Learning from data: Model Selection

Christian Forssén

Department of Physics, Chalmers University of Technology, Sweden

Sep 30, 2019

So far, we have been concerned with the problem of parameter estimation. In studying the linear relationship between two quantities, for example, we discussed how to infer the slope and the offset of the associated straight-line model. Often, however, there is a question as to whether a quadratic, or even cubic, function might be a more appropriate model. In this lecture, we will consider the broad class of scientific problems when there is uncertainty as to which one of a set of alternative models is most suitable. In the Bayesian terminology these can be labeled as **Model Selection** problems and we will discuss them in some depth.

We will start, however, with a brief discussion on the frequentist approach to hypothesis testing.

1 Frequentist hypothesis testing

One of the main objectives in science is that of inferring the truth of one or more hypotheses about how some aspect of nature works. Because we are always in a state of incomplete information, we can never prove any hypothesis (theory) is true.

1.1 Basic idea

Recall that in frequentist statistics, probability statements are restricted to random variables. A hypothesis can not be considered a random variable, and therefore we are restricted to a much more indirect approach when trying to infer its truth, or rather when attempting to falsify it. The basic idea is the following:

- The standard sampling theory approach to hypothesis testing is to construct a statistical test.
- The test usually compares a null hypothesis ($\mathcal{H}_0$) with an alternative hypothesis ($\mathcal{H}_A$). We will see an example below.
- The null hypothesis is accepted or rejected purely on the basis of how unexpected the data were to $\mathcal{H}_0$, not on how much better $\mathcal{H}_A$ predicted the data.

- The degree of "unexpectedness" is based on a chosen statistic, such as the sample mean or the χ^2 statistic. This statistic is first computed for the observed data set.
- Then it is computed for a very large number of hypothetical repeated measurements under the assumption that the null hypothesis is true.
- The value of the statistic from our actual data set is compared with the distribution that is associated with the truth of the null hypothesis.
- If the statistic from the observed data falls in a very unlikely spot on this distribution (the threshold is to be defined beforehand) we choose to reject the null hypothesis at some confidence level on the basis of the measured data set.

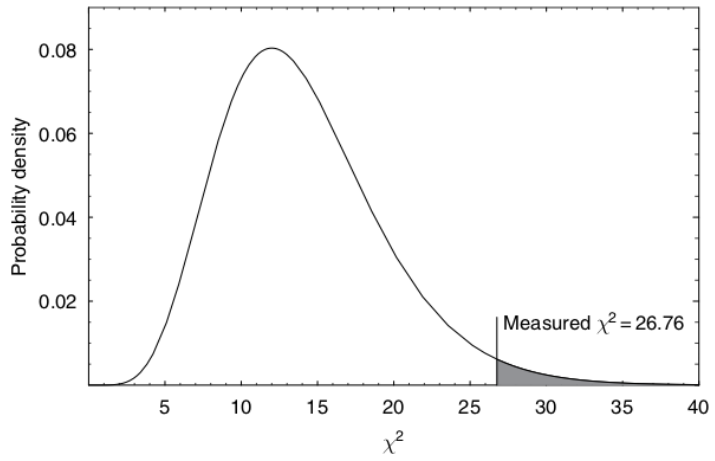
Hypothesis testing with the chi-squared statistic. A very common statistic to use is the chi-square. A good example is found in Gregory, ch 7.2.1, with the measurements of flux density from a distant galaxy over a period of 6000 days.

- Choose a null hypothesis that the galaxy has an unknown, but constant, flux density. If we can reject this hypothesis at e.g. the 95% confidence level, then this provides indirect evidence(?) for the alternative hypothesis that the radio emission is variable.
- In this example, it is assumed that he measurement errors are independently normal with a fixed standard deviation σ that is known beforehand.
- The χ^2 statistic from the data set is evaluated (where x_i is the data and $\bar{x}$ is the average from the sample)

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2}.$$

- In our example we had 15 data points, but we are using it to estimate the mean μ. Therefore, we loose one degree of freedom and are left with 14. This number will determine the form of the χ^2 distribution that will be used for comparison with our actual χ^2 statistic.
- The question of how unlikely is this value of χ^2 is by convention interpreted in terms of the area in the tail of the χ^2 distribution beyond this line. This is called the P-value or significance.

At the point of performing this comparison, and making a final statement, the sampling theory school divides itself into two camps:



Figur 1: The χ^2 distribution for 14 degrees of freedom. The value computed from the measurements of flux density from a galaxy is indicated by a vertical line. The shaded area is the P-value. (Gregory, Fig. 7.2)

1. One camp uses the following protocol: first, before looking at the data, pick the significance level of the test (e.g. 5%), and determine the critical value of χ^2 above which the null hypothesis will be rejected. (The significance level is the fraction of times that the statistic χ^2 would exceed the critical value, if the null hypothesis were true.) Then, compare the actual χ^2 with the critical value, and declare the outcome of the test, and its significance level (which was fixed beforehand).

2. The second camp looks at the data, finds χ^2, then looks in the table of χ^2-distributions for the significance level, P, for which the observed value of χ^2 would be the critical value. The result of the test is then reported by giving this value of p, which is the fraction of times that a result as extreme as the one observed, or more extreme, would be expected to arise if the null hypothesis were true.

2 Bayesian model selection

2.1 The Story of Dr. A and Prof. B

[Reproduced, with some modifications, from Sivia, 2006].

> *Dr. A has a theory; Prof. B also has a theory, but with an adjustable parameter λ. Whose theory should we prefer on the basis of data D?*—
> Jefferys (1939), Gull (1988), Sivia (2006)

It is clear that we need to evaluate the posterior probabilities for A and B being correct to ascertain the relative merit of the two theories. If the ratio of the posterior probabilities,

$$\text{posterior ratio} = \frac{p(A|D,I)}{p(B|D,I)} \tag{1}$$

is very much greater than one, then we will prefer A's theory; if it is very much less than one, then we prefer that of B; and if it is of order unity, then the current data are insufficient to make an informed judgement.

To estimate the odds, let us start by applying Bayes' theorem to both the numerator and the denominator; this gives

$$\frac{p(A|D,I)}{p(B|D,I)} = \frac{p(D|A,I)p(A|I)}{p(D|B,I)p(B|I)} \tag{2}$$

because the term $p(D|I)$ cancels out, top and bottom. As usual, probability theory warns us immediately that the answer to our question depends partly on what we thought about the two theories before the analysis of the data. To be fair, we might take the ratio of the prior terms, on the far right of Eq. (2), to be unity; a harsher assignment could be based on the past track records of the theorists! To assign the probabilities involving the experimental measurements, $p(D|A,I)$ and $p(D|B,I)$, we need to be able to compare the data with the predictions of A and B: the larger the mismatch, the lower the corresponding probability. This calculation is straightforward for Dr A, but not for Prof B; the latter cannot make predictions without a value for λ.

To circumvent this difficulty, we can use the sum and product rule to relate the probability we require to other pdfs which might be easier to assign. In particular, marginalization and the product rule allow us to express $p(D|B,I)$ as

$$p(D|B,I) = \int d\lambda p(D,\lambda|B,I) = \int d\lambda p(D|\lambda,B,I)p(\lambda|B,I). \tag{3}$$

The first term in the integral $p(D|\lambda,B,I)$, where the value of λ is given, is now just an ordinary likelihood function; as such, it is on a par with $p(D|A,I)$. The second term is B's prior pdf for λ; the onus is, therefore, on the theorist to articulate his or her state of knowledge, or ignorance, before getting access to the data.

To proceed further analytically, let us make some simplifying approximations. Assume that, a priori, Prof B is only prepared to say that λ must lie between the limits $\lambda_{\min}$ and $\lambda_{\max}$; we can then naively assign a uniform prior within this range:

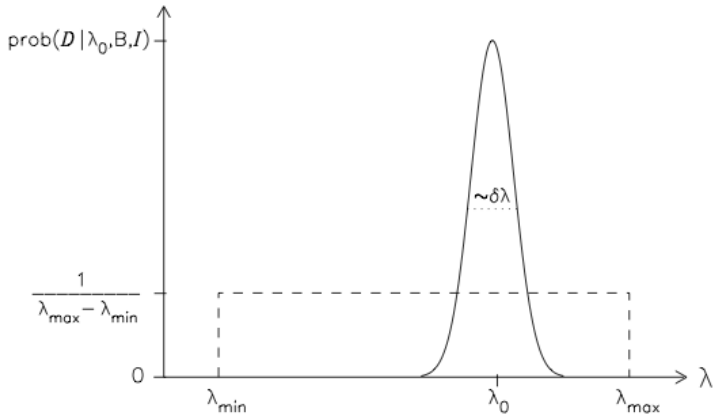
$$p(\lambda|B,I) = \frac{1}{\lambda_{\max} - \lambda_{\min}} \quad \text{for } \lambda_{\min} \leq \lambda \leq \lambda_{\max}, \tag{4}$$

and zero otherwise. Let us also take it that there is a value λ_0 which yields the closest agreement with the measurements; the corresponding probability $p(D|\lambda_0,B,I)$ will be the maximum of B's likelihood function. As long as this

adjustable parameter lies in the neighbourhood of the optimal value, $\lambda_0 \pm \delta\lambda$, we would expect a reasonable fit to the data; this can be represented by the Gaussian pdf

$$p(D|\lambda, B, I) = p(D|\lambda_0, B, I) \exp\left[-\frac{(\lambda - \lambda_0)^2}{2\delta\lambda^2}\right]. \tag{5}$$

The assignments of the prior (4) and the likelihood (5) are illustrated in Fig. 2. We may note that, unlike the prior pdf $p(\lambda|B, I)$, B's likelihood function need not be normalized with respect to λ; in other words, $p(D|\lambda_0, B, I)$ need not equal $1/\delta\lambda\sqrt{2\pi}$. This is because the λ in $p(D|\lambda, B, I)$ appears in the conditioning statement, whereas the normalization requirement applies to quantities to the left of the '|' symbol.



Figur 2: A schematic representation of the prior pdf (dashed line) and the likelihood function (solid line) for the parameter λ in Prof B's theory.

In the evaluation of $p(D|B, I)$, we can make use of the fact that the prior does not depend explicitly on λ; this enables us to take $p(\lambda|B, I)$ outside the integral in Eq. (3)

$$p(D|B, I) = \frac{1}{\lambda_{\max} - \lambda_{\min}} \int_{\lambda_{\min}}^{\lambda_{\max}} d\lambda p(D|\lambda, B, I), \tag{6}$$

having set the limits according to the specified range. Assuming that the sharp cut-offs do not cause a significant truncation of the Gaussian pdf of the likelihood, its integral will be equal to $\delta\lambda\sqrt{2\pi}$ times $p(D|\lambda_0, B, I)$. The troublesome term then reduces to

$$p(D|B, I) = \frac{1}{\lambda_{\max} - \lambda_{\min}} p(D|\lambda_0, B, I)\delta\lambda\sqrt{2\pi}.$$

Substituting this into Eq. (2), we finally see that the ratio of the posteriors required to answer our original question decomposes into the product of three

terms:

$$\frac{p(A|D,I)}{p(B|D,I)} = \frac{p(A|I)}{p(B|I)} \times \frac{p(D|A,I)}{p(D|\lambda_0,B,I)} \times \frac{\lambda_{\max} - \lambda_{\min}}{\delta\lambda\sqrt{2\pi}}. \tag{7}$$

The first term on the right-hand side reflects our relative prior preference for the alternative theories; to be fair, we can take it to be unity. The second term is a measure of how well the best predictions from each of the models agree with the data; with the added flexibility of his adjustable parameter, this maximum likelihood ratio can only favour B. The goodness-of-fit, however, cannot be the only thing that matters; if it was, we would always prefer more complicated explanations. Probability theory tells us that there is, indeed, another term to be considered. As assumed earlier in the evaluation of the marginal integral of Eq. (3), the prior range $\lambda_{\max} - \lambda_{\min}$ will generally be much larger than the uncertainty $\pm\delta\lambda$ permitted by the data. As such, the final term in Eq. (7) acts to penalize B for the additional parameter; for this reason, it is often called an Ockham factor. That is to say, we have naturally encompassed the spirit of Ockham's Razor: *'Frustra fit per plura quod potest fieri per pauciora'* or, in English, *'it is vain to do with more what can be done with fewer'* .

Although it is satisfying to quantify the everyday guiding principle attributed to the thirteenth-century Franciscan monk William of Ockham (or Occam, in Latin), that we should prefer the simplest theory which agrees with the empirical evidence, we should not get too carried away by it. After all, what do we mean by the simpler theory if alternative models have the same number of adjustable parameters? In the choice between Gaussian and Lorentzian peak shapes, for example, both are defined by the position of the maximum and their width. All that we are obliged to do, and have done, in addressing such questions is to adhere to the rules of probability.

While accepting the clear logic leading to Eq (7), many people rightly worry about the question of the limits $\lambda_{\min}$ and $\lambda_{\max}$. Jeffreys (1939) himself was concerned and pointed out that there would be an infinite penalty for any new parameter if the range was allowed to go to $\pm\infty$. Stated in the abstract, this would appear to be a severe limitation. In practice, however, it is not generally such a problem: since the analysis is always used in specific contexts, a suitable choice can usually be made on the basis of the relevant background information. Even in uncharted territory, a small amount of thought soon reveals that our state of ignorance is always far from the $\pm\infty$ scenario. If λ was the coupling constant (or strength) for a possible fifth force, for example, then we could put an upper bound on its magnitude because everybody would have noticed it by now if it had been large enough! We should also not lose sight of the fact that the precise form of Eq (7) stems from our stated simplifying approximations; if these are not appropriate, then eqns (2) and (3) will lead us to a somewhat different formula.

In most cases, our relative preference for A or B is dominated by the goodness of the fit to the data; that is to say, the maximum likelihood ratio in eqn (7) tends to overwhelm the contributions of the other two terms. The Ockham factor can play a crucial role, however, when both theories give comparably good agreement with the measurements. Indeed, it becomes increasingly important if B's theory

fails to give a significantly better fit as the quality of the data improves. In that case, $\delta\lambda$ continues to become smaller but the ratio of best-fit likelihoods remains close to unity; according to Eq. (7), therefore, A's theory is favoured ever more strongly. By the same token, the Ockham effect disappears if the data are either few in number, of poor quality or just fail to shed new light on the problem at hand. This is simply because the posterior ratio of Eq. (7) is then roughly equal to the complementary prior one, since the empirical evidence is very weak; hence, there is no inherent preference for A's theory unless it is explicitly encoded in $p(A|I)/p(B|I)$. This property can be verified formally by going back to Eqs. (2), (5) and (6), and considering the poor-data limit in which $\delta\lambda \gg \lambda_{\max} - \lambda_{\min}$ and $p(D|\lambda_0, B, I) \approx p(D|A, I)$.

One adjustable parameter each. Some further interesting features arise when we consider the case where Dr A also has one adjustable parameter; call it μ. If we make the same sort of probability assignments, and simplifying approximations, as for Prof B, then we find that

$$\frac{p(A|D,I)}{p(B|D,I)} = \frac{p(A|I)}{p(B|I)} \times \frac{p(D|\mu_0, A, I)}{p(D|\lambda_0, B, I)} \times \frac{\delta\mu(\lambda_{\max} - \lambda_{\min})}{\delta\lambda(\mu_{\max} - \mu_{\min})}. \tag{8}$$

This could represent the situation where we have to choose between a Gaussian and Lorentzian shape for a signal peak, but one associated parameter is not known. The position of the maximum may be fixed at the origin by theory, for example, and the amplitude constrained by the normalization of the data; A and B could then be the hypotheses favouring the alternative lineshapes, where $\delta\mu$ and $\delta\lambda$ are their related full-width-half-maxima. If we give equal weight to A and B before the analysis, and assign a similar large prior range for both μ and λ, then Eq. (8) reduces to

$$\frac{p(A|D,I)}{p(B|D,I)} \approx \frac{p(D|\mu_0, A, I)}{p(D|\lambda_0, B, I)} \times \frac{\delta\mu}{\delta\lambda}.$$

For data of good quality, the dominant factor will tend to be the best-fit likelihood ratio. If both give comparable agreement with the measurements, however, then the shape with the larger error-bar for its associated parameter will be favoured. At first sight, it might seem rather odd that the less discriminating theory can gain the upper hand. It appears less strange once we realize that, in the context of model selection, a larger 'error-bar' means that more parameter values are consistent with the given hypothesis; hence its preferential treatment.

One adjustable parameter each; different prior ranges. Finally, we can also consider the situation where Mr A and Mr B have the same physical theory but assign a different prior range for λ (or μ). Although Eq. (7) can be seen as representing the case when $(\mu_{\max} - \mu_{\min})$ is infinitesimally small, so that A has no flexibility, Eq. (8) is more appropriate when the limits set by both theorists are large enough to encompass all the parameter values giving a reasonable fit

to the data. With equal initial weighting towards A and B, the latter reduces to

$$\frac{p(A|D,I)}{p(B|D,I)} = \frac{\lambda_{\max} - \lambda_{\min}}{\mu_{\max} - \mu_{\min}}.$$

because the best-fit likelihood ratio will be unity (since $\lambda_0 = \mu_0$) and $\delta\lambda = \delta\mu$. Thus, our analysis will lead us to prefer the theorist who gives the narrower prior range; this is not unreasonable as he must have had some additional insight to be able to predict the value of the parameter more accurately.

2.2 Comparison with parameter estimation

The dependence of the result in Eq. (7) on the prior range ($\lambda_{\max} - \lambda_{\min}$) can seem a little strange, since we haven't encountered such behaviour in the preceding chapters. It is instructive, therefore, to compare the model selection analysis with parameter estimation. To infer the value of λ from the data, given that B's theory is correct, we use Bayes' theorem:

$$p(\lambda|D,B,I) = \frac{p(D|\lambda,B,I)p(\lambda|B,I)}{p(D|B,I)}. \tag{9}$$

The numerator is the familiar product of a prior and likelihood, and the denominator is usually omitted since it does not depend explicitly on λ; hence this relationship is often written as a proportionality. From the story of Dr A and Prof B, however, we find that the neglected term on the bottom plays a crucial role in ascertaining the merit of B's theory relative to a competing alternative. In recognition of its new-found importance, the denominator in Bayes' theorem is sometimes called the **'evidence'** for B; it is also referred to as the 'marginal likelihood', the 'global likelihood' and the 'prior predictive'. Since all the components necessary for both parameter estimation and model selection appear in Eq. (9), we are not dealing with any new principles; the only thing that sets them apart is that we are asking different questions of the data.

A simple way to think about the difference between parameter estimation and model selection is to note that, to a good approximation, the former requires the location of the maximum of the likelihood function whereas the latter entails the calculation of its average value. As long as $\lambda_{\min}$ and $\lambda_{\max}$ encompass the significant region of $p(D|\lambda,B,I)$ around λ_0, the precise bounds do not matter for estimating the optimal parameter and need not be specified. Since the prior range defines the domain over which the mean likelihood is computed, due thought is necessary when dealing with model selection. Indeed, it is precisely this act of comparing 'average' likelihoods rather than 'maximum' ones which introduces the desired Ockham balance to the goodness- of-fit criterion. Any likelihood gain from a better agreement with the data, allowed by the greater flexibility of a more complicated model, has to be weighed against the additional cost of averaging it over a larger parameter space.

3 Evidence calculations

The actual computation of Bayesian evidences can be a challenging task. Recall that we often have knowledge of the posterior distribution only through sampling. In many cases, the simple Laplace method can be used to compute the evidence approximately, while in other cases we have to rely on special sampling algorithms such as nested sampling or parallel tempering with thermodynamic integration.

3.1 Laplace's method

The idea behind the Laplace approximation is simple. We assume that an unnormalized probability density $P^*(\theta)$ has a peak at a point θ_0. We are interested in the evidence, Z_P, which is given by the normalizing constant

$$Z_P = \int P^*(\theta) d^K\theta,$$

where we consider the general case in which θ is in a K-dimensional space.

We Taylor-expand the logarithm $\log P^*$ around the peak:

$$\log P^*(\theta) = \log P^*(\theta_0) - \frac{1}{2}(\theta-\theta_0)^T\Sigma^{-1}(\theta-\theta_0) + \ldots,$$

where $\Sigma^{-1} = H$ is the (Hessian) matrix of second derivatives at the maximum

$$H_{ij} = -\left.\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log P^*(\theta)\right|_{\theta=\theta_0}.$$

We then approximate $P^*(\theta)$ by an unnormalized Gaussian,

$$Q^*(\theta) \equiv P^*(\theta_0)\exp\left[-\frac{1}{2}(\theta-\theta_0)^T\Sigma^{-1}(\theta-\theta_0)\right],$$

and we approximate the normalizing constant Z_P by the normalizing constant of this Gaussian,

$$Z_P \approx Z_Q = P^*(\theta_0)\sqrt{\frac{(2\pi)^K}{\det\Sigma^{-1}}}.$$

Predictions can then be made using this approximation Q. Physicists also call this widely-used approximation the saddle-point approximation.

Note, in particular, that if we consider a chi-squared pdf: $P^*(\theta) = \exp\left(-\frac{1}{2}\chi^2(\theta)\right)$, then we get

$$Z_P \approx \exp\left(-\frac{1}{2}\chi^2(\theta_0)\right)\sqrt{\frac{(4\pi)^K}{\det\Sigma^{-1}}},$$

where there is a factor $2^{K/2}$ that comes from the extra factor $1/2$ multiplying the covariance matrix Σ^{-1} and therefore appearing in all K eigenvalues.

The fact that the normalizing constant of a Gaussian is given by

$$\int d^K\theta\exp\left[-\frac{1}{2}\theta^T\Sigma^{-1}\theta\right] = \sqrt{\frac{(2\pi)^K}{\det\Sigma^{-1}}},$$

can be proved by making an orthogonal transformation into the basis u in which Σ is transformed into a diagonal matrix. The integral then separates into a product of one-dimensional integrals, each of the form

$$\int du_i \exp\left[-\frac{1}{2}\lambda_i u_i^2\right] = \sqrt{\frac{2\pi}{\lambda_i}}$$

The product of the eigenvalues λ_i is the determinant of Σ^{-1}.

Note that the Laplace approximation is basis-dependent: if θ is transformed to a nonlinear function $u(\theta)$ and the density is transformed to $P(u) = P(\theta)|d\theta/du|$ then in general the approximate normalizing constants Z_Q will be different. This can be viewed as a defect—since the true value Z_P is basis-independent in this approximation—or an opportunity, because we can hunt for a choice of basis in which the Laplace approximation is most accurate.