Data Analysis Using Bayesian Statistics

Jens Bratten Due

Spring 2020

Introduction

This is a report based on the book *Data Analysis - a Bayesian tutorial* by D.S. Sivia with J. Skilling. In it, the author describes a wish for a set of underlying basic principles to statistics, something he felt was missing from his undergraduate lectures on the subject. He suggests a different approach to data analysis, based entirely on probability theory. This report will cover the main takeaways from the book, and help me reach a greater understanding for the topic of Bayesian statistics.

1 The basics

First we are introduced to probability theory and its basic algebra which includes the sum rule

$$P(X|I) + P(\overline{X}|I) = 1 \tag{1}$$

and the product rule

$$P(X, Y|I) = P(X|Y, I) \times P(Y|I). \tag{2}$$

Here P stands for probability, the bar "|" means "given" and $\overline{X}$ means "not X". Lastly, we have the symbol I, meaning all relevant background information. The sum rule can then be stated as "the probability of X being true plus the probability of X not being true, both given all relevant background, equals 1".

Using the product rule, and the fact that $P(X, Y|I) = P(Y, X|I)$ we get the following.

$$P(X|Y, I) \times P(Y|I) = P(Y|X, I) \times P(X|I)$$

Rearranging this leads to *Bayes' theorem*

$$P(X|Y, I) = \frac{P(Y|X, I) \times P(X|I)}{P(Y|I)} \tag{3}$$

To get a clearer picture of the significance of Bayes' theorem, we can replace X and Y with *hypothesis* and *data*. $P(hypothesis|data, I)$ is then given the formal name *posterior probability*, $P(data|hypothesis, I)$ is called the *likelihood function* and $P(hypothesis|I)$ is called the *prior probability*, representing our knowledge about the truth of the hypothesis before any data has been analysed. The term in the denominator, $P(data|I)$, often called the *evidence*, is in many cases not shown, thereby replacing the equality sign with a proportionality.

$$P(hypothesis|data, I) \propto P(data|hypothesis, I) \times P(hypothesis|I) \tag{4}$$

This is due to it often being absorbed by a normalization constant. In summary, Bayes' theorem describes a learning process, showing how a probability is augmented by the introduction of data.

Another useful result from using the sum and product rule is the *marginalization* equation

$$P(X|I) = \int_{-\infty}^{\infty} P(X, Y|I)dY \tag{5}$$

with a normalization condition

$$\int_{-\infty}^{\infty} P(Y|X, I)dY = 1. \tag{6}$$

The marginalization equation gives us the ability to integrate out so-called nuisance parameters, values of no interest to a specific problem, such as background signals and measurement byproducts.

These rules of probability are widely applicable and provide a strong foundation for tackling data analysis problems. We are encouraged to share Laplace's view that "Probability theory is nothing but common sense reduced to calculation".

2 Parameter estimation I

This chapter focuses on the act of estimating a single parameter using Bayes' theorem, such as the mass of a planet, or the charge of the electron. We will firstly go through the example of deducting the fairness of a coin. This can be represented by the *bias-weighting* H. $H = 1/2$ will mean the coin is fair, while $H = 1$ and $H = 0$ means the coin is showing only heads or tails every flip. This value is continuous on the range $[0, 1]$, and $P(H|\{data\}, I)$ describes how much we believe H to be true. For a range of H-values, $P(H|\{data\}, I)$ is a *probability density function* (pdf). To find this, we use Bayes' theorem.

$$P(H|\{data\}, I) \propto P(\{data\}|H, I) \times P(H|I) \tag{7}$$

We can, if needed, find the normalization constant using equation (6). To express ultimate ignorance, we can assign a flat pdf for the prior.

$$P(H|I) = \begin{cases} 1 & 0 \leq H \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

meaning we assume every value of H to be equally probable. Assuming each flip is an independent event, the likelihood function takes the form of the binomial distribution.

$$P(\{data\}|H, I) \propto H^R(1-H)^{N-R} \tag{9}$$

where R is the number of heads and N is the number of flips.

Plugging (8) and (9) into Bayes' theorem results in the posterior probability, the shape of which varies significantly for the first few data points. When the number of data increases however, the pdf becomes sharper and converges to the most likely value. The choice of prior becomes mostly irrelevant when we have a large of number of data, as the majority of propositions will converge to the same solution, but the speed of convergence may vary. A very confident, but wrong prior will often approach the correct solution more slowly than an ignorant one.

2.1 Reliabilities: best estimates, error-bars and confidence intervals

One way to summarize the posterior pdf is with two quantities: the best estimate and its reliability. The best estimate is given by the maximum value of the pdf

$$\left.\frac{dP}{dX}\right|_{X_O} = 0 \tag{10}$$

where X_O denotes the best estimate. To make sure we have a maximum, we also need to check the second derivative

$$\left.\frac{d^2P}{dX^2}\right|_{X_O} = 0. \tag{11}$$

Using derivatives like this assumes X is continuous. If this is not the case, the best estimate will still be the value corresponding to the max of the pdf.

The reliability of the best estimate is found by considering the width of the pdf about X_O. We take the logarithm of the pdf as this varies more slowly with X, making it easier to work with.

$$L = ln[P(X|\{data\}, I)]. \tag{12}$$

Doing a Taylor expansion about X_O and using the condition

$$\left.\frac{dL}{dX}\right|_{X_O} = 0, \tag{13}$$

which is equivalent to (10), leads to

$$P(X|\{data\}, I) \approx A\ exp\left[\frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_O}(X-X_O)^2\right]. \tag{14}$$

Here, we only show the dominating quadratic term of the expansion, with A being a normalization constant. We have now approximated our pdf by the *normal distribution*, typically taking the form

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\ exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{15}$$

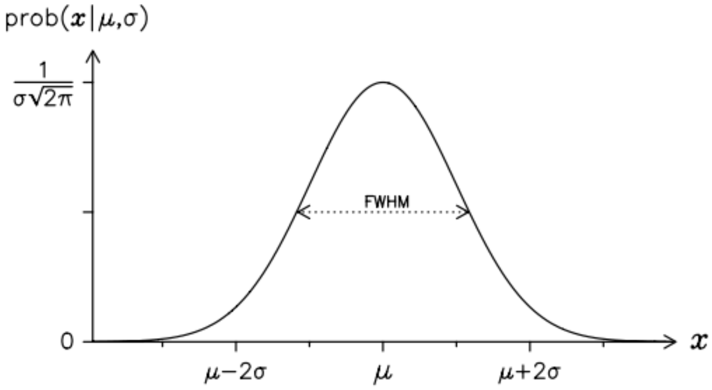

Figure 1: The normal distribution with a maximum at $x = \mu$ and a full width at half maximum (FWHM) of 2.35σ. (Sivia, 2006, p. 22) [1]

The parameter σ is called the *error-bar* and is defined as

$$\sigma = \left(-\frac{d^2L}{dX^2}\bigg|_{X_O}\right)^{-1/2}. \tag{16}$$

We then infer the quantity of interest by the following

$$X = X_O \pm \sigma. \tag{17}$$

By calculating the integral of the normal distribution in this range, we get a 67% chance that X lies within $X_O \pm \sigma$ and a 95% chance that it lies within $X_O \pm 2\sigma$.

2.1.1 Asymmetric pdfs

The error-bar needs a symmetric pdf to be valid, something that is often not the case. This is solved by replacing the error-bar with a *confidence interval* as a measure of reliability. It is defined as the shortest interval that encloses 95% of the area of the pdf. In short, we find X_1 and X_2 such that

$$P(X_1 \leq X \leq X_2|\{data\}, I) = \int_{X_1}^{X_2} P(X|\{data\}, I)dX \approx 0.95, \tag{18}$$

assuming the pdf is normalized. The 95% confidence level is conventionally seen as a sensible value, being a rather conservative estimate.

In the case of an asymmetric pdf, we may consider using the *mean* or *expectation* as the best estimate. This quantity takes the skewness of the pdf into account, and is given by

$$\langle X \rangle = \int X P(X|\{data\}, I) dX. \tag{19}$$

If the pdf is not normalized, we also need to divide the right-hand side by $\int P(X|\{data\}, I)dX$.

If the pdf is *multimodal*, meaning it has multiple maxima, it becomes more difficult to calculate a best estimate and its reliability. If one maximum is much greater than the others, we can ignore those other contributions and focus on the largest. However, if multiple peaks are of similar size, we would be better off displaying the pdf itself.

Another common pdf which is often used is the Cauchy distribution, shown below.

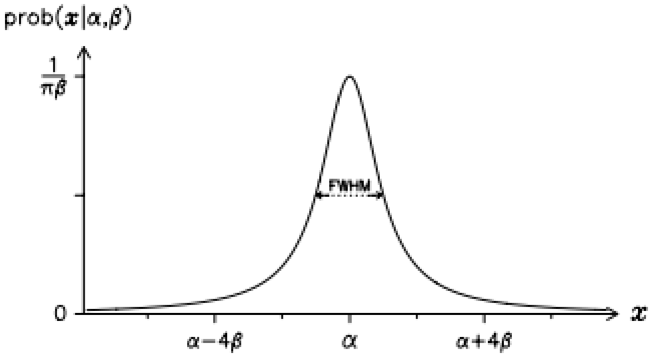

Figure 2: The Cauchy distribution, symmetric about $x = \alpha$ and has a FWHM of 2β. (Sivia, 2006, p. 31) [1]

This distribution has very wide wings and is in this case given by

$$P(x|\alpha, \beta, I) = \frac{\beta}{\pi[\beta^2 + (x-\alpha)^2]} \tag{20}$$

3 Parameter estimation II

This chapter covers cases with multiple parameters, a generalization of the previous chapter. We will learn how to estimate parameters of interest and how to deal with nuisances.

3.1 Example: Amplitude of a signal in the presence of background

This is a two-parameter problem, consisting of a flat background signal B and a peak signal A, both of unknown magnitude. We receive an integer-valued set of counts $\{N_k\}$ at experimental settings $\{x_k\}$. Assuming a Gaussian shape of A, one ideal sample is given by

$$D_K = n_o[Ae^{-(x_k-x_o)^2/2w^2} + B]. \tag{21}$$

Here, n_o is a constant related to time elapsed, x_o is the center and w is the width of the peak. D_k will not necessarily be an integer, meaning the actual sample will be an integer close to this value. This is achieved with the discretely valued *Poisson distribution*:

$$P(N|D) = \frac{D^N e^{-D}}{N!}. \tag{22}$$

Its expectation value is, as wished, equal to D:

$$\langle N \rangle = \sum_{N=0}^{\infty} NP(N|D) = D. \tag{23}$$

We then have the probability of finding one number of counts N_k

$$P(N_k|A,B,I) = \frac{D_k^{N_k} e^{-D_k}}{N_k!}, \tag{24}$$

and if the data are independent, the probability for the entire set of counts is then the product of the probabilities of each count:

$$P(\{N_k\}|A,B,I) = \prod_{k=1}^{M} P(N_k|A,B,I). \tag{25}$$

This is our likelihood function. For a suitable prior, we know that none of the signals can be negative, meaning we can use the following:

$$P(A,B|I) = \begin{cases} const. & A \geq 0, B \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{26}$$

Now we multiply the prior and the likelihood function according to Bayes' theorem and take the logarithm of the result:

$$L = ln[P(A,B|\{N_k\},I)] = const. + \sum_{k=1}^{M}[N_k ln(D_k) - D_k] \tag{27}$$

The constant contains all terms not dependent on A and B, and we find the best estimate for A and B by maximizing L. This can be shown in a two-dimensional plot with contours representing points of equal probability. As the number of parameters increase we can still do the above procedure and find the best estimate. Plotting the probabilities will however be more and more difficult as we increase the number of dimensions, eventually becoming impossible.

3.1.1 Marginal distributions

As B is a background signal, it can be considered to be of no interest, a nuisance parameter. By using the marginalization equation (5), we can safely ignore B.

$$P(A|\{N_k\},I) = \int_0^\infty P(A,B|\{N_k\},I)dB. \tag{28}$$

This can be done multiple times for multiple parameters.

3.2 Reliabilities

As we now have multiple parameters, the best estimate is now given by the set of simultaneous equations

$$\left.\frac{\partial P}{\partial X_i}\right|_{\{X_{Oj}\}} = 0. \tag{29}$$

Assuming the set of parameters consists of X and Y, we do a procedure similar to the one parameter case. We take the logarithm of the probability and find its Taylor expansion.

Assuming we only want the value of X, we can integrate out the Y-parameter. The resulting posterior pdf will then have the error bar

$$\sigma_X = \sqrt{\frac{-B}{AB - C^2}}. \tag{30}$$

Similarly, if we were only interested in Y:

$$\sigma_Y = \sqrt{\frac{-A}{AB - C^2}}. \tag{31}$$

where

$$A = \left.\frac{\partial^2 L}{\partial X^2}\right|_{X_O,Y_O}, B = \left.\frac{\partial^2 L}{\partial Y^2}\right|_{X_O,Y_O}, C = \left.\frac{\partial^2 L}{\partial X \partial Y}\right|_{X_O,Y_O}, \tag{32}$$

from the Taylor expansion of L.

For a more complete picture of the reliability we can use the *variance*. This quantity can be calculated for pdfs of multiple variables:

$$\sigma_X^2 = \langle (X - X_O)^2 \rangle = \iint (X - X_O)^2 P(X,Y|\{data\},I)dXdY. \tag{33}$$

We can also consider the *covariance*, the simultaneous deviations of X and Y:

$$\sigma_{XY}^2 = \langle (X - X_O)(Y - Y_O) \rangle = \iint (X - X_O)(Y - Y_O)P(X,Y|\{data\},I)dXdY \tag{34}$$

This is a measure of the correlation between X and Y. From this, we can make the *covariance matrix*:

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \sigma_{XY}^2 & \sigma_Y^2 \end{pmatrix} = \frac{1}{AB - C^2} \begin{pmatrix} -B & C \\ C & -A \end{pmatrix} = - \begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1} \tag{35}$$

In the case that $\sigma_{XY}^2 = 0$, there is no correlation between the two parameters.

3.2.1 Generalization of the quadratic approximation

This is the general case of the quadratic approximation of the posterior pdf, valid for several parameters:

$$P(\boldsymbol{X}|\{data\}, I) \propto exp\left[\frac{1}{2}(\boldsymbol{X} - \boldsymbol{X_O})^T \boldsymbol{\nabla\nabla} L(\boldsymbol{X_O})(\boldsymbol{X} - \boldsymbol{X_O})\right] \tag{36}$$

Additionally, the general covariance matrix is given by

$$[\boldsymbol{\sigma^2}]_{ij} = \big\langle (X_i - X_{Oi})(X_j - X_{Oj}) \big\rangle = -\Big[(\boldsymbol{\nabla\nabla L})^{-1}\Big]_{ij} \tag{37}$$

3.3 Algorithms

In many cases, we are not able to find the maximum of a pdf analytically. That is when we need numerical algorithms to help us.

3.3.1 Brute force and ignorance

For problems consisting of up to three parameters, an easy way to find the maximum is to plot the pdf and look for it. The whole pdf can then be displayed and we gain complete overview, but as the number of parameters and resolution increase, this procedure rapidly reaches extreme time requirements.

3.3.2 Linearity

The maximum is found by demanding

$$\boldsymbol{\nabla} L(\boldsymbol{X_O}) = 0. \tag{38}$$

If we are lucky and this is linear, meaning we can write

$$\boldsymbol{\nabla} L = \boldsymbol{HX} + \boldsymbol{C} \tag{39}$$

we can turn to matrix operations for finding the maximum:

$$\boldsymbol{X_O} = -\boldsymbol{H}^{-1}\boldsymbol{C}. \tag{40}$$

As the covariance matrix is given by the inverse of $\boldsymbol{\nabla\nabla L}$, we get a complete summary of the pdf:

$$[\boldsymbol{\sigma^2}]_{ij} = \big\langle (X_i - X_{Oi})(X_j - X_Oj) \big\rangle = -[\boldsymbol{H}^{-1}]_{ij}. \tag{41}$$

We can avoid having to calculate the inverse in (40) by instead using a matrix decomposition method, for example the *Cholesky decomposition*.

If the determinant of H is close or equal to zero, this may result in explosion. To avoid this, we need more relevant data or prior information.

3.3.3 Iterative linearization

If we are not able to write the gradient of L as the linear form above, we may use the *Newton-Raphson* algorithm, realized by the Taylor-expansion of $\boldsymbol{\nabla} L$, to find an estimate of the maximum:

$$\boldsymbol{X}_{N+1} = \boldsymbol{X}_N - \left[\boldsymbol{\nabla\nabla} L(\boldsymbol{X}_N) + c\boldsymbol{I}\right]^{-1} \boldsymbol{\nabla L}(\boldsymbol{X}_N) \tag{42}$$

where c is a small negative number multiplied with the identity matrix $\boldsymbol{I}$. This is to slow down the algorithm, making it more accurate.

For complex, multimodal pdfs, Monte Carlo methods are good approaches for finding the maximum values. These problems are much harder to solve, and we are not guaranteed an answer in a reasonable time.

3.3.4 Changing variables

Given an estimation of multiple variables, we are able to relate them to each other with this general formula:

$$P(\{X_j\}|I) = P(\{Y_j\}|I) \times \left|\frac{\partial(\{Y_j\})}{\partial(\{X_j\})}\right| \tag{43}$$

where the right-most term is called the *Jacobian*.

4 Model selection

Assuming we have two possible theories A and B, we compare their validity by taking the ratio of their posterior pdfs

$$\text{posterior ratio} = \frac{P(A|D,I)}{P(B|D,I)}. \tag{44}$$

Theory A is better if this quantity is much larger than one, and B is better if it is much smaller. If it is roughly one, we need better data to make a clear choice. We can apply Bayes' theorem to these posteriors and arrive at

$$\frac{P(A|D,I)}{P(B|D,I)} = \frac{P(D|A,I)}{P(D|B,I)} \times \frac{P(A|I)}{P(B|I)}. \tag{45}$$

If we were to include an adjustable parameter λ to B, we would get the following:

$$\frac{P(A|D,I)}{P(B|D,I)} = \frac{P(A|I)}{P(B|I)} \times \frac{P(D|A,I)}{P(D|\lambda_O,B,I)} \times \frac{\lambda_{\max} - \lambda_{\min}}{\delta\lambda\sqrt{2\pi}} \tag{46}$$

with the assumptions that λ lies between λ_{max} and λ_{min}, and that the likelihood function $P(D|\lambda, B, I)$ is Gaussian in shape. This can be extended to multiple paramaters for both A and B, and shows the fact that the best models are almost always the ones with the lowest degree of complexity that still fit the data well.

This method of using the ratio of posteriors is useful for being able to quantify the validity of a hypothesis.

5 Assigning probabilities

So far, we have seen how to make use of pdfs for tackling problems in data analysis. We will now look at how to assign these probabilities.

5.0.1 The binomial distribution

In problems with two possible outcomes, often called success and failure, the *binomial distribution* gives the probability of having r successes in N trials:

$$P(r|N, I) = \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r} \tag{47}$$

where p is the probability of a single success. The mean value of r and its mean squared error is given by

$$\langle r \rangle = Np \quad \text{and} \quad \langle (r - Np)^2 \rangle = Np(1-p). \tag{48}$$

5.0.2 Location and scale parameters

Assuming X is a location parameter, a complete ignorance entails assigning a uniform prior:

$$P(X|I) = const. \tag{49}$$

If a small change in X would result in a significant change in our prior probability, it would mean we have some non-ignorant knowledge of the situation beforehand.

When it comes to *scale* parameters, meaning we are interested in relative change as opposed to absolute change, the most ignorant prior is

$$P(X|I) \propto 1/X. \tag{50}$$

This corresponds to a uniform pdf for the logarithm of X.

5.0.3 Maximum entropy

In general, we should choose the prior that satisfies all available conditions and which contains the most entropy S. In other words, we should maximize

$$S = -\sum_{i=1}^{M} p_i log \left[\frac{p_i}{m_i} \right] \tag{51}$$

for discrete parameters, and

$$S = -\int p(x) log \left[\frac{p(x)}{m(x)} \right] dx \tag{52}$$

for continuous parameters. Here, $p(x)$ denotes the normalized probability constrained to all known conditions. The $m(x)$ term is called the *Lebesgue measure* and transforms in the same way as $p(x)$, ensuring S is invariant with an eventual variable change. $m(x)$ is described as any multiple of $p(x)$ that has complete ignorance about x.

6 Non-parametric estimation

In this chapter, we take a look at the difficult case where we do not have enough information to describe the object of interest by a specific function.

6.1 Free-form solutions

When we have too little information about the object of interest, we have to make a choice allowing for a large amount of flexibility. This means we have to deal with a lot of parameters, which is problematic. We must then try to search for weak constraints that can make such a free-form solution $f(x)$ satisfactory. A good way to start is to discretize the function. Assuming $f(x)$ is a one-dimensional spectrum, we can divide the x-axis into a set of small pieces, and define a_j as the average value of $f(x)$ over such a sub-interval.

One candidate for a free-form solution is a linear combination of basis functions:

$$f(x) = \sum_{l=1}^{M_b} c_l \eta_l(x) \tag{53}$$

and the discretized version:

$$f(x_j) = a_j = \sum_{l=1}^{M_b} c_l \eta_l(x_j) \tag{54}$$

The aim of this is to reduce the problem to parameter estimation, and the best choice of basis functions is the one that makes M_b as small as possible.

6.2 MaxEnt

Assuming the object of interest is comprised of a positive and additive distribution, we may use maximum entropy to make a suitable prior:

$$P(\{a_j\}|\{m_j\}, \alpha, I) \propto \exp(\alpha S) \tag{55}$$

with S being the entropy of $\{a_j\}$ relative to the Lebesgue measure $\{m_j\}$. A generalized form of S is

$$S = \sum_{j=1}^{M} \left(a_j - m_j - a_j log \left[\frac{a_j}{m_j} \right] \right). \tag{56}$$

If we however cannot know if we are dealing with an additive and positive pdf, an alternative prior is:

$$P(\{a_j\}|\{m_j\}, \{\gamma_j\}, \alpha, I) \propto \exp\left[-\alpha \sum_{j=1}^{M} \frac{(a_j - m_j)^2}{\gamma_j^2} \right], \tag{57}$$

a multivariate Gaussian distribution.

7 Expiremental design

This chapter involves the optimization of experiments and effectiveness of data. From Bayes' theorem we know that the likelihood function is the term that introduces data to the calculation. We want the data to be as influential as possible, leading us to learn the most. To achieve this, we need the likelihood function to be sharply peaked, dominating the prior term, the information we already know. This can often be difficult.

To fulfill these conditions, we need to investigate the likelihood functions dependency on parameters. Such parameters include the amount of time used by the experiment, the number of counts of data, any background signals, as well as the shape of any additional functions included in the analysis (e.g. resolution functions). In the case of model selection, an optimized experiment is usually one that gives rise to the largest possible difference between the hypotheses tested.

7.1 Quantifying the worth of an experiment

Using entropy, we are able to put a number to the information gain from an experiment:

$$\mathcal{H}(D) = \int \frac{P(D|X, I)P(X, I)}{P(D|I)} \log_2 \left[\frac{P(D|X, I)}{P(D|I)} \right] dX, \tag{58}$$

measured in *bits*. We can also estimate the expectation value of this before receiving the data:

$$\langle \mathcal{H} \rangle = \int \mathcal{H}(D)P(D|I)dD, \tag{59}$$

also called the *benefit*.

8 Least-squares extensions

Least-squares is a very popular method when tackling problems in data analysis. We will in this chapter look at the basic procedure of least-squares and relevant considerations.

8.1 Constraints and restraints

The least-squares method is based on minimizing the χ^2-function given by:

$$\chi^2 = \sum_{k=1}^{N} R_k^2, \tag{60}$$

where $(F_k - D_k)/\sigma_k$ is the difference between the predicted (F_k) and real (D_k) data divided by the expected mismatch σ_k. This follows from a uniform prior and a Gaussian likelihood function:

$$P(\mathbf{D}|\mathbf{X}, I) = \prod_{k=1}^{N} \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(- \frac{R_k^2}{2} \right) \propto \exp\left(- \frac{\chi^2}{2} \right). \tag{61}$$

The logarithm of the posterior is then:

$$L = \log_e[P(\mathbf{X}|\mathbf{D}, I)] = \text{const.} - \frac{\chi^2}{2}. \tag{62}$$

If we were to know beforehand that $X_j = x_{oj} \pm \epsilon_j$, we can assign a Gaussian pdf for the prior:

$$P(\mathbf{X}|I) = \prod_{j=1}^{M} \frac{1}{\epsilon_j \sqrt{2\pi}} \exp\left[- \frac{(X_j - x_{oj})^2}{2\epsilon_j^2} \right] \propto \exp\left(- \frac{C}{2} \right) \tag{63}$$

where

$$C = \sum_{j=1}^{M} \left(\frac{X_j - x_o j}{\epsilon_j} \right)^2 \tag{64}$$

is called the constraint function. The logarithm then takes the form:

$$L = \exp[P(\mathbf{X}|\mathbf{D}, I)] = \text{const.} - \frac{1}{2}[\chi^2 + C]. \tag{65}$$

When the uncertainties $\{\epsilon_j\}$ are very large, C goes to zero and we get the ordinary least-squares procedure. When some of the ϵ_j goes to zero, meaning we have strong prior information, the constraints are called restraints and holds the relevant X_j to their assumed values.

8.2 Noise scaling

In the case where we have no estimates for the uncertainties $\{\sigma_k\}$, we need to make some assumptions. For instance that all σ_k are equal to an unknown constant, or that they are related to the square root of the data $\sigma_k = \sigma\sqrt{D_k}$, or any other fitting guesses.

If we only have a single σ, the likelihood in (61) is slightly modified to have the following proportionality:

$$P(\mathbf{D}|\sigma, \mathbf{X}, I) \propto \frac{1}{\sigma^N} \exp\left(-\frac{\chi^2}{2\sigma^2} \right). \tag{66}$$

We then use marginalization to integrate out the uncertainty and arrive at the new likelihood:

$$P(\mathbf{D}|\mathbf{X}, I) = \int_0^\infty P(\mathbf{D}|\sigma, \mathbf{X}, I) P(\sigma|I) d\sigma \propto \int_0^\infty \left(\frac{2t}{\chi^2} \right)^{N/2-1} e^{-t} \frac{dt}{\chi^2} \tag{67}$$

where $t = \chi^2/2\sigma^2$. The logarithm of the new posterior then becomes:

$$L = \log_e[P(\mathbf{X}|\mathbf{D}, I)] = \text{const.} - \frac{N}{2} \log_e(\chi^2). \tag{68}$$

8.3 Outliers

We may in many cases have a few data points that deviate greatly from the rest and do not fit to a straight line approximation. In cases where these can not be ignored as faults in measurements, we need a way to deal with them.

8.3.1 A conservative formulation

We may start by assuming the estimated uncertainties represent lower bounds of the real noise. This can be represented by the pdf:

$$P(\sigma|\sigma_o, I) = \frac{\sigma_o}{\sigma^2} \tag{69}$$

for $\sigma \geq \sigma_o$. From this, the likelihood is as follows:

$$P(D|F, \sigma_o, I) = \frac{1}{\sigma_o\sqrt{2\pi}} \left[\frac{1 - e^{-R^2/2}}{R^2} \right], \tag{70}$$

and assuming independent noise, the logarithm of the posterior becomes

$$L = \text{const.} + \sum_{k=1}^{N} \log_e \left[\frac{1 - e^{-R_k^2/2}}{R_k^2} \right]. \tag{71}$$

Other possible techniques include using the *good-and-bad* data model, meaning we allow two options: Either we believe the estimated σ, or something has gone wrong and we largely increase the noise-value:

$$P(\sigma|\sigma_o, \beta, \gamma, I) = \beta\delta(\sigma - \gamma\sigma_o) + (1 - \beta)\delta(\sigma - \sigma_o) \tag{72}$$

where $0 \leq \beta \ll 1$ and $\gamma \gg 1$.

Another possibility is the *Cauchy formulation* where we expect that σ is of the same order as σ_o, but may have a different width:

$$P(\sigma|\sigma_o,) = \frac{2\sigma_o}{\sqrt{\pi}\sigma^2}\exp\left(-\frac{\sigma_o^2}{\sigma^2}\right). \tag{73}$$

Other techniques may be used to help the least-squares procedure along, such as background removal (separating the background and signal before analysing the data), or the inclusion of a covariance matrix when dealing with correlations between uncertainties.

Final thoughts

This book has given me new insight towards the field of statistics and more specifically, the Bayesian way. Several ideas of varying degrees of complexity have been presented, and I have appreciated the authors efforts to achieve elegance and simplicity where possible, and tried my best to understand when the difficulty has risen. I do share the desire for a general, underlying set of principles for tackling data analysis, as memorizing countless specialized recipes is not by any means my forte. This book has shown me that this desire is achievable and the result is very potent and versatile, of course not without requiring effort.

References

[1] Sivia D.S., Skilling, J. (2006) *Data Analysis - a Bayesian tutorial, 2nd edition.* New York: Oxford University Press