Tutorial 5: Fitting Black Hole - Galaxy Relations using Bayesian Inference

Name: Felix Martinez
Partner: Mahir Pirmohammed
Date: 5/4/21

1 Overview

This tutorial is an introduction to performing a fit using the nested sampling code “Dynesty" to determine posteriors and evidences. Furthermore, we will also go over how to conduct a comparison between models of varying complexity, and how to select the most appropriate model given a particular data set.

This will all be done with the fitting of black hole (BH) demographics that we can use with measurements made over the past $\sim$ 25 years. From these measurements, we should hopefully be able to determine a clear correlation between black hole mass ($M_{\rm BH}$) and the large-scale galaxy bulge properties, including the stellar velocity dispersion ($\sigma_\star$), luminosity ($L_{\rm bulge}$), and mass ($M_{\rm bulge}$).

Throughout this tutorial, we will be using terms like posteriors, likelihoods, priors, and evidences. For a set of data, d, and a model, $\mathcal{M}$, and any relevant information assumed to be true, I, Bayes' theorem states,

$$p(\mathcal{M}|d,I) = \frac{p(d|\mathcal{M},I)p(\mathcal{M}|I)}{p(d|I)} \tag{1}$$

where $p(\mathcal{M}|d,I)$ is the posterior probability of the model taking the data into account. it is proportional to $p(d|\mathcal{M},I)$, the sampling distribution of the data assuming the model is true, times the prior probability for the model, $p(\mathcal{M}|I)$. The prior represents our state of knowledge before seeing the data. The sampling distribution encodes how the degree of plausibility of the model changes when we acquire new data. When considered as a function of the model, for fixed data, it is called the likelihood function and we'll use the shortcut notation $\mathcal{L} = p(\mathcal{M}|d,I)$. The denominator on the right side of the equation, $p(d|I)$, is the evidence and is a normalization term. It is equal to the likelihood times the prior, summed over all possible models:

$$p(d|I) \equiv \sum_{\mathcal{M}} p(d|\mathcal{M},I)p(\mathcal{M}|I) \tag{2}$$

The posterior is the central quantity for model inference, while the evidence is the relevant quantity for model comparison.

2 Results

Below we summarize our work for each section of the tutorial and address the questions posed in the manual.

2.1 Data

Using the compilation of BH masses and host galaxy properties from Saglia et al. 2016, we can find relationships between $M_{\rm BH} - \sigma_\star$ and $M_{\rm BH} - M_{\rm bul}$. The next page holds the results we found.

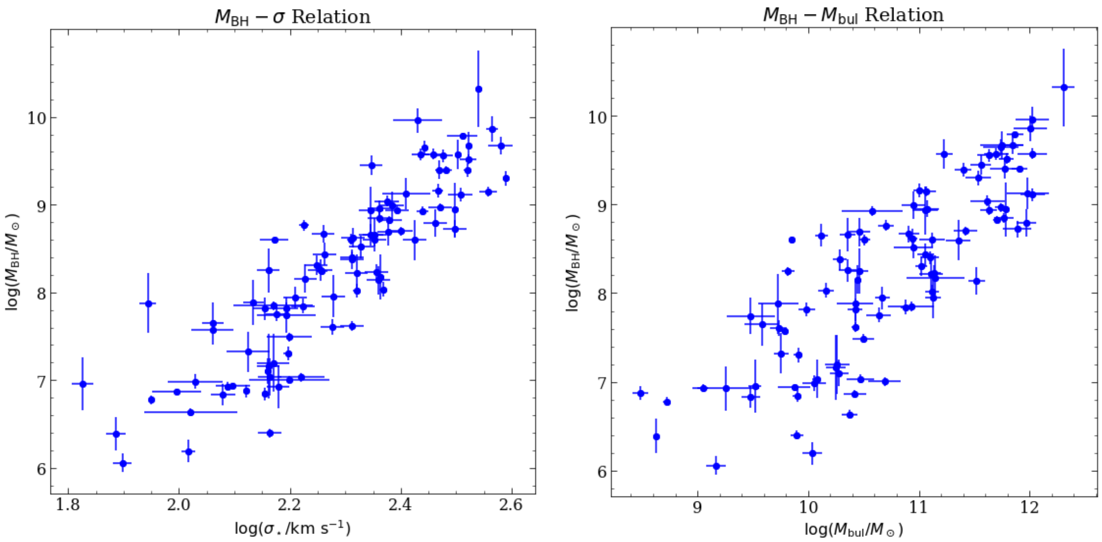


Figure 1: The left plot shows our $M_{\rm BH} - \sigma_\star$ relationship, the right plot shows our $M_{\rm BH} - M_{\rm bul}$ relationship.

As can be seen by the results found from the data taken from Saglia et al. 2016 in Figure 1, an increase in $\log_{10}(M_{\rm BH}/M_\odot)$ results in an increase in both $\log_{10}(\sigma_\star/\text{km s}^{-1})$ and $\log_{10}(M_{\rm bul}/M_\odot)$.

2.2 Parameter Estimation Using Nested Sampling

Dynesty is a public, open-source Python package used to estimate Bayesian posteriors and evidences via dynamic nested sampling. To use Dynesty, we will be defining two functions in addition to writing our code. One function sets the priors and the other function evaluates the natural logarithm of the likelihood [ln($\mathcal{L}$); referred to as the log-likelihood].

Following the recommendations in the Dynesty documentation, we will examine 2σ confidence intervals. We will also be fitting our prior function to three parameters: the slope of the line, the intercept of the line, and the intrinsic scatter. By including the intrinsic scatter as a model parameter, we are saying that even if we had measured all BH masses perfectly there would be some scatter about the line, which is physical.

During the fit, we will also account for uncertainties on both $\log_{10}(M_{\rm BH}/M_\odot)$ and $\log_{10}(\sigma_\star/\text{km s}^{-1})$. The function we will be using to define our likelihood is:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2 + \alpha^2\epsilon_{x,i}^2)}} \exp\left[\frac{-(y_i - (\alpha x_i + \beta))^2}{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2 + \alpha^2\epsilon_{x,i}^2)}\right] \tag{3}$$

where the log-likelihood function is,

$$\ln(\mathcal{L}) = -\frac{1}{2}X^2 + A \tag{4}$$

where,

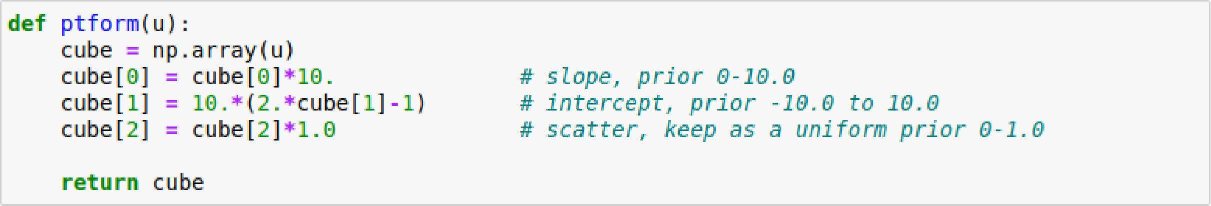
$$X^2 = \sum_{i=1}^{N} \frac{(y_i - (\alpha x_i + \beta))^2}{(\epsilon_{y,i}^2 + \epsilon_0^2 + \alpha^2\epsilon_{x,i}^2)} \tag{5}$$

and,

$$A = \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2 + \alpha^2\epsilon_{x,i}^2)}}\right]. \tag{6}$$

Where, x is $\log_{10}(\sigma_\star/\text{km s}^{-1})$ and then $\log_{10}(M_{\text{bul}}/M_\odot)$, y is always going to be our $\log_{10}(M_{\text{BH}}/M_\odot)$, ϵ_x is the uncertainty on $\log_{10}(\sigma_\star/\text{km s}^{-1})$ and then $\log_{10}(M_{\text{bul}}/M_\odot)$, ϵ_y is the uncertainty on $\log_{10}(M_{\text{BH}}/M_\odot)$, α is the slope of the line, β is the intercept of the line, and ϵ_0 is the intrinsic scatter.

Below is the function we wrote in python to define our priors.

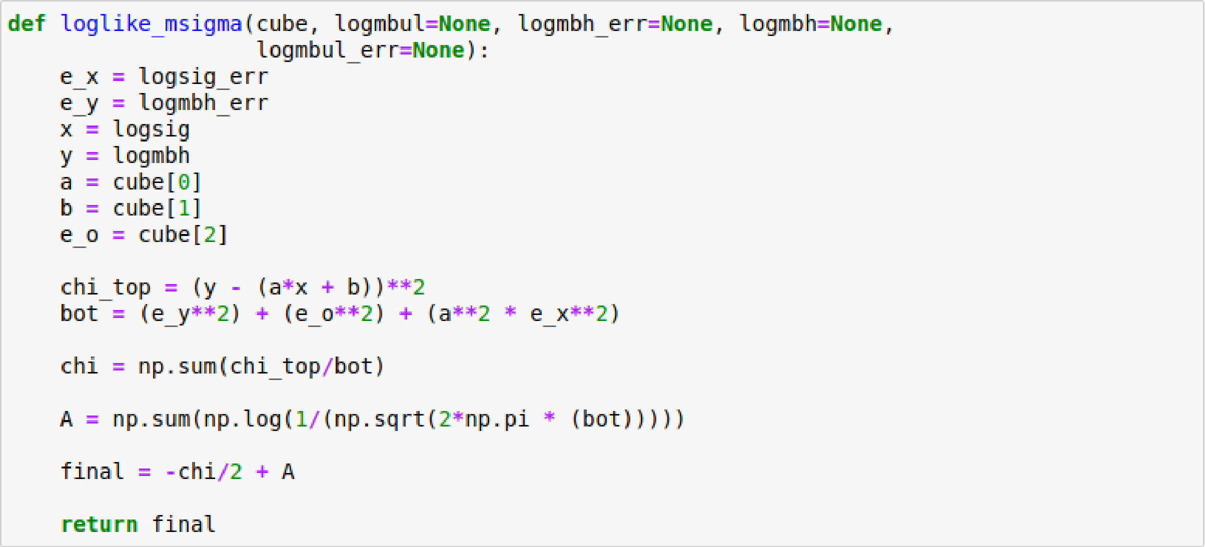


```
def ptform(u):
    cube = np.array(u)
    cube[0] = cube[0]*10.              # slope, prior 0-10.0
    cube[1] = 10.*(2.*cube[1]-1)       # intercept, prior -10.0 to 10.0
    cube[2] = cube[2]*1.0              # scatter, keep as a uniform prior 0-1.0

    return cube
```

Figure 2: Our code showing us defining our prior and its parameters.

We also include our log-likelihood function we wrote in python below.



```
def loglike_msigma(cube, logmbul=None, logmbh_err=None, logmbh=None,
                   logmbul_err=None):
    e_x = logsig_err
    e_y = logmbh_err
    x = logsig
    y = logmbh
    a = cube[0]
    b = cube[1]
    e_o = cube[2]

    chi_top = (y - (a*x + b))**2
    bot = (e_y**2) + (e_o**2) + (a**2 * e_x**2)

    chi = np.sum(chi_top/bot)

    A = np.sum(np.log(1/(np.sqrt(2*np.pi * (bot)))))

    final = -chi/2 + A

    return final
```

Figure 3: Our code showing us defining our log-likelihood function using equations (4), (5), and (6).

When using the data found in Saglia et. al. 2016, with our prior and log-likelihood functions, we were able to get the following corner plots for the $M_{\text{BH}} - \sigma_\star$ and $M_{\text{BH}} - M_{\text{bul}}$ relationships. They are show on the next page.

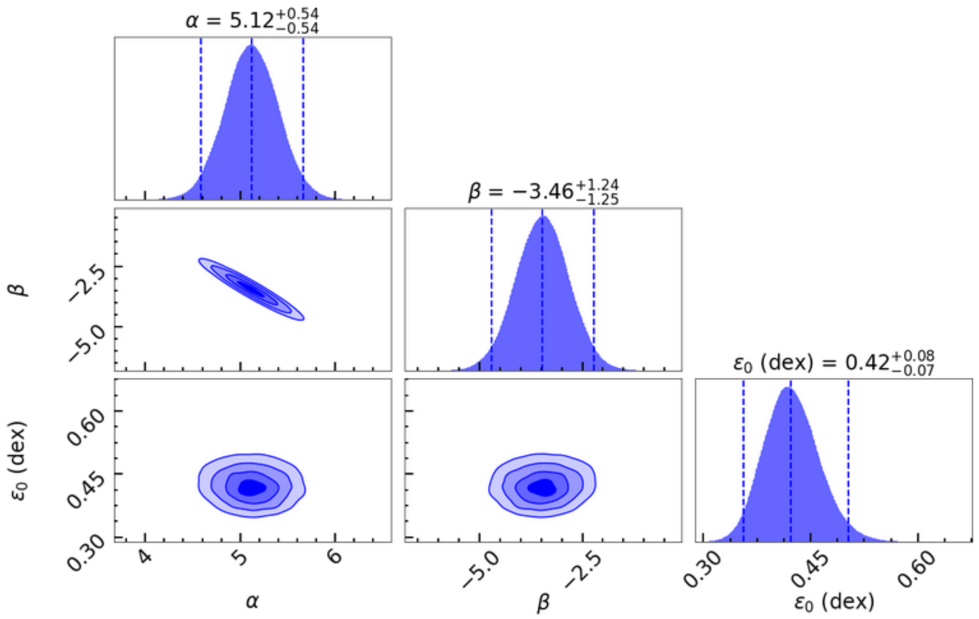


Figure 4: Our corner plot for our $M_{\rm BH} - \sigma_\star$ relationship, we also show the median values, and the 2σ uncertainties for our posteriors.

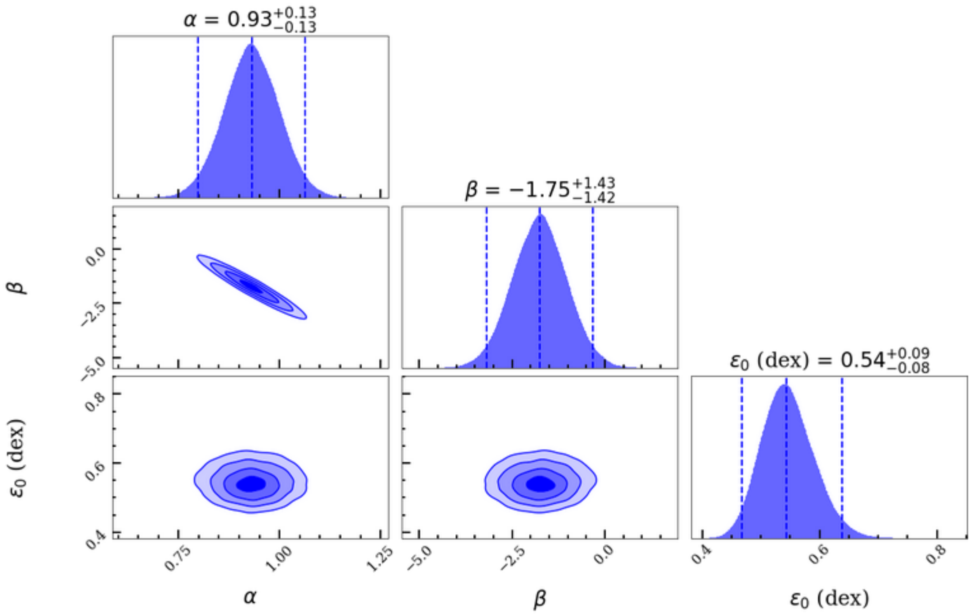


Figure 5: Our corner plot for our $M_{\rm BH} - M_{\rm bul}$ relationship, we also show the median values, and the 2σ uncertainties for our posteriors.

Finally, we also show our $M_{\rm BH}-\sigma_\star$ and $M_{\rm BH}-M_{\rm bul}$ plots along with their best-fit lines to indicate their intrinsic scatter below.

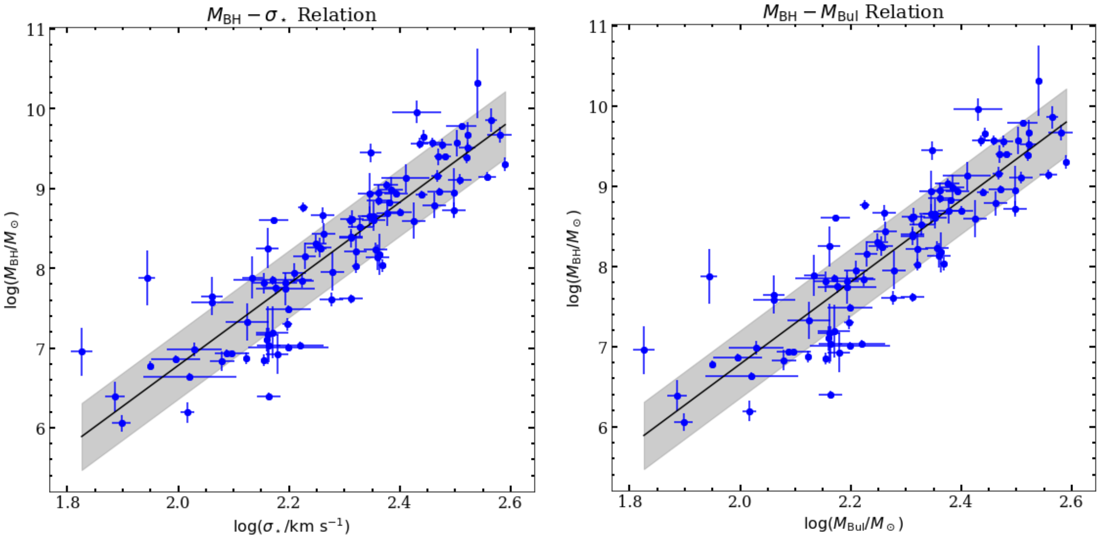


Figure 6: The left plot shows our $M_{\rm BH}-\sigma_\star$ relationship, the right plot shows our $M_{\rm BH}-M_{\rm bul}$ relationship.

As seen above, the $\sigma_\star$ and $M_{\rm bul}$ galaxy properties are related to each other, for example galaxies with a larger $M_{\rm bul}$ tend to have a larger $L_{\rm bul}$ and $\sigma_\star$, although the relation is not one-to-one. Therefore, we can deduce that if $M_{\rm BH}$ correlates to one host galaxy property, it also correlates with another. In order to determine which galaxy property is the main driver, we can examine the intrinsic scatter. The relationship with the smallest intrinsic scatter should be the most fundamental relationship. So, when comparing the intrinsic scatter (ϵ_0) in Figures 4 and 5, we can determine that our $M_{\rm BH}-\sigma_\star$ relationship is the most fundamental. as it has the smaller intrinsic scatter value.

2.3 Model Comparison Using Nested Sampling

For this section, we will be using different log-likelihood and prior functions to fit our data and compare which models for our given data set result in the best fit. We will be returning to the $M_{\rm BH}-\sigma_\star$ relationship for three models of varying complexity using Dynesty. To simplify the problem, we will consider uncertainties on the dependent variable ($\log_{10}(M_{\rm BH}/M_\odot)$) only, and not the uncertainties on the independent variable $\log_{10}(\sigma_\star/{\rm km\ s}^{-1})$). Our different models are given below.

2.3.1 Model A.

Model A will only constrain an single constant with an intrinsic scatter term. This means the likelihood function becomes:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi(\epsilon_{y,i}^2+\epsilon_0^2)}} \exp\left[\frac{-(y_i-\mu)^2}{2\pi(\epsilon_{y,i}^2+\epsilon_0^2)}\right] \tag{7}$$

where the log-likelihood function is,

$$\ln(\mathcal{L}) = -\frac{1}{2}X^2 + A \tag{8}$$

where,

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{(\epsilon_{y,i}^2 + \epsilon_0^2)} \tag{9}$$

and,

$$A = \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2)}}\right]. \tag{10}$$

Where μ is our constant, and the rest of the values follow that in which was explained when discussing our original log-likelihood function in section 2.2, the only value we are missing is $\epsilon_{x,i}$, which we assumed to be zero.

2.3.2 Model B.

Model B will be a line with intrinsic scatter. The likelihood function is given by:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2)}} \exp\left[\frac{-(y_i - (\alpha x_i + \beta))^2}{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2)}\right] \tag{11}$$

where the log-likelihood function is,

$$\ln(\mathcal{L}) = -\frac{1}{2}X^2 + A \tag{12}$$

where,

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - (\alpha x_i + \beta))^2}{(\epsilon_{y,i}^2 + \epsilon_0^2)} \tag{13}$$

and,

$$A = \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2)}}\right]. \tag{14}$$

Where the rest of the values follow that in which was explained when discussing our original log-likelihood function in section 2.2, the only value we are missing is $\epsilon_{x,i}$, which we assumed to be zero.

2.3.3 Model C.

Model C will be a more complex model; it involves a quadratic function with three parameters in addition to the intrinsic scatter. The likelihood function is:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2)}} \exp\left[\frac{-(y_i - (\delta x_i^2 + \alpha x_i + \beta))^2}{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2)}\right] \tag{15}$$

where the log-likelihood function is,

$$\ln(\mathcal{L}) = -\frac{1}{2}X^2 + A \tag{16}$$

where,

$$X^2 = \sum_{i=1}^{N} \frac{-(y_i - (\delta x_i^2 + \alpha x_i + \beta))^2}{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2)} \tag{17}$$

and,

$$A = \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi(\epsilon_{y,i}^2 + \epsilon_0^2)}}\right]. \tag{18}$$

Where δ is the coefficient in from of the quadratic term. The rest of the values follow that in which was explained when discussing our original log-likelihood function in section 2.2, the only value we are missing is $\epsilon_{x,i}$, which we assumed to be zero.

2.3.4 Plotting our Models.

We will have to adopt the same uniform prior for a given parameter across all models. We will be using a uniform prior of [-40,60] for the constant (μ in Model A, and β in Model B), and [0,5] for the intrinsic scatter in all three models. We will also adopt a uniform prior of [-50,50] for the coefficient in the front of the linear term, α, in models B and C, and a [-20,20] uniform prior for the quadratic term coefficient δ in Model C. These priors should be large enough so that the posteriors, regardless of the model, don't get cut off. Below, we show our results from each of our three models in our $M_{\rm BH} - \sigma_\star$ plots.

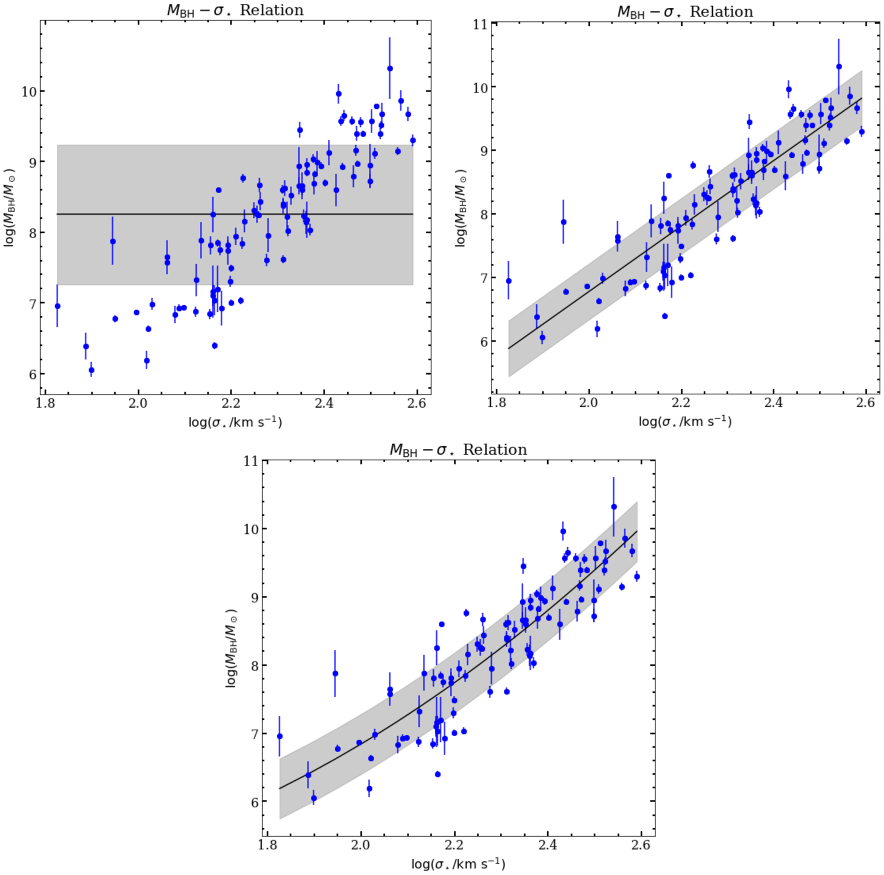


Figure 7: Our $M_{\rm BH} - \sigma_\star$ relationship plots for our different Models. The figure on the upper left is our Model A, the figure on the upper right is our Model B, and the figure on the lower row is our Model C.

Table 1. Our Best-fit Parameters and their 2σ uncertainties

Range	Model	Constant (μ or β)	Intrinsic Scatter (ϵ_0)	Linear Term (α)	Quadratic Term (δ)
Original	A	$\mu = 8.25^{+0.20}_{-0.20}$	$0.99^{+0.16}_{-0.13}$	NA	NA
	B	$\beta = -3.52^{+1.26}_{-1.26}$	$0.44^{+0.08}_{-0.06}$	$5.15^{+0.55}_{-0.54}$	NA
	C	$\beta = 6.68^{+13.47}_{-12.77}$	$0.44^{+0.08}_{-0.06}$	$-3.94^{+11.36}_{-11.98}$	$2.01^{+2.66}_{-2.52}$
Smaller	A	$\mu = 8.25^{+0.20}_{-0.20}$	$0.99^{+0.17}_{-0.13}$	NA	NA
	B	$\beta = -3.52^{+1.24}_{-1.27}$	$0.44^{+0.08}_{-0.06}$	$5.14^{+0.55}_{-0.55}$	NA
	C	$\beta = 6.64^{+13.71}_{-13.64}$	$0.44^{+0.08}_{-0.06}$	$-3.89^{+12.05}_{-12.20}$	$2.01^{+2.69}_{-2.68}$

Note. — Our data is split between two sections, the first being our measurements in the original range covered in section 2.3.4, the second being our smaller range covered in section 2.4.

2.4 Bayes Factor

When comparing models with different numbers of parameters and using diffuse priors, there is a tendency to favor the model with fewer parameters unless the data really are not compatible with the simpler model. Ideally, using Bayes factors for model selection should involve comparing physically motivated models and using meaningful priors that truly reflect our beliefs before the arrival of data. Since that is not the case here, we will decrease our priors to see if there is a change to the preferred model. We will be using a uniform prior of [-20,40] for the constant, and [0,1.5] for the intrinsic scatter. We will also adopt a smaller uniform prior of [-30 to 30] for the coefficient in front of the linear term, and [-10,10] uniform prior for the quadratic term coefficient. We will also be refitting all three models to $M_{\rm BH}-\sigma_\star$ with these smaller priors, which should still be enough so that the posteriors, regardless of the model, don't get cut off. Our results can be seen in Table 1.

In model selection, an important feature is that an alternative model must be specified against which the comparison is made. In other words, it is pointless to reject a theory unless an alternative explanation is available that fits the observed data better, taking into account any additional complicity in the new model. We can find a formal framework for comparing models by using the Bayes factor, B_{01}. A value $B_{01} > 1$ represents an increase of the support in favor of Model 0 versus Model 1 given the observed data, while a value $B_{01} < 1$ represents a decrease of the support in favor of Model 0 versus Model 1. We can find Bayes factors for Model B compared to Model A (B_{BA}), and for model B compared to Model C (B_{BC}).

An important distinction for working with Bayes factors is the number of parameters. When comparing models with different numbers of parameters and using diffuse priors, there is a tendency to favor the model with fewer parameters unless the data really are not compatible with the simpler model.

For the original range, we found the evidence for Model A to be, 5.4957×10^{-64}, with the log-likelihood at -136.269. We found the evidence for Model B to be, 2.3434×10^{-34}, with the log-likelihood at -61.822. We found the evidence for Model C to be, 1.1405×10^{-34}, with the log-likelihood at -60.718.

For the smaller range, we found the evidence for Model A to be 4.1740×10^{-63}, with the log-likelihood at -136.269. We found the evidence for Model B to be 1.9184×10^{-33}, with the log-likelihood at -61.823. We found the evidence for Model C to be 7.0834×10^{-34}, with the log-likelihood at -60.714. All of our data in both ranges have 97 data points.

When using the **original range** for our set of priors and their evidence values, we were able to find Bayes factors of B_{BA} and B_{BC} to be:

$$B_{BA} = 4.264 \times 10^{29}$$

$$B_{BC} = 2.054$$

When using the **smaller range** for our priors and their evidence values, we were able to find the Bayes factors of B_{BA} and B_{BC} to be:

$$B_{BA} = 4.596 \times 10^{29}$$

$$B_{BC} = 2.708$$

When analyzing our values for B_{BA}, we see an extremely large number. When comparing this to Jefferys' scale (Table 1 found on the manual report) we can see that there is very strong evidence in support of our Model B in comparison to our Model A. As our Model B only had one extra parameter when comparing to our Model A, and our Bayes factor for these two models was so high, we can be fairly confident that the extra parameter is more than worth the trouble to get a better fit.

When analyzing our values for B_{BC}, we get a much smaller number. One that does in fact fall between 1 - 3 on Jefferys' scale, resulting in a change that is barely worth a mention. When comparing the amount of our parameters from our Model B to our Model C, we can note that that while Model C had a slightly better fit with four parameters than our Model B with a fit of three parameters, the difference is so small that it is barely worth a mention. Meaning that Model B will give us roughly the same fit as Model C, and will ultimately be simpler, as a result of using one less parameter.

While our inferred model parameter values changed slightly when comparing the original range to the smaller range, the change is so small that our results are left unchanged.

2.5 Bayesian Information Criterion

Another commonly used criterion for model selection is the Bayesian information criterion (BIC). It is defined as:

$$\mathrm{BIC} \equiv -2\ln\mathcal{L}_{\max} + k\ln N \tag{19}$$

Where $\mathcal{L}_{\max}$ is the maximum likelihood value, k is the number of fitted parameters, and N is the number of data points. The second term is a default penalty term, working against more complex models. We should pick the model with the smaller value of the BIC, and compare that BIC value to the BIC values of the other models. The strength of evidence against a model with a higher BIC is summarized in Table 2, Interpreting BIC Values, found in the manual.

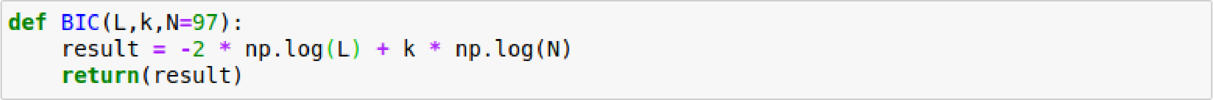
In general, the BIC is easier to use than doing a full integration to obtain the evidence and compute a Bayes factor. The BIG is actually an approximation to the full Bayesian evidence. However, there are some simplifying assumptions inherent in the BIC, such as the data points are independent and the posterior distribution is Gaussian or near-Gaussian. Also, the BIC penalizes extra parameters regardless of whether they are constrained by the data or not, unlike the Bayesian evidence.

Table 2. Our BIC Values and the numbers we used to calculate them

Range	Model	BIC	$\mathcal{L}_{\max}$	k	N
Original	A	281.6874	6.593×10^{-60}	2	97
	B	137.368	1.415×10^{-27}	3	97
	C	139.7348	4.270×10^{-27}	4	97
Smaller	A	281.6874	6.593×10^{-60}	2	97
	B	137.3701	1.414×10^{-27}	3	97
	C	139.7268	4.287×10^{-27}	4	97

Note. — Our data is split between two sections, the first being our measurements in the original range covered in section 2.3.4, the second being our smaller range covered in section 2.4.

When calculating the BIC from our own models with the original and smaller priors, we got the following results shown above in Table 2. We used the following code to easily find our BIC Values.



```
def BIC(L,k,N=97):
    result = -2 * np.log(L) + k * np.log(N)
    return(result)
```

Figure 8: Our code that lists our BIC equation.

Ultimately, our BIC for Model B within our original range is the smallest value, resulting in it being the preferred model. From this, we can measure how strongly preferred Model B is when finding Δ BIC for each of our Models. Right off the back, when comparing with the original range, the Δ BIC$_{\mathrm{BA}}$ value favors a very strong preference towards Model B with a value of 144.319. Then, the Δ BIC$_{\mathrm{BC}}$ value favors a positive preference towards Model B with a value of 2.366. As can be assumed, this relationship carries into the smaller range with it barely being worth a mention for the B model, a positive preference for the C model, and a very strong preference towards Model B in the original range when compared to the A model.

There is agreement in preference towards Model B in the original range with both Bayes factors and BIC values. While our priors make some minor changes, they ultimately do not change the results we find, that Model B is the best Model for our fit.