

1.2.1. *Bayes's theorem*

Consider two propositions, X and Y (these may be thought of as statements that are either true or false). The probability that both are true may be written

$$\begin{aligned} P(X,Y) &= P(X|Y) \times P(Y) \\ &= P(Y|X) \times P(X), \end{aligned} \tag{1.1}$$

where $P(X|Y)$ is the probability X is true, given that Y is true (a conditional probability).

The Reverend Thomas Bayes [1] applied Eq. (1.1) to inference by identifying one of the propositions with a hypothesis or model (labelled H), and the other with available data (labelled D), and writing the equation in the form

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}. \tag{1.2}$$

In many cases it is sufficient to omit the evidence term and use the state-

ment of proportionality rather than equality:

$$P(H|D) \propto P(D|H) \times P(H), \tag{1.3}$$

and then the requirement that the probabilities sum to unity over all possible hypotheses:

$$\sum_i P(H_i|D) = 1 \tag{1.4}$$

is used to determine the missing factor.

The terms in Eq. (1.2) are given names: $P(H|D)$ is called the "posterior" probability, $P(D|H)$ is the "likelihood," $P(H)$ is the "prior" probability, and $P(D)$ is sometimes called the "evidence." Eqs. (1.2) or (1.3) may be interpreted as statements of how an initial estimate of the probability of a hypothesis (the prior) is modified by new information (the likelihood), to give an updated estimate of the probability of a hypothesis (the posterior). Eq. (1.1) is a fact about conditional probability. However, in the application to inference there is some ambiguity because of the subjectivity inherent in the choice of the prior. The probability that one person assigns to a hypothesis being true, a priori, may not match that of another person.

1.2.2. *Bayesian parameter estimation*

In inference there are two basic problems: parameter estimation, i.e. deciding the best values for the parameters of a given model, and hypothesis testing or model selection, i.e. deciding between competing models. First we consider parameter estimation.

The basic approach is to express the model H in terms of model parameters, labelled $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_N]$. The functional form of the likelihood $P(D|\boldsymbol{\theta})$ must be identified in terms of these parameters, based on the model, and possibly details of how the data were obtained (the likelihood may incorporate observational uncertainties). A prior $P(\boldsymbol{\theta})$ also needs to be chosen, based on existing knowledge. Bayes's theorem is then applied in the form

$$P(\boldsymbol{\theta}|D) \propto P(D|\boldsymbol{\theta})P(\boldsymbol{\theta}) \tag{1.5}$$

to give the posterior as a function of the model parameters.

In many cases the posterior will have a single maximum, as a function of the model parameters, and the parameter values corresponding to the maximum provide "best estimates" for the parameters. The width of the

posterior in the vicinity of the maximum (how localized the maximum is) provides an estimate of the uncertainties in the best estimates.

If the interest is with only one parameter, say θ_1, then it is possible to integrate over the other parameters, to produce a univariate posterior:

$$P(\theta_1|D) = \int P(\boldsymbol{\theta}|D) d\theta_2 d\theta_3 ... d\theta_N. \tag{1.6}$$

This process of integrating over unwanted or "nuisance" parameters is called "marginalization."

In the Bayesian method, the posterior is taken to provide complete information about parameters, and methods of obtaining best estimates of parameters from the posterior are of secondary importance. Correspondingly, there are many ways to obtain best estimates. Expected values are often used. The expected value of a function $f(\theta_1)$ is

$$E\left[f(\theta_1)\right] = \int f(\theta_1) P(\theta_1|D) d\theta_1, \tag{1.7}$$

or

$$E\left[f(\theta_1)\right] = \int f(\theta_1) P(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \tag{1.8}$$

in the multi-dimensional case. Expected values of powers of θ_1 provide means and standard deviations which may be used as best estimates and uncertainties:

$$\begin{aligned} \theta_{1,\mathrm{est}} &= E\left[\theta_1\right], \\ \sigma^2_{1,\mathrm{est}} &= E\left[\theta_1^2\right] - \left(E\left[\theta_1\right]\right)^2. \end{aligned} \tag{1.9}$$

Alternatively, the location of the maximum of the posterior (the mode) is often used as a best estimate:

$$\left.\frac{d}{d\theta_1} P(\theta_1|D)\right|_{\theta_{1,\mathrm{est}}} = 0. \tag{1.10}$$

If the posterior function is approximately Gaussian, then the following formula may be used to estimate the associated uncertainty:

$$\sigma^{-2}_{1,\mathrm{est}} = -\left.\frac{d^2}{d\theta_1^2} \ln P(\theta_1|D)\right|_{\theta_{1,\mathrm{est}}}. \tag{1.11}$$

The right hand side of Eq. (1.11) is the coefficient of $\frac{1}{2}(\theta - \theta_{1,\mathrm{est}})^2$ in the Taylor expansion of $-\ln P(\theta_1|D)$ around $\theta_{1,\mathrm{est}}$. If the posterior is Gaussian this is equal to σ^{-2}, where σ is the usual width (standard deviation). The formula uses only the behaviour of the posterior function at the peak, so it

is important to check that the global behaviour is approximately Gaussian, to ensure the estimate is meaningful.

Fig. 1.1 illustrates these two approaches to obtaining best estimates and uncertainties.

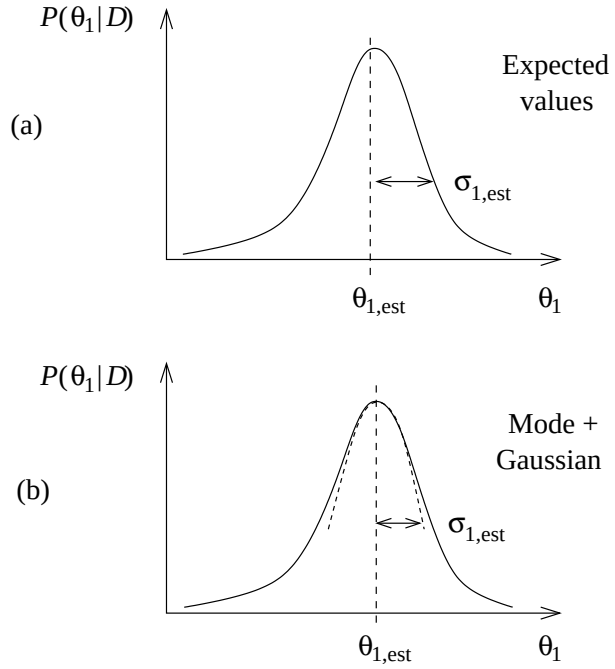

Fig. 1.1. Best estimates based on expected values [panel (a)], and on the mode and local Gaussian behaviour [panel (b)].

1.2.3. *Bayesian hypothesis testing*

Bayesian hypothesis testing involves taking ratios of Bayes's theorem. For two competing hypotheses, H_1 and H_2, we have

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \frac{P(H_1)}{P(H_2)}. \tag{1.12}$$

It should be noted that the common evidence term $P(D)$ in the two statements of Bayes's theorem has cancelled, and plays no further role. The ratio of posteriors $O_{12} = P(H_1|D)/P(H_2|D)$ is called the "odds ratio", and is equal to the ratio of the likelihoods, modulated by the ratio of the priors.

For exclusive hypotheses $[P(H_1|D) + P(H_2|D) = 1]$ it is possible to assign an absolute probability for a model, e.g. $P(H_1|D) = O_{12}/(1 + O_{12})$. More generally there may be many competing hypotheses, and it is necessary to order the relative probabilities.

1.2.4. *Is this coin fair?*

To illustrate the methods of Bayesian parameter estimation and hypothesis testing, we consider a simple example often used in text books [2]: coin tossing. Suppose that you have a coin, and that you would like to determine, on the basis of tossing the coin, whether it is fair. For example, in 10 tosses of the coin you observe two heads. Is the coin fair?

As a problem in Bayesian parameter estimation, we can consider trying to infer the "bias" θ of the coin, which we define as the probability of obtaining a head in a single toss. For a fair coin, $\theta = \frac{1}{2}$.

If r heads are observed in n tosses of the coin, the likelihood of this data D is given by the binomial distribution

$$P(D|\theta) = \frac{n!}{r!(n-r)!}\theta^r(1-\theta)^{n-r}, \tag{1.13}$$

although all we really need is the statement of proportionality:

$$P(D|\theta) \propto \theta^r(1-\theta)^{n-r}. \tag{1.14}$$

In Bayesian inference it is necessary to choose a prior, describing the state of knowledge or ignorance about parameters in the absence of data. If you were suspicious about the coin before you started tossing it, you might consider a uniform prior, assigning equal probability to all possible values of θ:

$$P(\theta) = \begin{cases} 1 \text{ if } 0 \leq \theta \leq 1, \\ 0 \text{ otherwise.} \end{cases} \tag{1.15}$$

Alternatively, if you were somewhat confident of fairness, but still wanted to admit other possibilities, you might consider a Gaussian prior, peaked about one half:

$$P(\theta) \propto \begin{cases} \exp\left[-\frac{1}{2}(\theta - \frac{1}{2})^2/\sigma^2\right] & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{1.16}$$

with a width σ chosen to reflect your suspicion about the coin.

The posterior for the problem is then given by the product of the likelihood and the chosen prior:

$$P(\theta|D) \propto \begin{cases} \theta^r(1-\theta)^{n-r}P(\theta) \text{ if } 0 \leq \theta \leq 1, \\ 0 \qquad\qquad\qquad\quad \text{otherwise,} \end{cases} \tag{1.17}$$

and the normalization condition $\int_0^1 P(H|D)dH = 1$ is used to determine the constant of proportionality.

Fig. 1.2 illustrates the evaluation of this posterior, for both choices of the prior. Each panel shows a probability distribution function (PDF) for θ. Panel (a) illustrates the two priors, with the uniform prior shown by the solid curve and the Gaussian prior (with $\theta = 0.2$) shown by the dashed curve. Panel (b) shows the corresponding posterior distributions for the observation of two heads in ten tosses of the coin. Panel (c) shows the corresponding posterior distributions for the observation of 28 heads in 100 tosses of the coin.

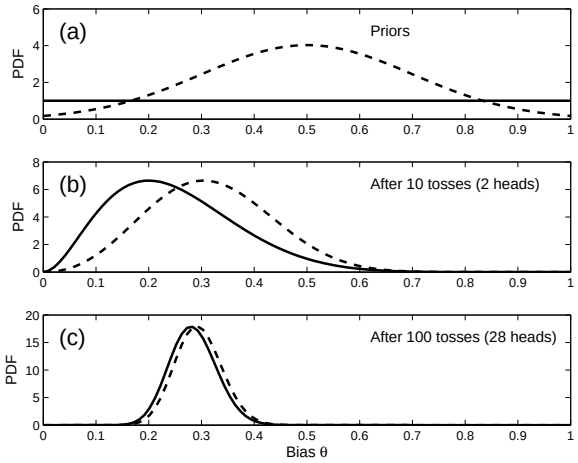

Fig. 1.2. Bayesian inference applied to coin tossing. Panel (a) shows two possible choices for the prior distribution of the probability θ of a coin landing heads: a uniform prior (solid), and a Gaussian prior (dashed). Panel (b) shows the corresponding posterior distributions based on the observation of two heads in ten tosses. Panel (c) shows the corresponding posterior distributions based on 28 heads in 100 tosses.

After 10 tosses of the coin [panel (b) in Fig. 1.2], the posterior distributions are quite broad. The value $\theta = 0$ is ruled out, because heads have been observed. Values of θ above about 0.7 are unlikely, but only $\theta = 1$

is strictly impossible (since tails have been observed). The two choices of prior lead to somewhat different posterior distributions. On the basis of these results, it is hard to make a very definitive estimate of θ, and the prior plays an important role.

After 100 tosses of the coin [panel (c) in Fig. 1.2], the posterior distributions are much narrower. Values of θ less than about 0.1 and larger than about 0.5 are very unlikely, based on the data. The two choices of prior lead to quite similar posterior distributions. On the basis of these result, more definitive estimates of θ may be made, and the role of the prior is less important.

To illustrate more quantitative parameter estimation, we can consider the case of the uniform prior, which is straightforward to evaluate analytically. The normalisation of the posterior is achieved using the Eulerian integral

$$\int_0^1 \theta^r (1-\theta)^{n-r} d\theta = \frac{r!(n-r)!}{(n+1)!}, \tag{1.18}$$

so that

$$P(\theta|D) = \frac{(n+1)!}{r!(n-r)!}\theta^r (1-\theta)^{n-r}. \tag{1.19}$$

Evaluating Eq. (1.9) for this distribution gives

$$\theta_{\rm est} = \frac{r+1}{n+2}, \qquad \sigma^2_{\rm est} = \frac{\theta_{\rm est}(1-\theta_{\rm est})}{n+3}. \tag{1.20}$$

(The result for $\theta_{\rm est}$ is known as Laplace's rule of succession, and was famously used by Laplace to estimate the probability that the Sun will rise tomorrow [3].) For the case of two heads in 10 tosses, we have $\theta_{\rm est} \approx 0.27$ and $\sigma_{\rm est} \approx 0.12$, which is suggestive of a departure from fairness but not definitive. For 28 heads in 100 tosses we have $\theta_{\rm est} \approx 0.28$ and $\sigma_{\rm est} \approx 0.045$, which is becoming quite definitive. Alternatively, we can use Eqs. (1.10) and (1.11), leading to

$$\theta_{\rm est} = \frac{r}{n}, \qquad \sigma^2_{\rm est} = \frac{\theta_{\rm est}(1-\theta_{\rm est})}{n}. \tag{1.21}$$

For $r = 2$ and $n = 10$ we have $\theta_{\rm est} = 0.2$ and $\sigma_{\rm est} \approx 0.13$, and for $r = 28$ and $n = 100$ we have $\theta_{\rm est} = 0.28$ and $\sigma_{\rm est} \approx 0.045$.

As an example of a hypothesis test, we consider the question of whether, on the basis of the data, the coin is more likely to be heads biased ($\theta > \frac{1}{2}$),

or tails-biased ($\theta < \frac{1}{2}$). For a uniform prior, the odds ratio of the two models may be evaluated analytically:

$$\begin{aligned} O_{ht}(r,n) &= \frac{\int_{\frac{1}{2}}^{1} \theta^r (1-\theta)^{n-r} d\theta}{\int_0^{\frac{1}{2}} \theta^r (1-\theta)^{n-r} d\theta} \\ &= \frac{I_{\frac{1}{2}}(n-r+1, r+1)}{1 - I_{\frac{1}{2}}(n-r+1, r+1)} \end{aligned} \tag{1.22}$$

where $I_x(a,b)$ is the incomplete Beta function. [This example does not correspond exactly to Eq. (1.12) because here we have integrated each posterior over the relevant values of the model parameter θ.] Evaluating this expression for the examples of interest gives $O_{ht}(2,10) = 67/1981 \approx 3.4 \times 10^{-2}$, and $O_{ht}(28,100) \approx 4.3 \times 10^{-6}$. For the case of 10 tosses, the coin is more likely to be tails-biased, although the result is not definitive, but for 100 tosses the tails bias is very strongly favoured.

1.2.5. *Markov chain Monte Carlo (MCMC)*

Normalisation and calculation of expected values involves evaluating integrals, for example of the form of Eq. (1.7), which may be multi-dimensional. Until recently, this presented a practical problem for Bayesian inference. However, "Markov chain Monte Carlo" (MCMC) methods now provide a powerful, general solution to the problem. Here we present only the basic idea, which is particularly simple. (For details, see e.g. Ref. [4].)

If a sample $\{\theta_{1i}, i = 1, 2, ..., n\}$ of random variables from a probability distribution $P(\theta_1|D)$ is available, then an estimate of an expected value may be constructed via a sum:

$$\begin{aligned} E\left[f(\theta_1)\right] &= \int f(\theta_1) P(\theta_1|D) d\theta_1, \\ &\approx \frac{1}{n} \sum_i f(\theta_{1i}). \end{aligned} \tag{1.23}$$

Markov chain Monte Carlo methods provide ways to generate appropriate sets $\{\theta_{11}, \theta_{12}, ...\}$, using only uniformly-distributed random variables, which are simple to generate (approximately) on a computer, and evaluations of the function $P(\theta_1|D)$. The methods produce Markov chains (sequences of random numbers, such that each number depends only on the previous number) with the property that, after an initial "burn-in" period of non-stationarity, the Markov Chain becomes stationary, and then approximates

a sequence of samples from $P(\theta_1|D)$. A number of different MCMC algorithms are commonly used, including the Metropolis, Metropolis-Hastings, and Gibbs sampler methods.

1.2.6. *Relationship to maximum likelihood and least squares*

"Maximum likelihood" and "least squares" are methods commonly used for parameter estimation, which are closely related to the Bayesian approach. We briefly discuss the relationship, and the approximations and assumptions being made these methods.

Consider a model involving parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_N]$, and a set of data $\mathbf{D} = [D_1, D_2, ..., D_M]$. Bayes's theorem may be stated

$$P(\boldsymbol{\theta}|\mathbf{D}) \propto P(\mathbf{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}). \tag{1.24}$$

Assuming a uniform prior gives

$$P(\boldsymbol{\theta}|\mathbf{D}) \propto P(\mathbf{D}|\boldsymbol{\theta}), \tag{1.25}$$

i.e. the posterior is proportional to the likelihood. The "maximum likelihood estimate" $\boldsymbol{\theta}_{\rm ML} = [\theta_{\rm ML}1, \theta_{\rm ML}2, ..., \theta_{\rm ML}N]$ is the set of model parameters which maximizes the likelihood, i.e. satisfies

$$\left.\frac{\partial}{\partial\theta_i}P(\mathbf{D}|\boldsymbol{\theta})\right|_{\theta_{{\rm ML}i}} = 0, \tag{1.26}$$

for $i = 1, 2, ..., N$. In conventional statistical inference, the only justification for this estimate is that it makes the observed data most probable, or most likely [5]. However, given Eq. (1.25), this estimate also maximizes the posterior, i.e. makes the model most probable. Hence we see that the maximum likelihood estimate is the Bayesian modal estimate, assuming a uniform prior.

Assuming that the data points are independent, we have

$$P(\mathbf{D}|\boldsymbol{\theta}) = P(D_1|\boldsymbol{\theta})P(D_2|\boldsymbol{\theta})\ldots P(D_M|\boldsymbol{\theta}). \tag{1.27}$$

If the model gives data values $\mathbf{F}(\boldsymbol{\theta}) = [F_1(\boldsymbol{\theta}), F_2(\boldsymbol{\theta}), ..., F_M(\boldsymbol{\theta})]$ in the absence of observational errors (uncertainties), and the errors are assumed to be Gaussian distributed, then the likelihood of each datum is

$$P(D_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{[F_i(\boldsymbol{\theta}) - D_i]^2}{2\sigma_i^2}\right\} \tag{1.28}$$

for $i = 1, 2, ..., M$, where the $\sigma_i = \sigma_i[F_i(\boldsymbol{\theta})]$ are uncertainties which depend on the model data values in a specified way[a] Combining Eqs. (1.27) and (1.28), the overall likelihood is

$$P(\mathbf{D}|\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}\chi^2(\boldsymbol{\theta})\right], \tag{1.29}$$

where

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{M} \frac{[F_i(\boldsymbol{\theta}) - D_i]^2}{\sigma_i^2}. \tag{1.30}$$

The quantity $\chi^2 = \chi^2(\boldsymbol{\theta})$ is usually called "chi-square," or the "chi-square statistic." The log-likelihood is

$$\ln P(\mathbf{D}|\boldsymbol{\theta}) = \text{const} - \frac{1}{2}\chi^2(\boldsymbol{\theta}), \tag{1.31}$$

and clearly the likelihood/log-likelihood is a maximum when chi-square is a minimum. The estimate for the model parameters obtained by minimizing chi-square, which we label $\boldsymbol{\theta}_{\rm LS}$, is usually called the "least squares" estimate.

From this derivation, we see that the least squares estimate is also a Bayesian estimate, subject to additional of assumptions and approximations. In principle these assumptions may be relaxed in the Bayesian approach. For example, it is possible to incorporate errors other than Gaussian errors, which is sometimes appropriate. Also, a non-uniform prior may be introduced to reflect prior knowledge about the model parameters. For example, if one parameter represents energy, than the prior may be used to enforce the requirement of non-negativity of this parameter. More generally, the prior may be used to "bias" certain areas of the parameter space, if it is known a priori that certain values are more likely to be correct. This process has no counterpart in classical methods.

Bayesian methods are also distinct from classical methods in other specific ways. The Bayesian approach provides a posterior distribution, rather than the limited information afforded by best estimates and uncertainties. This may be of particular use if the posterior distribution has an unusual shape (for example is multi-modal, or otherwise departs significantly from a Gaussian). The posterior distribution contains the totality of information available from inference, and this may be scrutinized in different ways.

[a]The Gaussian or "normal" distribution is often appropriate to describe observational uncertainties. Some insight into the almost ubiquitous success of the Gaussian to describe errors is provided by the central limit theorem [5], which states (roughly) that the sum of a large number of independent random variables from a variety of distributions is normally distributed. For a more detailed explanation, see Ref. [6].

Bayesian methods also place a fundamentally different emphasis on the roles of data and of models. Classical methods work with the likelihood, which presupposes a model, and assesses the probability of the data given the model. The model is essentially treated as being perfect, and the data imperfect. From the Bayesian perspective, the roles are reversed. The posterior assesses the probability of the model given the data, so the data is presupposed, or perfect, and the model imperfect. For many scientists this may appear to be a more natural perspective: science involves the construction and refinement of models based on available observations.

1.2.7. *Classical hypothesis testing*

Classical hypothesis testing is quite different to the Bayesian approach presented in Section 1.2.3. The classical method involves the choice of a "statistic,"and here we consider the use of chi-square.

A large value of $\chi^2_{\rm LS} = \chi^2(\boldsymbol{\theta}_{\rm LS})$ (where $\boldsymbol{\theta}_{\rm LS}$ is the least squares estimate, obtained as explained in Section 1.2.6), may be an indication that something is wrong. One possibility is that the model is incorrect. The "chi-square test" involves calculating $P_d(\chi^2 > \chi^2_{\rm LS})$, the probability of obtaining a larger value of χ^2 than $\chi^2_{\rm LS}$, for the given data, assuming that the model is correct. The quantity $P_d(\chi^2 > \chi^2_{\rm LS})$ is called the significance, and depends on $d = M - N$, the "number of degrees of freedom." It is straightforward to calculate this quantity based on the likelihood defined by Eqs. (1.27)-(1.30) [7]. The calculation evaluates the probability of getting data that departs further from the (fixed) model than the data that was observed. The classical approach to hypothesis testing then involves "rejecting" the model if the significance is too small, say less than 1%.

A variety of criticisms of this procedure have been raised. First, it is not possible to accept a model, only to reject it. Given a suitably aberrant set of data, any model will be rejected. As Harold Jeffreys stated, "There has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected all laws and left us with no law." [8] A related criticism is that the method does not consider alternative hypotheses. Finally, there is a degree of arbitrariness in the choice of the statistic, and also in the choice of a significance level for rejection. The Bayesian method explicitly deals with these problems. If a hypothesis is generally accepted, then the prior should reflect this, and a test based on a single set of aberrant data will not lead to the model being rejected. The Bayesian method forces consideration of competing hypotheses, and does

not involve the arbitrary choice of a statistic, or of a significance level.

1.3. An application to solar flare prediction

1.3.1. *Background*

Solar flares are magnetic explosions in the ionised outer atmosphere of the Sun, the solar corona. Flares occur in and around sunspots, where intense magnetic fields penetrate the visible surface of the Sun, and thread the overlying coronal plasma. During a flare some of the energy stored in the magnetic field is released and appears in accelerated particles, radiation, heating, and bulk motion. The flare mechanism is accepted to be magnetic reconnection, a process involving a change in connectivity of magnetic field lines, but many aspects of the process remain poorly understood. Flares occur suddenly, and there are no (known) infallible indicators that a flare is about to occur. Hence flare prediction is probabilistic.

Large flares strongly influence our local "space weather." They can lead, for example, to enhanced populations of energetic particles in the Earth's magnetosphere (the region magnetically connected to the Earth), and these particles can damage satellite electronics, and pose radiation risks to astronauts and to passengers on polar aircraft flights. The space weather effects of large flares motivate a need for accurate solar flare prediction.

A variety of properties of active regions are correlated with flare occurrence. For example, certain sunspot classifications [9], qualitative measures of magnetic complexity [10], and moments of quantitative photospheric magnetic field maps [11, 12] provide flare predictors, of varying reliability. Operational flare forecasters refer to the tendency of a region which has produced large flares in the past to produce large flares in the future as "persistence," and this provides one of the most reliable predictors for large flare occurrence in 24-hour forecasts [13]. The US National Oceanic and Atmospheric Administration (NOAA) uses an "expert system" based on sunspot classification and other properties of active regions to assign probabilities for the occurrence of large flares [9].

Flares are commonly classified by their peak flux at X-ray wavelengths, in the 1-8 Å band measured by the Geostationary Observational Environmental (GOES) satellites. Moderate size flares correspond to "M-class" events, with a peak flux in the range $\geq 10^{-5}\,\mathrm{W\,m^{-2}}$ to $\geq 10^{-4}\,\mathrm{W\,m^{-2}}$. Large flares correspond to "X-class" events, with peak flux above $\geq 10^{-4}\,\mathrm{W\,m^{-2}}$. The NOAA predictions assign corresponding probabilities

$\epsilon_{\rm M}$ and $\epsilon_{\rm X}$ for the occurrence of at least one event with peak flux above these levels within 24 hours.

Existing methods of flare prediction are not very accurate. One measure of success of probabilistic event forecasts is provided by the "skill score," defined as

$$\mathrm{SS}(f,x) = 1 - \frac{\mathrm{MSE}(f,x)}{\mathrm{MSE}(\langle x\rangle, x)}, \tag{1.32}$$

where f denotes the forecast value, x denotes the observation (a one or a zero, according to whether an event did or did not occur, respectively), $\langle ...\rangle$ denotes an average over the forecasts, and

$$\mathrm{MSE}(f,x) = \langle (f-x)^2\rangle \tag{1.33}$$

denotes the mean-square error. The skill score quantifies the improvement of the forecasts over a prediction of the average in every case. The maximum of the skill score is one, representing perfect prediction, and negative values of the skill score indicate predictions worse than forecasting the average. The NOAA published statistics describing the success of its forecasts for 1986–2006[b] The skill score for one-day forecasting of X-class flares is positive for only 7 of the 21 years.

1.3.2. *Flare statistics*

Flare occurrence follows a power-law frequency-size distribution, where "size" denotes some measure of the flare magnitude, for example the peak flux in X-rays measured by GOES. In other words, the number of events per unit time and per unit size S, denoted $N(S)$, obeys

$$N(S) = \lambda_1(\gamma-1)S_1^{\gamma-1}S^{-\gamma}, \tag{1.34}$$

where λ_1 is the total rate of events above size S_1 oberved, and $\gamma \approx 1.5-2$ is a constant, which depends on the specific choice of S. Although the distribution is typically constructed based on all flaring active regions present on the Sun over some period of time, it also appears to hold in individual active regions [14]. The appearance of this power law in flare occurrence motivated the avalanche model for flares, in which the energy release mechanism consists of a sequence of elementary events which trigger one another, and in which the system is in a self-organised critical state [15, 16].

[b]See http://www.swpc.noaa.gov/forecast_verification/.

Flare occurrence in time may be modelled as a Poisson process [17, 18]. For intervals in which the mean rate of flaring λ does not vary greatly, the distribution of waiting times τ is then

$$P(\tau) = \lambda \exp(-\lambda\tau). \tag{1.35}$$

Over longer time scales, the rate will vary with time, and the distribution is more complex.

Fig. 1.3 illustrates these properties of flare statistics. Panel (a) shows a schematic of a sequence of events, as size versus time, and also illustrates a waiting time τ. Panel (b) shows the power-law frequency-size distribution, and panel (c) shows the Poisson waiting-time distribution.

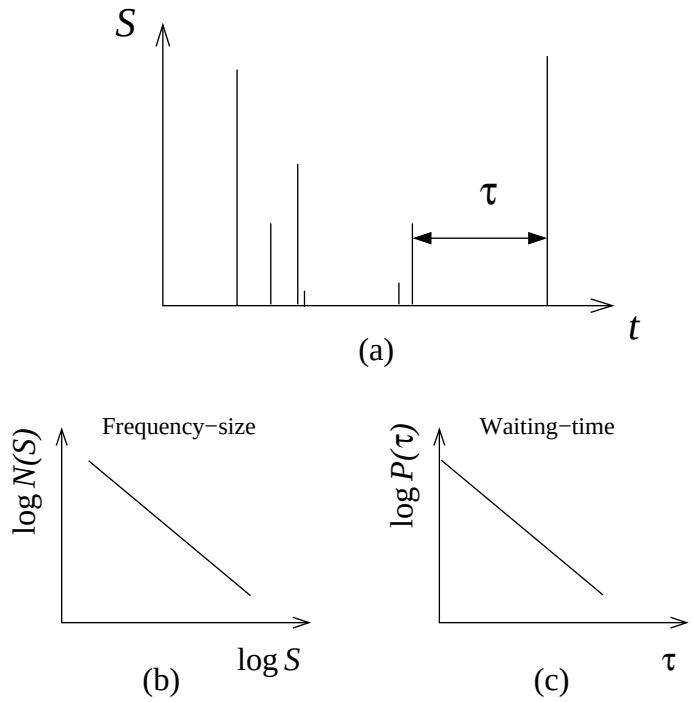

Fig. 1.3. Schematic illustration of flare statistics. Panel (a): flare events, showing size S versus time, and indicating a waiting time τ. Panel (b): frequency-size distribution. Panel (c): waiting-time distribution.

1.3.3. *Event statistics method of prediction*

Given the relative success of persistence as a flare predictor, and the simple statistical rules describing flare occurrence, it is worthwhile to consider methods of prediction relying on flare statistics alone. Refs. [19] and [20] develop such an approach, using the Bayesian method.

The basic idea is as follows. If S_1 is the size of a "small" event (chosen such that small events are well observed), and S_2 is the size of a "big" event (which you would like to predict), then the power-law frequency-size distribution Eq. (1.34) implies that the rates λ_1 and λ_2 of events above the two sizes are related according to

$$\lambda_2 = \lambda_1 \left(S_1/S_2\right)^{\gamma-1}. \tag{1.36}$$

Eq. (1.36) allows estimation of the rate of big events even if none have been observed. Given this estimate, the probability of at least one big event in a time T_P is

$$\begin{aligned}\epsilon &= 1 - \exp(-\lambda_2 T_P) \\ &= 1 - \exp\left[-\lambda_1 \left(S_1/S_2\right)^{\gamma-1} T_P\right],\end{aligned} \tag{1.37}$$

using Eq. (1.35). If M events are involved in the estimation of the rate λ_1, then it follows that $\sigma_\epsilon/\epsilon \approx M^{-1/2}$ [19]. Hence the prediction becomes accurate if many small events are observed.

The Bayesian aspect of the method concerns the estimation of γ and λ_1 from the observed data. Specifically, if the data D consists of events $s_1, s_2, ..., s_M$ at times $t_1 < t_2 < ... < t_M$, then the problem is to calculate posterior distributions $P_\gamma(\gamma|D)$ and $P_1(\lambda_1|D)$. Given these, the posterior distribution for λ_2 is

$$\begin{aligned}P_2(\lambda_2|D) = \int_1^\infty d\gamma \int_0^\infty d\lambda_1 P_1(\lambda_1|D) P_\gamma(\gamma|D) \\ \times\, \delta\left[\lambda_2 - \lambda_1 (S_1/S_2)^{\gamma-1}\right],\end{aligned} \tag{1.38}$$

using Eq. (1.36). Finally, the posterior distribution for ϵ is obtained using

$$P_\epsilon(\epsilon|D) = P_2\left[\lambda_2(\epsilon)|D\right] \left|\frac{d\lambda_2}{d\epsilon}\right|, \tag{1.39}$$

where $\lambda_2(\epsilon) = -\ln(1-\epsilon)/T_P$, from Eq. (1.37).

The inference of the power-law index γ follows from Eq. (1.34), which implies the likelihood [21]:

$$P(D|\gamma) \propto \prod_{i=1}^{M} (\gamma - 1)(s_i/S_1)^{-\gamma}. \tag{1.40}$$

A uniform prior is used. If $M \gg 1$, the posterior/likelihood is sharply peaked in the vicinity of the modal/maximum likelihood estimate

$$\gamma_{\rm ML} = \frac{M}{\ln \pi} + 1, \quad \text{where} \quad \pi = \prod_{i=1}^{M} \frac{s_i}{S_1}. \tag{1.41}$$

In this case a suitable approximation to the posterior in Eq. (1.38) is provided by $P_\gamma(\gamma|D) = \delta(\gamma - \gamma_{\rm ML})$.

The inference of the rate λ_1 of small events is complicated by the time variation of the rate. The procedure used is to estimate the rate at the time the prediction is made using the "Bayesian blocks" procedure from Ref. [22]. This procedure is a Bayesian change-point algorithm for decomposing a point process into a piecewise-constant Poisson process by iterative comparison of one- versus two-rate Poisson models.

Fig. 1.4 illustrates the procedure. Panel (a) shows a sequence of data, consisting of point events on a time line, during an observation interval T. The prediction interval T_P is also shown. The Bayesian blocks procedure compares the relative probability of one- and two-rate models for the observation interval T, for all choices of change point corresponding to an event time. If a two-rate model is more probable, then the data in each of the two chosen intervals is used for comparison of one- and two-rate models, and these intervals may be further sub-divided. An interval for which the one-rate model is more probable is a Bayesian block. The procedure continues iteratively in this way, until a sequence of Bayesian blocks is decided on, as shown in panel (b). The data D' in the last block, consisting of M' events in time T', then supplies a likelihood for the current data given the rate λ_1:

$$P_1(D'|\lambda_1) \propto \lambda_1^{M'} \mathrm{e}^{-\lambda_1 T'}, \tag{1.42}$$

based on the assumption of Poisson occurrence. The prior may be taken to be uniform [19], or a prior may be constructed based on the rates in the other blocks [20].

1.3.4. *Whole-Sun prediction of GOES flares*

To illustrate the method, we consider whole-Sun prediction of GOES soft X-ray flares, as described in detail in Ref. [20].

The largest soft X-ray flare of the modern era occurred on 4 November 2003, and saturated the GOES detectors at X28 (a peak flux in the 1-8 Å GOES band of $2.8 \times 10^{-3}\,\mathrm{W\,m^{-2}}$), although it was later estimated to

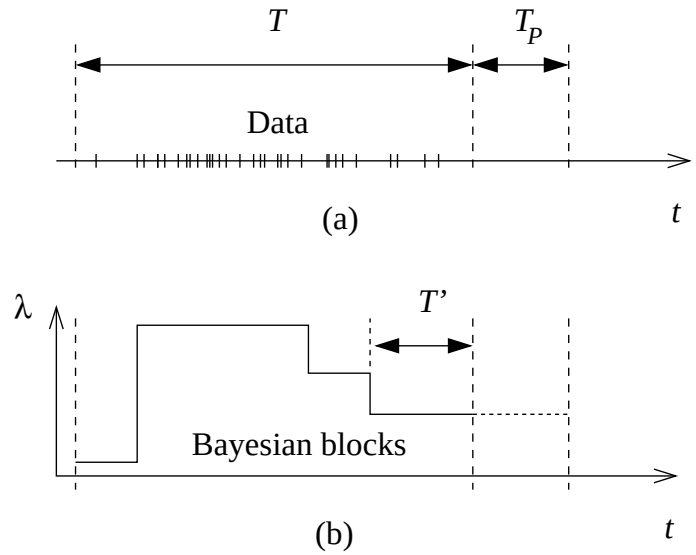

Fig. 1.4. Schematic illustration of Bayesian blocks determination of current rate. Panel (a): data, comnsisting of point events in time line during an observation interval T. The prediction interval T_P is also shown. Panel (b): Bayesian blocks decomposition of the rate λ, and identification of the most recent interval T' when the rate is approximately constant.

be as large as X45 [23]. It is interesting to consider applying the method for that day, following [20].

The data D consists of one year of events prior to the day from the whole Sun, above peak flux $S_1 = 4 \times 10^{-6}\,\mathrm{W\,m^{-2}}$ (corresponding to a GOES C4 event). This gives 480 events. Probabilities ϵ_{MX} and ϵ_{X}, for the occurrence of at least one event in the range M to X, and at least one X-class event, respectively, were inferred for the 24 hours of 4 November 2003.

Fig. 1.5 illustrates the Bayesian blocks procedure in application to the data. Panel (a) shows the 480 events plotted as peak flux versus time. Panel (b) shows the Bayesian blocks decomposition: there are 13 blocks, and the last block has a duration of $T' = 15$ days and contains $M' = 104$ events.

Fig. 1.6 shows the posteriors for the predictions. The solid curve corresponds to ϵ_{MX}, and the dashed curve corresponds to ϵ_{X}. The best estimates (using expected values) are shown by short vertical lines at the bottom, and are $\epsilon_{\mathrm{MX}} \approx 0.73 \pm 0.03$, and $\epsilon_{\mathrm{X}} \approx 0.19 \pm 0.02$. These values are quite high, reflecting the recent high rate of flaring on the Sun. However, the estimates

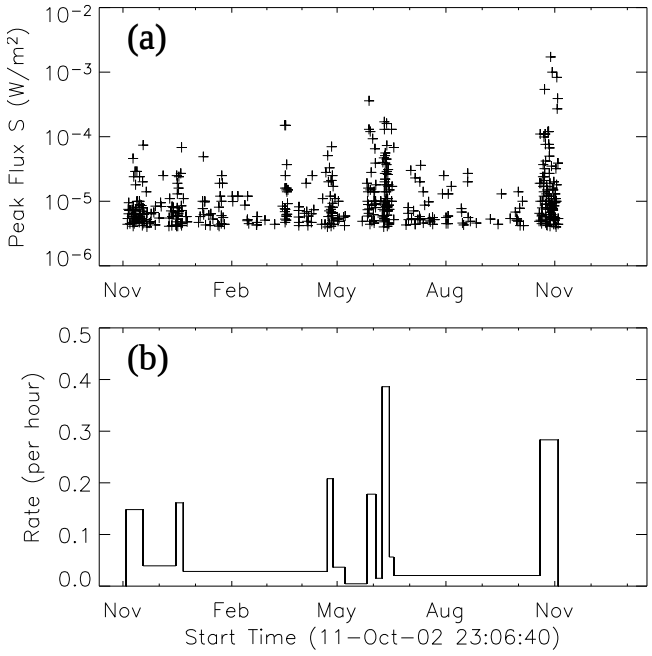

Fig. 1.5. Bayesian blocks applied to one year of GOES events prior to 4 November 2003.

also highlight the limitations of probabilistic forecasting: the prediction for X-class events is only 20%, yet the largest flare of the last three decades is about to occur. (Incidentally, three M-class events were also recorded on 4 November 2003.)

The whole-Sun implementation of the method was tested on the GOES record for 1976-2003 [20]. For each day a prediction was made based on one year of data prior to the day, following the procedure outlined for 4 November 2003. Comparison was then made with whether or not events occurred on each day, and the success of the method was evaluated statistically. Table 1.1 provides statistics for the predictions for 1987-2003, for which years NOAA predictions are also available. The mean-square errors [see Eq. (1.33)], and the skill scores [see Eq. (1.32)] are listed. The event statistics method achieves very comparable results to the NOAA method, and even performs somewhat better, in terms of the skill score, for prediction of X-class flares.

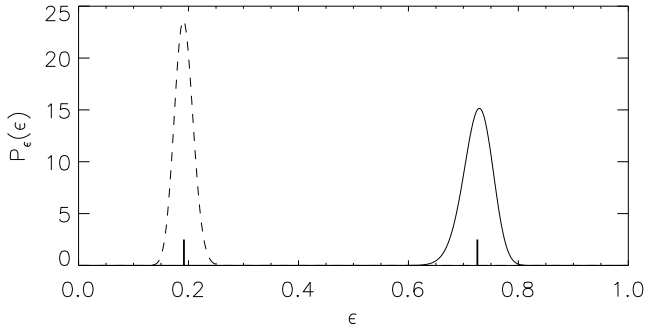

Fig. 1.6. Posterior distributions for predictions for 4 November 2003. The solid curve is the posterior for $\epsilon_{\rm MX}$, the probability of getting at least one flare in the range M to X, and the dashed curve is the posterior for $\epsilon_{\rm X}$, the probability of getting at least one X flare. The short vertical lines at the bottom indicate best estimates, using expected values.

Table 1.1. Comparison with NOAA predictions, for 1987-2003.

	Event statistics		NOAA	
	M-X	X	M-X	X
MSE(f,x)	0.143	0.031	0.139	0.032
SS(f,x)	0.258	0.078	0.262	-0.006

1.4. Summary

This chapter presents a tutorial on Bayesian methods, and an example of application to solar flare prediction. The emphasis has been on the basic principles, and on the relationship to conventional methods. For more details on Bayesian approaches, I recommend Refs. [2, 4, 6, 24].

References

[1] T. Bayes, An essay towards solving a problem in the doctrine of chances, *Phil. Trans. Roy. Soc.* **53**, 370–418, (1763).
[2] D. Sivia, *Data Analysis, A Bayesian Tutorial.* (Oxford University Press, Oxford, 1996).
[3] P.-S. Laplace, *A Philosophical Essay on Probabilities (1819; Translated from the 6th French edition by Frederick Wilson Truscott and Frederick Lincoln Emory).* (Dover Publications, New York, 1951).

[4] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. (Chapman and Hall/CRC, Boca Raton, 1996).
[5] J. Rice, *Mathematical Statistics and Data Analysis*. (Wadsworth and Brooks, Pacific Grove, California, 1988).
[6] E. Jaynes, *Probability Theory, The Logic of Science*. (Cambridge University Press, Cambridge, 2003).
[7] J. Mathews and R. Walker, *Mathematical Methods of Physics*. (Addison-Wesley, Redwood City, California, 1970), 2nd edition.
[8] H. Jeffreys, *Theory of Probability*. (Clarendon Press, Oxford, 1961), 3rd edition.
[9] P. McIntosh, The classification of sunspot groups, *Solar Phys.* **125**, 251–267, (1990).
[10] I. Sammis, F. Tang, and H. Zirin, The dependence of large flare occurrence on the magnetic structure of sunspots, *Astrophys. J.* **540**, 583–587, (2000).
[11] K. Leka and G. Barnes, Photospheric magnetic field properties of flaring versus flare-quiet active regions. IV. A statistically significant sample, *Astrophys. J.* **656**, 1173–1186, (2007).
[12] G. Barnes, K. Leka, E. Schumer, and D. Della-Rose, Probabilistic forecasting of solar flares from vector magnetogram data, *Space Weather*. **5**, S09002, (2007).
[13] D. Neidig, P. Wiborg, and P. Seagraves. The role of persistence in the 24-hour flare forecast. In eds. R. Thompson, D. Cole, P. Wilkinson, M. Shea, D. Smart, and G. Heckman, *Solar-Terrestrial Predictions: Workshop Proceedings, Leura, Australia, 16-20 October, 1989*, pp. 541–545, Boulder, Colorado, (1990).
[14] M. Wheatland, Flare frequency-size distributions for individual active regions, *Astrophys. J.* **532**, 1209–1214, (2000).
[15] E. Lu and R. Hamilton, Avalanches and the distribution of solar flares, *Astrophys. J.* **380**, L89–L92, (1991).
[16] P. P. Charbonneau, S. McIntosh, H.-L. Liu, and T. Bogdan, Avalanche models for solar flares (Invited review), *Solar Phys.* **203**, 321–353, (2001).
[17] M. Wheatland, Rates of flaring in individual active regions, *Solar Phys.* **203**, 87–106, (2001).
[18] M. Wheatland and Y. Litvinenko, Understanding solar flare waiting-time distributions, *Solar Phys.* **211**, 255–274, (2002).
[19] M. Wheatland, A Bayesian approach to solar flare prediction, *Astrophys. J.* **609**, 1134–1139, (2004).
[20] M. Wheatland, A statistical solar flare forecast method, *Space Weather*. **3**, S07003, (2005).
[21] T. Bai, Variability of the occurrence frequency of solar flares as a function of peak hard x-ray rate, *Astrophys. J.* **404**, 805–809, (1993).
[22] J. Scargle, Studies in astronomical time series analysis. V. Bayesian blocks, a new method to analyze structure in photon counting data, *Astrophys. J.* **504**, 405–418, (1998).
[23] N. Thomson, C. Rodger, and R. Dowden, Ionosphere gives size of greatest solar flare, *Geophys. Res. Lett.* **31**, L06803, (2004).

[24] P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences, A Comparative Approach with Mathematica Support*. (Cambridge University Press, Cambridge, 2005).