Learning from data: Error propagation and nuisance parameters

Christian Forssén

Department of Physics, Chalmers University of Technology, Sweden

Sep 21, 2020

1 Why Bayes is Better

Quotes from one pioneering and one renaissance Bayesian authority. Laplace:

> *Probability theory is nothing but common sense reduced to calculation."*

Sivia

> *Bayesian inference probabilities are a measure of our state of knowledge about nature, not a measure of nature itself."*

1.1 Advantages of the Bayesian approach

1. Provides an elegantly simple and rational approach for answering, in an optimal way, any scientific question for a given state of information. This contrasts to the recipe or cookbook approach of conventional statistical analysis. The procedure is well-defined:
   - Clearly state your question and prior information.
   - Apply the sum and product rules. The starting point is always Bayes' theorem.
2. For some problems, a Bayesian analysis may simply lead to a familiar statistic. Even in this situation it often provides a powerful new insight concerning the interpretation of the statistic.
3. Incorporates relevant prior (e.g., known signal model or known theory model expansion) information through Bayes' theorem. This is one of the great strengths of Bayesian analysis.

- For data with a small signal-to-noise ratio, a Bayesian analysis can frequently yield many orders of magnitude improvement in model parameter estimation, through the incorporation of relevant prior information about the signal model.

4. Provides a way of eliminating nuisance parameters through marginalization. For some problems, the marginalization can be performed analytically, permitting certain calculations to become computationally tractable.
5. Provides a way for incorporating the effects of systematic errors arising from both the measurement operation and theoretical model predictions.
6. Calculates probability of hypothesis directly: $p(H_i|D,I)$.
7. Provides a more powerful way of assessing competing theories at the forefront of science by automatically quantifying Occam's razor.

The Bayesian quantitative Occam's razor can also save a lot of time that might otherwise be spent chasing noise artifacts that masquerade as possible detections of real phenomena.

Occam's razor. Occam's razor is a principle attributed to the medieval philosopher William of Occam (or Ockham). The principle states that one should not make more assumptions than the minimum needed. It underlies all scientific modeling and theory building. It cautions us to choose from a set of otherwise equivalent models of a given phenomenon the simplest one. In any given model, Occam's razor helps us to shave offthose variables that are not really needed to explain the phenomenon. It was previously thought to be only a qualitative principle.

1.2 Nuisance parameters

Nuisance parameters (I): Bayesian Billiard. See demonstration notebook: A Bayesian Billiard game

Nuisance parameters (II): marginal distributions. Assume that we have a model with two parameters, θ_0, θ_1, although only one of them (say θ_1) is of physical relevance (the other one is them labeled a nuisance parameter). Through a Bayesian data analysis we have the joint, posterior pdf

$$p(\theta_0, \theta_1|D,I).$$

The marginal posterior pdf $p(\theta_1|D,I)$ is obtained via marginalization

$$p(\theta_1|D,I) = \int p(\theta_0, \theta_1|D,I) d\theta_0.$$

Assume that we have N samples from the joint pdf. This might be the Markov Chain from an MCMC sampler: $\{(\theta_0, \theta_1)_i\}_{i=0}^{N-1}$. Then the marginal distribution

Figur 1: Did the Leprechaun drink your wine, or is there a simpler explanation?

of θ_1 will be given by the same chain by simply ignoring the θ_0 column, i.e., $\{\theta_{1,i}\}_{i=0}^{N-1}$.

See the interactive demos created by Chi Feng for an illustration of this: The Markov-chain Monte Carlo Interactive Gallery.

1.3 Error propagation

Error propagation (I): marginalization. The Bayesian approach offers a straight-forward approach for dealing with (known) systematic uncertainties; namely by marginalization. Let us demonstrate this with an example

Inferring galactic distances with an imprecise knowledge of the Hubble constant The Hubble constant acts as a galactic ruler as it is used to measure astronomical distances according to $v = H_0 x$. An error in this ruler will therefore correspond to a systematic uncertainty in such measurements.

Here we use marginalization to obtain the desired posterior pdf $p(x|D,I)$ from the joint distribution of $p(x,H_0|D,I)$

$$p(x|D,I) = \int_{-\infty}^{\infty} dH_0 p(x,H_0|D,I). \tag{1}$$

Using Bayes' rule: $p(x,H_0|D,I) \propto p(D|x,H_0,I)p(x,H_0|I)$, the product rule: $p(x,H_0|I) = p(H_0|x,I)p(x|I)$, and the fact that H_0 is independent of x:

$p(H_0|x,I) = p(H_0|I)$, we find that

$$p(x|D,I) \propto p(x|I) \int dH_0 p(H_0|I) p(D|x,H_0,I),$$

which means that we have expressed the quantity that we want (the posterior of x) in terms of quantities that we know.

Assume that the pdf $p(H_0|I)$ is known via its N samples $\{H_i\}_{i=0}^{N-1}$ generated by the MCMC sampler.

This means that we can approximate

$$p(x|D,I) \propto \int dH_0 p(H_0|I) p(D|x,H_0,I) \approx \frac{1}{N} \sum_{i=1}^{N} p(D|x,H_i,I)$$

where we have used a uniform prior for the distance $p(x|I) \propto 1$.

Error propagation (II): changing variables and prior information. (Based on Sivia, ch 3.6.)

Assume that we have measured parameter $X = 10 \pm 3$ and $Y = 7 \pm 2$; what can we say about the difference $X - Y$ or the raio X/Y, or the sum of their squares $X^2 + Y^2$, etc? In essence, the problem is nothing more than an exercise in the change of variables: given the joint pdf $p(X,Y|I)$, where the information I might include the data if the pdf is a posterior from a data analysis, we need the corresponding pdf $p(Z|I)$, where $Z = X - Y$, or $Z = X/Y$, or whatever as appropriate.

Let us start with a single variable X and a function $Y = f(X)$. How is $p(X|I)$ related to $p(Y|I)$?

Consider a point X^* and a small interval δX around it. The probability that X lies within that interval can be written

$$p\left(X^* - \frac{\delta X}{2} \leq X < X^* + \frac{\delta X}{2} |I\right) \approx p(X = X^*|I)\delta X.$$

Assume now that the function f will map the point $X = X^*$ uniquely onto $Y = Y^* = f(X^*)$. Then there must be an interval δY around Y^* so that the probability is conserved

$$p(X = X^*|I)\delta X = p(Y = Y^*|I)\delta Y.$$

In the limit of infinitesimally small intervals, and with the realization that this should be true for any point X, we obtain the relationship

$$p(X|I) = p(Y = Y|I) \left| \frac{dY}{dX} \right|, \qquad (2)$$

where the term on the far right is called the *Jacobian*.

The generalization to several variables, relating the pdf for M variables $\{X_j\}$ in terms of the same number of quantities $\{Y_j\}$ related to them, is

$$p(\{X_j\}|I) = p(\{Y_j\}|I) \left| \frac{\partial(Y_1, Y_2, \ldots, Y_M)}{\partial(X_1, X_2, \ldots, X_M)} \right|, \tag{3}$$

where the multivariate Jacobian is given by the determinant of the $M \times M$ matrix of partial derivatives $\partial Y_i / \partial X_j$.

Summary

We have now seen the basic ingredients required for the propagation of errors: it either involves a transformation in the sense of Eq. (3) or an integration as in Eq. (1).

A useful short cut. For practical purposes, we are often satisfied to approximate pdfs with Gaussians. Within such limits there is an easier method that is often used for error propagation. Note, however, that there are instances when this method fails miserably as will be shown in the example further down.

Suppose that we have summarized the pdfs $p(X|I)$ and $p(Y|I)$ as two Gaussians with mean and standard deviation x_0, σ_x and y_0, σ_y, respectively. Assume further that these two variables are not correlated, i.e., $p(X, Y|I) = p(X|I)p(Y|I)$.

Suppose now that we are interested in $Z = X - Y$. Intuitively, we might guess that the best estimate $z_0 = x_0 - y_0$, but the error bar σ_z requires some more thought. Differentiate the relation

$$\delta Z = \delta X - \delta Y.$$

Square both sides and integrate to get the expectation value

$$\langle \delta Z^2 \rangle = \langle \delta X^2 + \delta Y^2 - 2\delta x \delta Y \rangle = \langle \delta X^2 \rangle + \langle \delta Y^2 \rangle - 2\langle \delta X \delta Y \rangle,$$

where we have employed the linear property for an integral over a sum of terms.

Since we assumed that the pdfs for X and Y were described by independent Gaussians we have

$$\langle \delta X^2 \rangle = \sigma_x^2; \qquad \langle \delta Y^2 \rangle = \sigma_y^2; \qquad \langle \delta X \delta Y \rangle = 0, \tag{4}$$

and we find that

$$\sigma_z = \sqrt{\langle \delta Z^2 \rangle} = \sqrt{\sigma_x^2 + \sigma_y^2}.$$

Consider, as a second example, the ratio of two parameters $Z = X/Y$. Differentiation gives

$$\delta Z = \frac{Y \delta X - X \delta Y}{Y^2} \quad \Leftrightarrow \quad \frac{\delta Z}{Z} = \frac{\delta X}{X} - \frac{\delta Y}{Y}.$$

Squaring both sides and taking the expectation values, we obtain

$$\frac{\langle \delta Z^2 \rangle}{z_0^2} = \frac{\langle \delta X^2 \rangle}{x_0^2} + \frac{\langle \delta Y^2 \rangle}{y_0^2} - 2\frac{\langle \delta X \rangle \langle \delta ZY \rangle}{x_0 y_0},$$

where the X, Y and Z in the denominator have been replaced by the constants x_0, y_0 and $z_0 = x_0/y_0$ because we are interested in deviations from the peak of the pdf.

Finally, substituting the information for the pdfs of X and Y as summarized in Eq. (4) we finally obtain the propagated error for the ratio

$$\frac{\sigma_z}{z_0} = \sqrt{\left(\frac{\sigma_x}{x_0}\right)^2 + \left(\frac{\sigma_y}{y_0}\right)^2}.$$

Despite its virtues, let us end our discussion of error-propagation with a salutary warning against the blind use of this nifty short cut.

Example: Taking the square root of a number. (Example 3.6.2 in Sivia)

- Assume that the amplitude of a Bragg peak is measured with an uncertainty $A = A_0 \pm \sigma_A$ from a least-squares fit to experimental data.
- The Bragg peak amplitude is proportional to the square of a complex structure function: $A = |F|^2 \equiv f^2$.
- What is $f = f_0 \pm \sigma_f$?

Obviously, we have that $f_0 = \sqrt{A_0}$. Differentiate the relation, square and take the expectation value

$$\langle \delta A^2 \rangle = 4 f_0^2 \langle \delta f^2 \rangle \quad \Leftrightarrow \quad \sigma_f = \frac{\sigma_A}{2\sqrt{A_0}},$$

where we have used the Gaussian approximation for the pdfs.

But what happens if the best fit gives $A_0 < 0$, which would not be impossible if we have weak and strongly overlapping peaks. The above equation obviously does not work since f_0 would be a complex number.

We have made two mistakes:

1. Likelihood is not posterior!
2. The Gaussian approximation around the peak does not always work.

Consider first the best fit of the signal peak. It implies that the likelihood can be approximated by

$$p(D|A,I) \propto \exp\left[-\frac{(A-A_0)^2}{2\sigma_A^2}\right].$$

However, the posterior for A is $p(A|D,I) \propto p(D|A,I)p(A|I)$ and we should use the fact that we know that $A \geq 0$.

We will incorporate this information through a simple step-function prior

$$p(A|I) = \left\{ \begin{array}{ll} \frac{1}{A_{\mathrm{max}}}, & 0 \leq A \leq A_{\mathrm{max}}, \\ 0, & \mathrm{otherwise}. \end{array} \right.$$

This implies that the posterior will be a truncated Gaussian, and its maximum will always be above zero.

This also implies that we cannot use the Gaussian approximation. Instead we will do the proper calculation using the transformation (2)

$$p(f|D,I) = p(A|D,I) \left| \frac{dA}{df} \right| = 2fp(A|D,I)$$

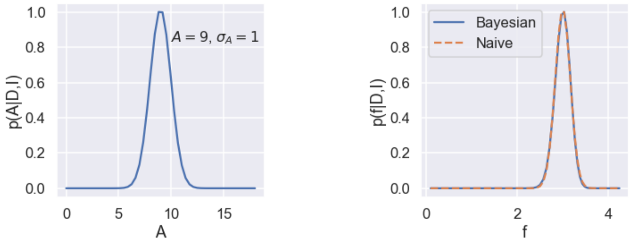
In the end we find the proper Bayesian error propagation given by the pdf

$$p(f|D,I) \propto \left\{ \begin{array}{ll} f \exp \left[-\frac{(A-A_0)^2}{2\sigma_A^2} \right], & 0 \leq f \leq \sqrt{A_{\mathrm{max}}}, \\ 0, & \mathrm{otherwise}. \end{array} \right.$$

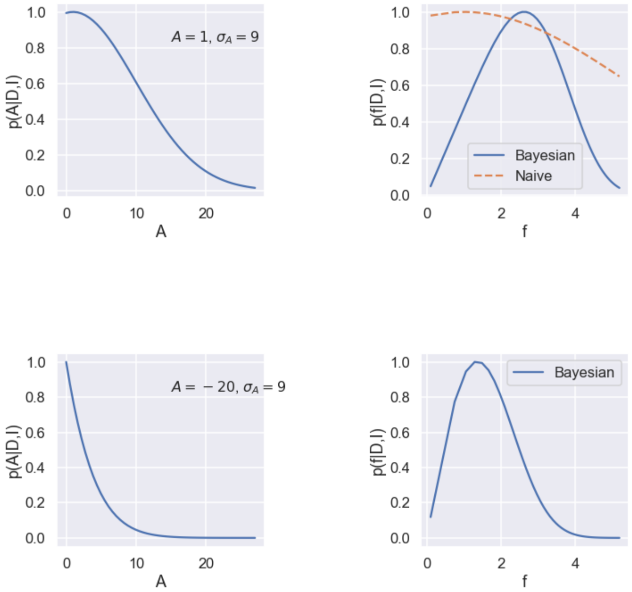
Let us visualize the difference between the Bayesian and the naive error propagation for a few scenarios.

```
def A_posterior(A,A0,sigA):
    pA = np.exp(-(A-A0)**2/(2*sigA**2))
    return pA/np.max(pA)

# Wrong analysis
def f_likelihood(f,A0,sigA):
    sigf = sigA / (2*np.sqrt(A0))
    pf = np.exp(-(f-np.sqrt(A0))**2/(2*sigf**2))
    return pf/np.max(pf)

# Correct error propagation
def f_posterior(f,A0,sigA):
    pf = f*np.exp(-(f**2-A0)**2/(2*sigA**2))
    return pf/np.max(pf)
```

```
for (A0,sigA) in [(9,1),(1,9),(-20,9)]:
    maxA = max(2*A0,3*sigA)
    A_arr = np.linspace(0.01,maxA)
    f_arr = np.sqrt(A_arr)
    fig,ax=plt.subplots(1,2,figsize=(10,4))
    ax[0].plot(A_arr,A_posterior(A_arr,A0,sigA))
    ax[1].plot(f_arr,f_posterior(f_arr,A0,sigA),label=’Bayesian’)
    if A0>0:
        ax[1].plot(f_arr,f_likelihood(f_arr,A0,sigA),’--’,label=’Naive’)
    ax[0].set(xlabel=’A’,ylabel=’p(A|D,I)’)
    plt.text(0.55,0.8,f’$A={A0}$, $\sigma_A={sigA}$’, transform=ax[0].transAxes,fontsize=16)
    ax[1].set(xlabel='f',ylabel='p(f|D,I)')
    ax[1].legend(loc='best')
    plt.tight_layout()
```



Figur 2: The left-hand panels show the posterior pdf for the amplitude of a Bragg peak in three different scenarios. The right-hand plots are the corresponding pdfs for the modulus of the structure factor $f = \sqrt{A}$. The solid lines correspond to a full bayesian error propagation, while the dashed lines are obtained with the short-cut error propagation.


1.0
0.8
0.6
0.4
0.2
0.0
p(A|D,I)
0
10
20
A
$A = 1, \sigma_A = 9$
1.0
0.8
0.6
0.4
0.2
0.0
p(f|D,I)
0
2
4
f
Bayesian
Naive
1.0
0.8
0.6
0.4
0.2
0.0
p(A|D,I)
0
10
20
A
$A = -20, \sigma_A = 9$
1.0
0.8
0.6
0.4
0.2
0.0
p(f|D,I)
0
2
4
f
Bayesian