Bayesian Framework

The Bayesian framework is a systematic way of incorporating data into prior knowledge/hypotheses to get new hypotheses. It can be derived from the basic formula of the product of probability.

$$\mathcal{P}(\mathcal{E})\mathcal{P}(\mathcal{H}|\mathcal{E}) = \mathcal{P}(\mathcal{E}|\mathcal{H})\mathcal{P}(\mathcal{H}) \tag{1}$$

Where the left side is read as the probability of $\mathcal{E}$ time probability of $\mathcal{H}$ given $\mathcal{E}$. By rearranging these we get the Bayes theorem.

$$\mathcal{P}(\mathcal{H}|\mathcal{E}) = \frac{\mathcal{P}(\mathcal{E}|\mathcal{H})\mathcal{P}(\mathcal{H})}{\mathcal{P}(\mathcal{E})} \tag{2}$$

The following names of these terms are often used in literature.

$$Posterior = \frac{Likelihood \ \times \ Prior}{Margeninal \, Likelihood} \tag{3}$$

Let's talk about these terms one by one. A more general form of Bayes theorem is

$$\mathcal{P}(\mathcal{H}_i|\mathcal{E}) = \frac{\mathcal{P}(\mathcal{E}|\mathcal{H}_i)\mathcal{P}(\mathcal{H}_i)}{\mathcal{P}(\mathcal{E})} \tag{4}$$

Where $\mathcal{H}$ stands for Hypothesis and i suggests there are many alternative hypotheses to be considered. $\mathcal{P}(\mathcal{H}_i)$ is number assigned to i_{th} hypothesis giving expected outcome of it. Or in the short prior probability of hypothesis or just Prior. The sum of all the Priors must be equal to one. The likelihood is usually calculated to tell how good the hypothesis is fit with the data. It is typically calculated from the residue of curve fitting. Hence read as the probability of the evidence-given hypothesis. While the Posterior or probability of a hypothesis gave evidence that quantifies the updated belief in the hypothesis after the evidence has been considered.

The Demonominator on the right side plays the role of the normalisation constant and can be calculated by the marginalisation rule of probability.

$$\mathcal{P}(\mathcal{X}) = \sum_y \mathcal{P}(\mathcal{X}, \mathcal{Y} = y) = \sum_y \mathcal{P}(\mathcal{X}|\mathcal{Y} = y)\mathcal{P}(\mathcal{Y})$$

(5)

where variable $\mathcal{Y}$ have discrete values at y and for continuous variable summation is replaced by integral.

$$\mathcal{P}(\mathcal{X}) = \int dy\, \mathcal{P}(\mathcal{X}, \mathbf{Y}) = \int dy\, \mathcal{P}(\mathcal{X}|\mathbf{Y})\mathcal{P}(\mathcal{Y}) \tag{6}$$

A more convenient practice is to use the variable D acronym for data in place of $\mathcal{E}$, which is more convenient for numerical calculation. Throughout the text, the following notation will be used $\mathcal{P}(D|a, b, c, I)$ where a, b and c are parameters of the model and character "I" is reserved for non-mentioned parameters. It's just for convenience and can be ignored if there are a few parameters in the model.

Bayesian Inference

Bayes theorem can be used to infer best-fit parameters of the model. following is the algorithm to make this inference:

- Choose a model with a set of parameters.
- Give prior value to these parameters.
- Choose likelihood function.
- Calculate Best fit parameters using Bayes theorem

This algorithm not only gives best-fit parameters but also gives the probability distribution of parameters.

Marginalization

Any parameter in the probability formula can be eliminated by integrating over all of its values. This method of integrating all the values is historically called Marginalization. A generalized formula for marginalization is:

$$\mathcal{P}(D|M, I) = \int da \int db \int dc \int \ldots\ \mathcal{P}(a|M, I)\mathcal{P}(b|M, I)\mathcal{P}(c|M, I)\mathcal{P}(D|a, b, c, \ldots M, I) \tag{7}$$

Where M is the Model with Parameters a,b,c...

Nuisance parameters

Bayesian framework allows us to infer one parameter without explicitly finding other parameters as they can be marginalised. For example in (7) where the probability distribution of D can be calculated by averaging over a,b and c. Which seems trivial but is the feature generally missing in frequentist statistics. This fact makes calculations easy, hence the variables which are integrated over are called nuisance parameters.

Model Comparison

Once the joint posterior distribution of models is estimated with all the parameters of the model explicitly, then any model comparison can be employed. But It is best to use Bayesian hierarchal structures for model comparison. We can compare models without explicitly finding parameters more details of hierarchical structures of the Bayesian framework will be discussed in a subsequent chapter for model comparison.

Gregory-Lorendo Methods

GL method is a Bayesian way of inferring the period and shape of the light curve when the noise sampling distribution is independent Gaussian and data is non-uniformly sampled. Following the general rule, GL methods have four steps introduced in the previous chapter.

Model

The model in the GL method is a stepwise model. Defined by

$$j(t) = int[1 + m(wt + \phi) mod\ 2\pi/2\pi] \tag{8}$$

This function will split the timeline into different bins. Repeating after time $\omega/2\pi$. Each Step can take an arbitrary height. Giving model $m + 2$ parameters to work with. The value of each bin can be called r_j with j varying from 1 to m. Here $m = 1$ signifies there is no periodic motion and data is independent noise around the constant. higher m means the periodic model.

Priors

After choosing a model with m steps, we are left with m + 2 parameters in the model. Priors values of these are to be specified, Folling priors are used in the original Gregory-Lorendo Method.

$$\mathcal{P}(\phi|M_m, I) = \frac{1}{2\pi} \tag{9}$$

$$\mathcal{P}(w|M_m, I) = \frac{1}{w \ln \frac{w_{hi}}{w_{lo}}} \tag{10}$$

$$\mathcal{P}(r_j|M_m, I) = \frac{1}{\Delta r} = \frac{1}{r_{max} - r_{min}} \tag{11}$$

where M_m says model (large M) with a number of bins (small m) is considered. These Priors come with a range from low to high, Which is another important consideration in the Bayesian framework and caution which suggest that we are biased to finding a solution in a specific range only.

Likelihood

Datapoints are denoted as $d_i(t)$ where i range from 1 to N total number of data points. The difference between specific data point d_i and r_j the bin it falls into is assumed to be Gaussian distribution with variance σ_i^2.

Given the w frequency, ϕ phase, m number of bins, and $\vec{r}$ set of m values of r. We can find the Likelihood as:

$$p(D|w,\phi,\vec{r},m,I) = \prod_{j=1}^{m}\left[(2\pi)^{-\frac{n_j}{2}}\left(\prod_{i=1}^{n_j}(s_i)^{-1}\right)exp\left(\frac{-\alpha}{2}\right)\right] \tag{12}$$

where n_j is number of samples in j_{th} bin. and s_i is expected noise in i_{th} data point.

$$\alpha = \sum_{i=1}^{n_j}\frac{(d_i - r_j)^2}{s_i^2} \tag{13}$$

Notice It's a product of residue from all the bins.

0.1 Calculation

Although we get all the components of the Bayes theorem to infer best-fit parameters, calculations are not practical at this point as we have to calculate m+2 variable convoluted integrals (7). Our Strategy is to integrate over all the m + 2 parameters and use analytical integrals whenever possible.

Isolating $\vec{r}$ term

To make the calculation economical first isolate r from the likelihood.

$$\alpha = \sum_{i=i}^{n_j}\frac{d_i^2}{s_i^2} - 2r_j\sum_{i=i}^{n_j}\frac{d_i^2}{s_i^2} + r_j^2\sum_{i=i}^{n_j}\frac{1}{s_i^2}$$

Let's call

$$\sum_{i=1}^{n_j}\frac{1}{s_i^2} = W_j$$

and introduce the first signature

$$\frac{\sum_{i=1}^{n_j}\frac{d_i}{s_i^2}}{W_j} = \vec{d_{W_j}}$$

and second signature

$$\frac{\sum_{i=1}^{n_j}\frac{d_i^2}{s_i^2}}{W_j} = \vec{d_{W_j}^2}$$

α can be simplified to

$$\alpha = W_j(\vec{d_{W_j}^2} - 2r_j\vec{d_{W_j}} + r_j^2)$$

To separate out r we can add and subtract $\vec{d_{W_j}}^{\,2}$ giving

$$\alpha = W_j \left[(r_j^2 - 2r_j \vec{d_{W_j}} + \vec{d_{W_j}}^{\,2}) + \vec{d_{W_j}^2} - \vec{d_{W_j}}^{\,2}) \right] \tag{14}$$

Where we can separate out $\chi^2 = \vec{d_{W_j}^2} - \vec{d_{W_j}}^{\,2}$ and complete the square of the extra term.

$$\alpha = W_j \left[(r_j - \vec{d_{W_j}})^2 + \chi^2 \right] \tag{15}$$

Here we have separated out r_j as a single variable to exponential where every other part can be treated as constant and we can perform analytical integral. In short, our likelihood is in form.

$$\mathcal{L} = e^{-r^2} \tag{16}$$

Marginalizing over $\vec{r}$

Using generalised tricks to marginalise:

$$\mathcal{P}(D|M_m, I) = \int dw \int d\phi \int d\vec{r}\ \ \mathcal{P}(\phi|M_m, I)\mathcal{P}(w|M_m, I)\mathcal{P}(\vec{r}|M_m, I)\mathcal{P}(D|w, \phi, \vec{r}, M_m, I)$$

Marginalizing $\vec{r}$ first.

$$\mathcal{P}(D|w, \phi, M_m, I) = \int d\vec{r}\ \ \mathcal{P}(\vec{r}|M_m, I)\mathcal{P}(D|\vec{r}, w, \phi, M_m, I)$$

The first term on the right-hand side is Prior and the second term we have already simplified.

$$\mathcal{P}(D|w, \phi, M_m, I) = \prod_{j=1}^{m} \left[(2\pi)^{-\frac{n_j}{2}} \left(\prod_{i=1}^{n_j} (s_i)^{-1} \right) exp\left(-\frac{\chi^2 W_j}{2} \right) R \right]$$

where R is

$$R = \int_{r_{min}}^{r_{max}} dr_j\ \mathcal{P}(r_j|M_m, I) exp\left(-\frac{Wj(r_j - \vec{d_{W_j}})^2}{2} \right)$$

see R is inside $\prod$ that's why $\vec{r}$ turns to r_j Now expanding along the product.

$$= (2\pi)^{-N/2}(\Delta r)^{-m} \left(\prod_{i=1}^{N} (s_i)^{-1} \right) exp\left(-\sum_{j=1}^{m} \frac{\chi^2 W_j}{2} \right) \times \prod_{j=1}^{m} \left[\int_{r_{min}}^{r_{max}} dr_j\ \ exp\left(-\frac{W_j(r_j - \vec{d_{W_j}})^2}{2} \right) \right]$$

Here terms that are not depending on j turn from the product (zero to n_j) to (zero to N). Prior to r used is $1/\Delta r$ and the product turns to sum in exponential.

Integrand is sufficiently isolated and we can perform analytical integral now.

***Complementary error function**

This integral is well-behaved in following limits and depends only on the limit of the integral.

$$erfc(y) = \frac{2}{\sqrt{\pi}} \int_{y}^{\inf} du \;\; exp(-u^2) \tag{17}$$

In above equation If we take $u^2 = W_j(r_j - \vec{d_{W_j}})^2/2$ we can Simplify:

$$\int_{r_{min}}^{r_{max}} dr_j \;\; exp\left(-\frac{W_j(r_j - \vec{d_{W_j}})^2}{2}\right) = \sqrt{\frac{\pi}{w}} W_j^{-\frac{1}{2}} \left[erfc(y_{jmin}) - erfc(y_{jmax})\right] \tag{18}$$

where

$$y_{jmin} = \sqrt{\frac{W_j}{2}}(r_{min} - \vec{d_{w_j}}) \;\; ; \;\; y_{jmax} = \sqrt{\frac{W_j}{2}}(r_{max} - \vec{d_{w_j}})$$

After performing the integral we are left with

$$\mathcal{P}(D|w, \phi, M_m, I) =$$

$$(2\pi)^{-N/2}(\Delta r)^m \left(\prod_{i=1}^{N}(s_i)^{-1}\right) \exp\left(-\sum_{j=1}^{m}\frac{\chi^2 W_j}{2}\right)(\pi/2)^{m/2} \times \prod_{j=1}^{m}(W_j^{1/2}\left[erfc(y_{jmin}) - erfc(y_{jmax})\right])$$

where $\sqrt{\pi/2}$ comes out of the product and gets a power of m. One advantage of the GL method is that we never have to perform integral over r again.

Marginalisation over m and ϕ

Again using the marginalisation trick

$$\mathcal{P}(D|M_m, I) = \int dw \int d\phi \; \mathcal{P}(D|w, \phi, M_m, I)\mathcal{P}(w|M_m, I)\mathcal{P}(\phi|M_m, I)$$

Substituting the priors specified earlier in the equation we get:

$$\mathcal{P}(D|M_m, I) =$$

$$\frac{(2\pi)^{-N/2}(\Delta r)^m \left(\prod_{i=1}^{N}(s_i)^{-1}\right)(\pi/2)^{m/2}}{2\pi \ln \frac{w_{hi}}{w_{lo}}}$$

$$\times \int \frac{dw}{w} \int d\phi \exp\left(-\sum_{j=1}^{m}\frac{\chi^2 W_j}{2}\right) \prod_{j=1}^{m}(W_j^{1/2}\left[erfc(y_{jmin}) - erfc(y_{jmax})\right]) \tag{19}$$

We will keep track of this Integral as it is key in GL methods and will be used multiple times.

Calculation of $\mathcal{P}(D|M_m, I)$ Probability of D given Model

A word about formula (19). It's called the probability of data given model. It's quite a powerful formula as it can quantify the likelihood of the model itself. Do note that it gives the likelihood of a model with m bins and there can be different models with different numbers of bins. This formula can quantify the relative probability of which model is more likely to fit the given data/evidence.

The integral in this formula is quite convoluted because wherever there is j dependency it means dependency of w and ϕ and m. revising $j(t) = int[1 + m(wt + \phi) mod 2\pi/2\pi]$. So we have done m integrals anaclitic but the remaining integrals are challenging and computationally expensive and one of the limitations of GL methods and Bayesian inference in general. Conceptually we can find all the posterior distribution of parameters once we do these integrals.

0.2 Parameter Estimation

The important thing to note is that we haven't explicitly found any parameters of models and can still say a lot about the model and compare them. This feature is unique to Bayesian inference and there is no general way to do it in frequentist statistics. Now we can explicitly find the posterior distribution of all the parameters of the model using the Bayes theorem. see the general formula:

$$\mathcal{P}(a|D, M, I) = \frac{\mathcal{P}(a|M, I) \times \mathcal{P}(D|a, M, I)}{\mathcal{P}(D|M, I)}$$

Here left-hand side is the posterior distribution of a. On the left-hand side, the first term is prior and the second is the probability of D marginalised not over a. Note the denominator plays the role of normalisation constant and can be ignored for comparison purposes. So effectively finding the posterior distribution of the parameter is more simpler as we have to do one less integral. A more practical formula for the posterior distribution of any parameter is.

$$Posterior(a) = Prior(a) \times \int db \int dc \int ...Prior(b) \times Prior(c) \times ...\mathcal{P}(D|b, c...)$$

Which is of course not normalised. Summarising the algorithm now. we need to perform integral to find the posterior distribution of Models and parameters.

- For Model integral is over all parameters
- For parameter integral is over all parameters except parameter itself

Coming back to the GL method, we have m + 2 parameters. We have strategically done m integrals over r so we are left with two integrals m and ϕ.

Estimation of frequency

The posterior of the frequency probability distribution can be calculated from:

$$\mathcal{P}(w|D, M_m, I) = \mathcal{P}(w|M_m, I)\frac{\mathcal{P}(D|w, M_m, I)}{\mathcal{P}(D|M_m, I)}$$

As denominator is just a normalisation constant. We will call it C. The first term is prior. And the Second term is (19) with integral not over w.

$$\mathcal{P}(w|D, M_m, I) = \frac{C}{w}\int_0^{2\pi} d\phi\ \exp\left(-\sum_{j=1}^{m}\frac{\chi^2 W_j}{2}\right)\prod_{j=1}^{m}(W_j^{1/2}\left[erfc(y_{jmin}) - erfc(y_{jmax})\right])$$

And w in the denominator comes out from prior, all are constant are consumed in C. This formula tells the probability of frequency w. We can perform this integral over a range of w to get a probability distribution of w for a given GL Model.

Averaging Over Different GL-models

As different GL models with different numbers of bins can satisfy the data so naturally we should find the optimal model and use it for calculation. But in the Bayesian framework, we can average over all the models to get an even better posterior distribution of parameters.

$$\mathcal{P}(w|m > 1, D, I) = \sum_{m=2}^{m_{max}} \mathcal{P}(M_m|D, I)\mathcal{P}(w|D, M_m, I)$$

The last term in the equation is the Posterior distribution calculated from one model. And the First term is the probability of the Model. Which we can calculate from Bayes theorem.

$$\mathcal{P}(M_m|D, I) = \frac{\mathcal{P}(D|M_m, I)}{\sum_{m=2}^{m_{max}}\mathcal{P}(D|M_m, I)}$$

We now have all the pieces required to do the calculation. Discussion and methods of calculation are discussed in the next chapter. The posterior distribution of ϕ doesn't give any insight so we don't need it, and we will not calculate it. It's remarkable we can skip this calculation in Bayesian inference.

Estimation of Curve shape

We can use a similar strategy to get the shape of the periodic light curve by calculating the posterior distribution of $\vec{r}$ and their averaging over models to get a smooth light curve shape with confidence interval included as Bayesian inference give probability distribution. But that calculation is multiple dimensional and we will do it in later sections.