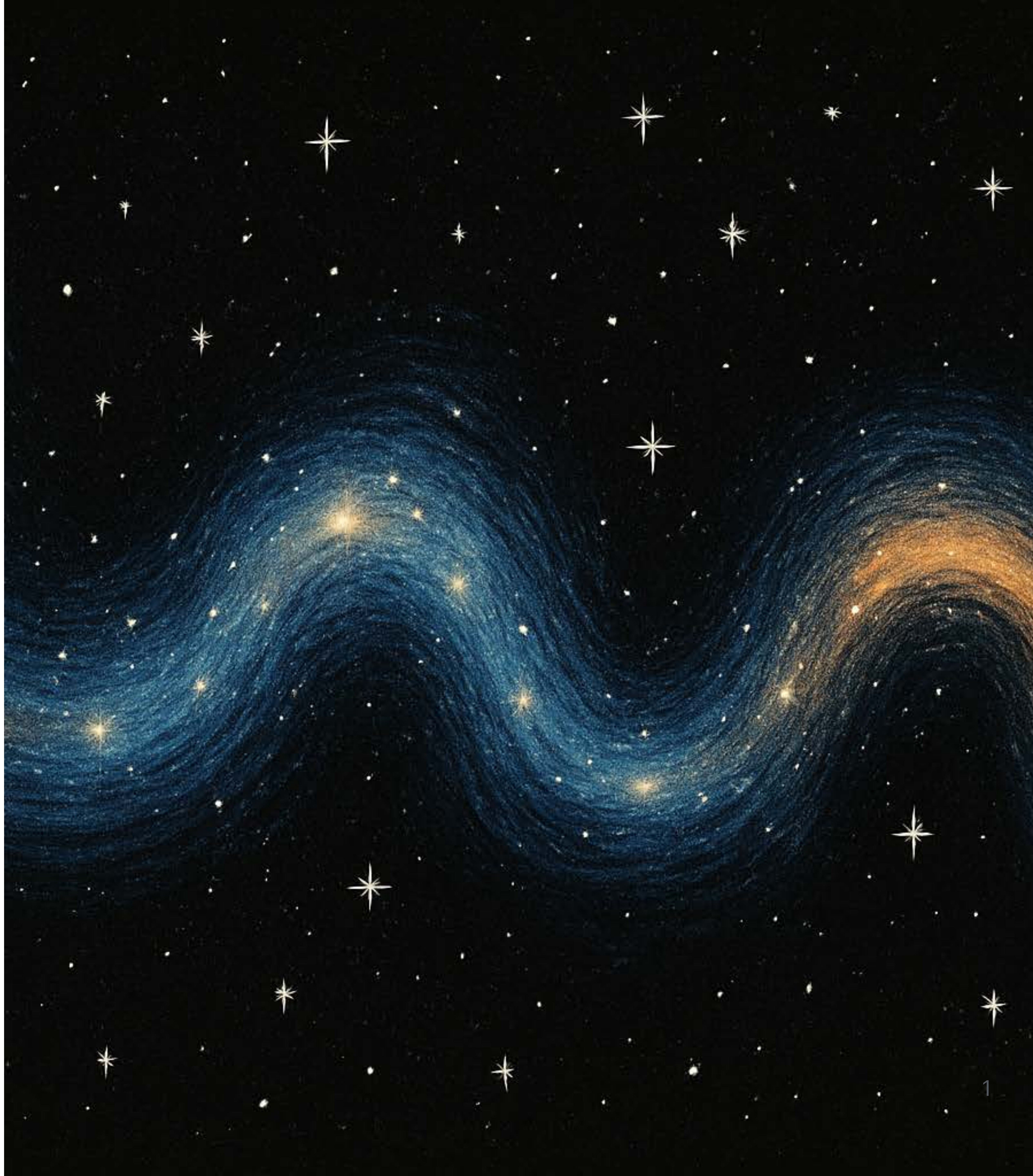
Nested Sampling for ARIMA Model Selection : A Novel Approach to Astronomical Time Series Analysis

Cambridge Mathematics Placement 2025

Ajinkya J. Naik
Supervisor : Dr. Will Handley

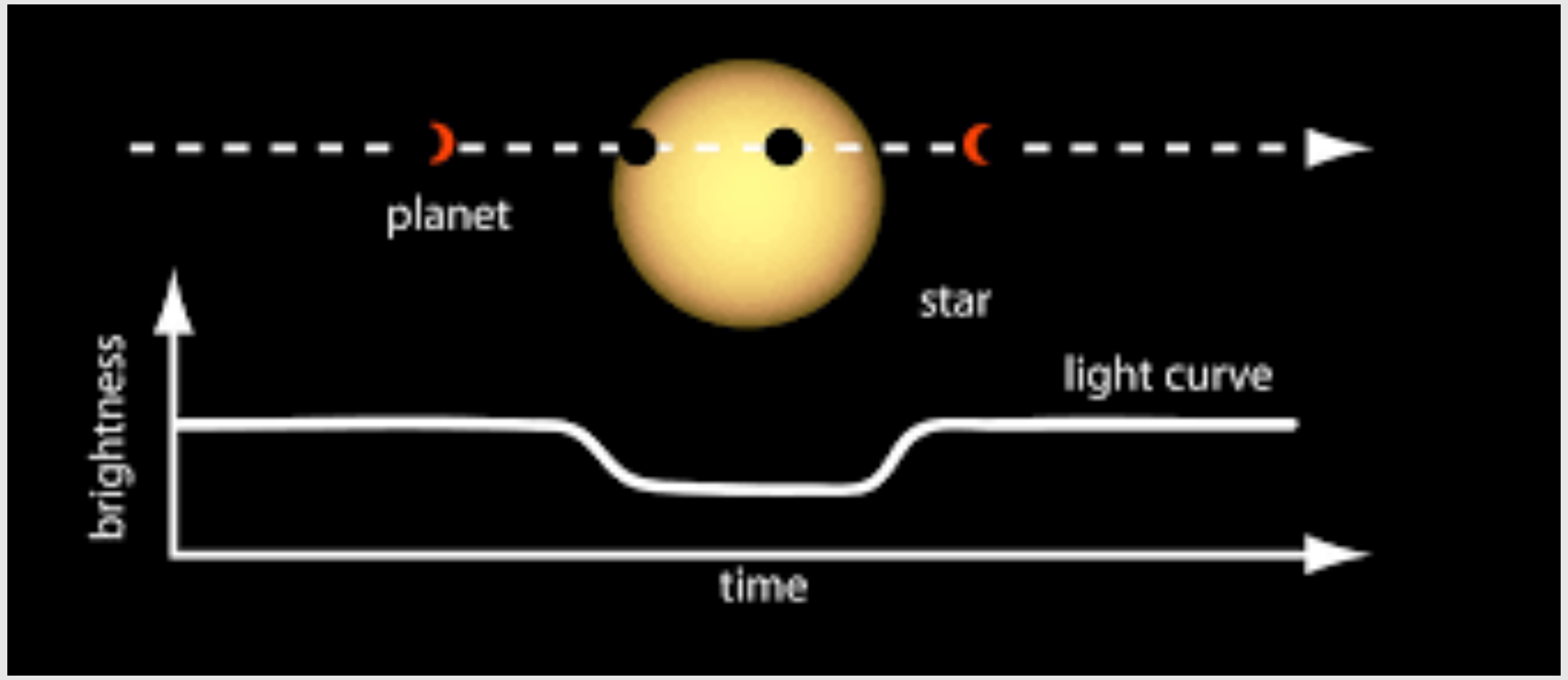
ioa

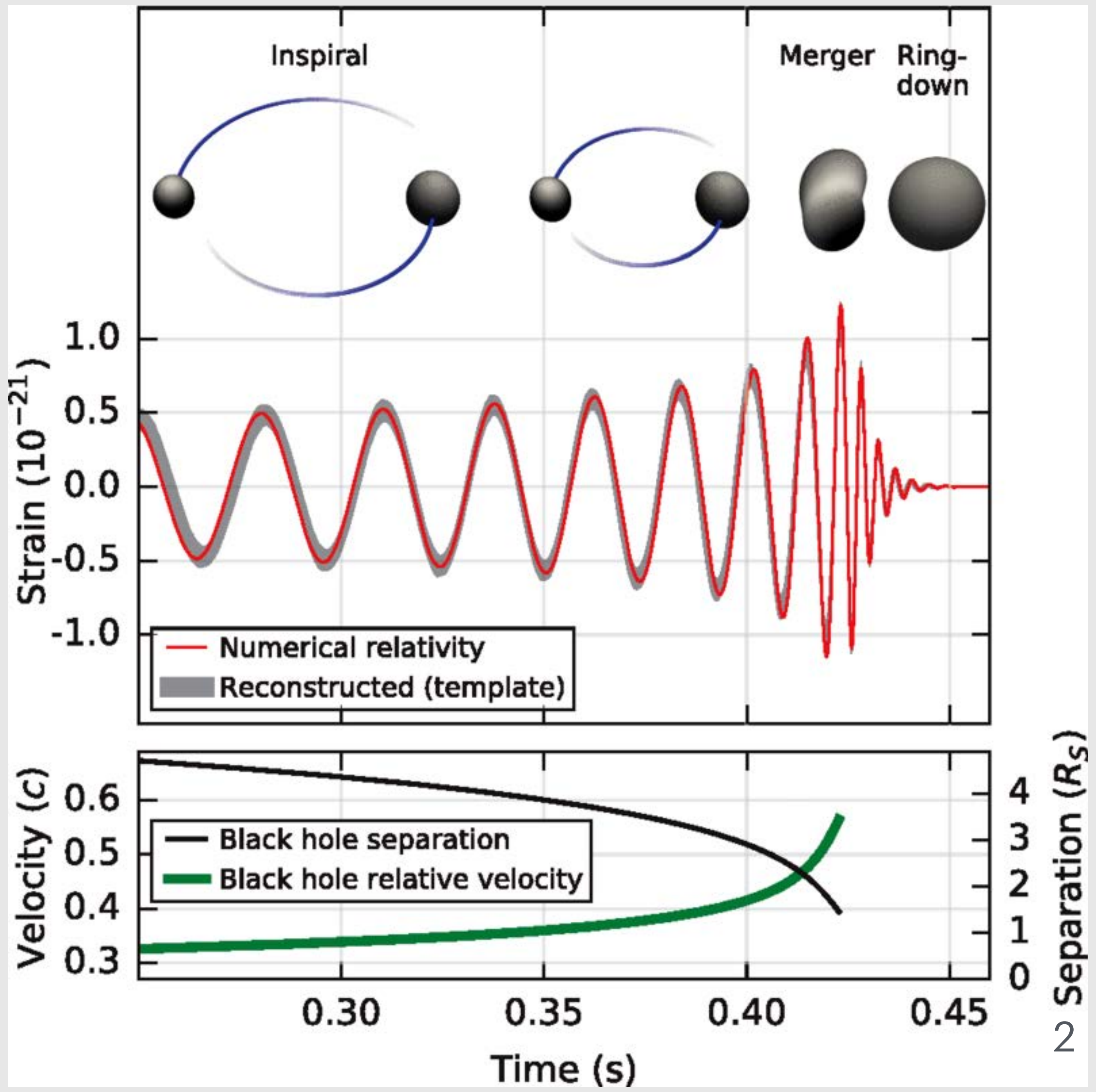
UNIVERSITY OF CAMBRIDGE

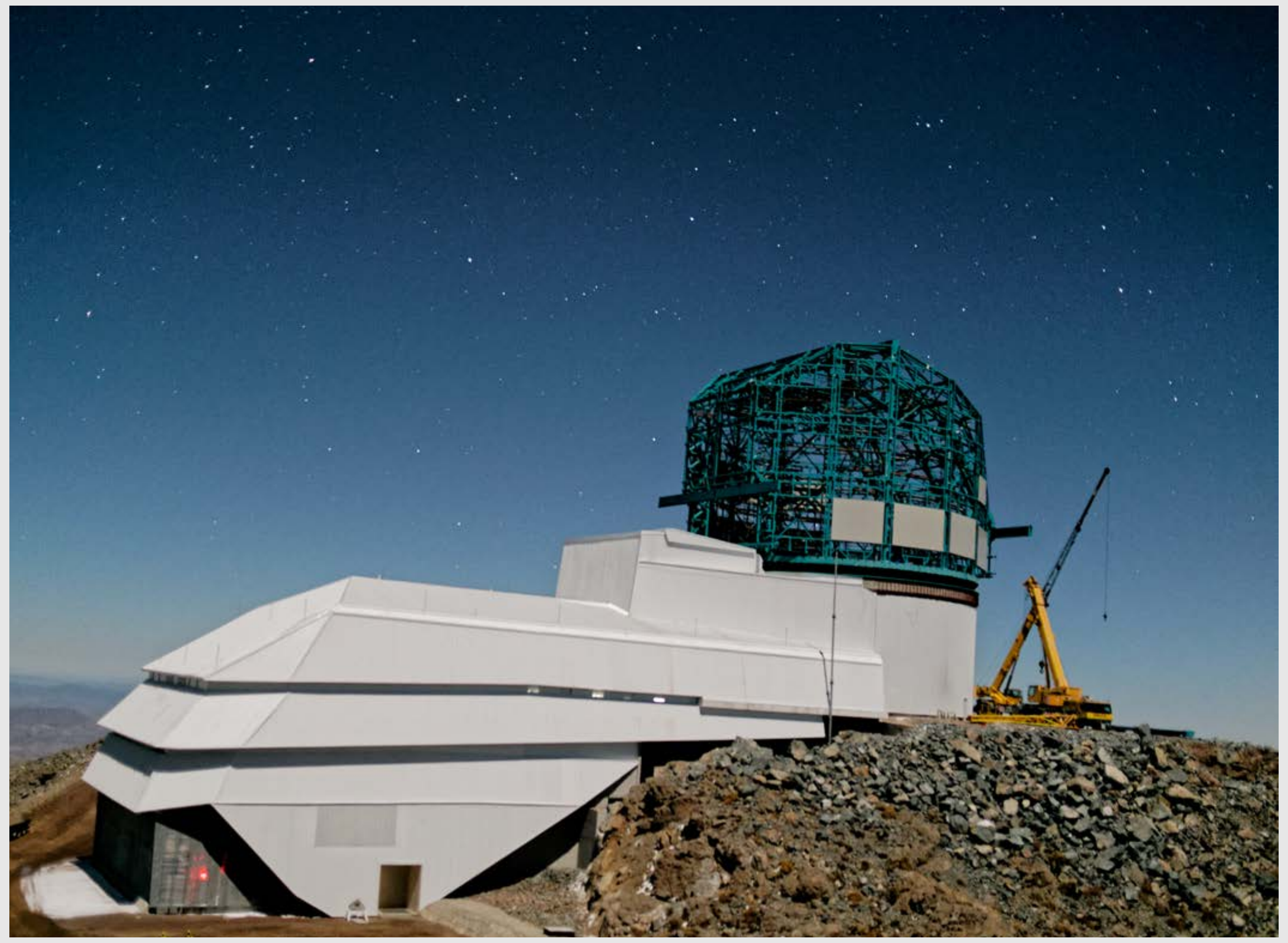


Introduction

- The era of time-domain astronomy.
- Common analysis methods : gaussian processes, polynomial models, machine learning etc.

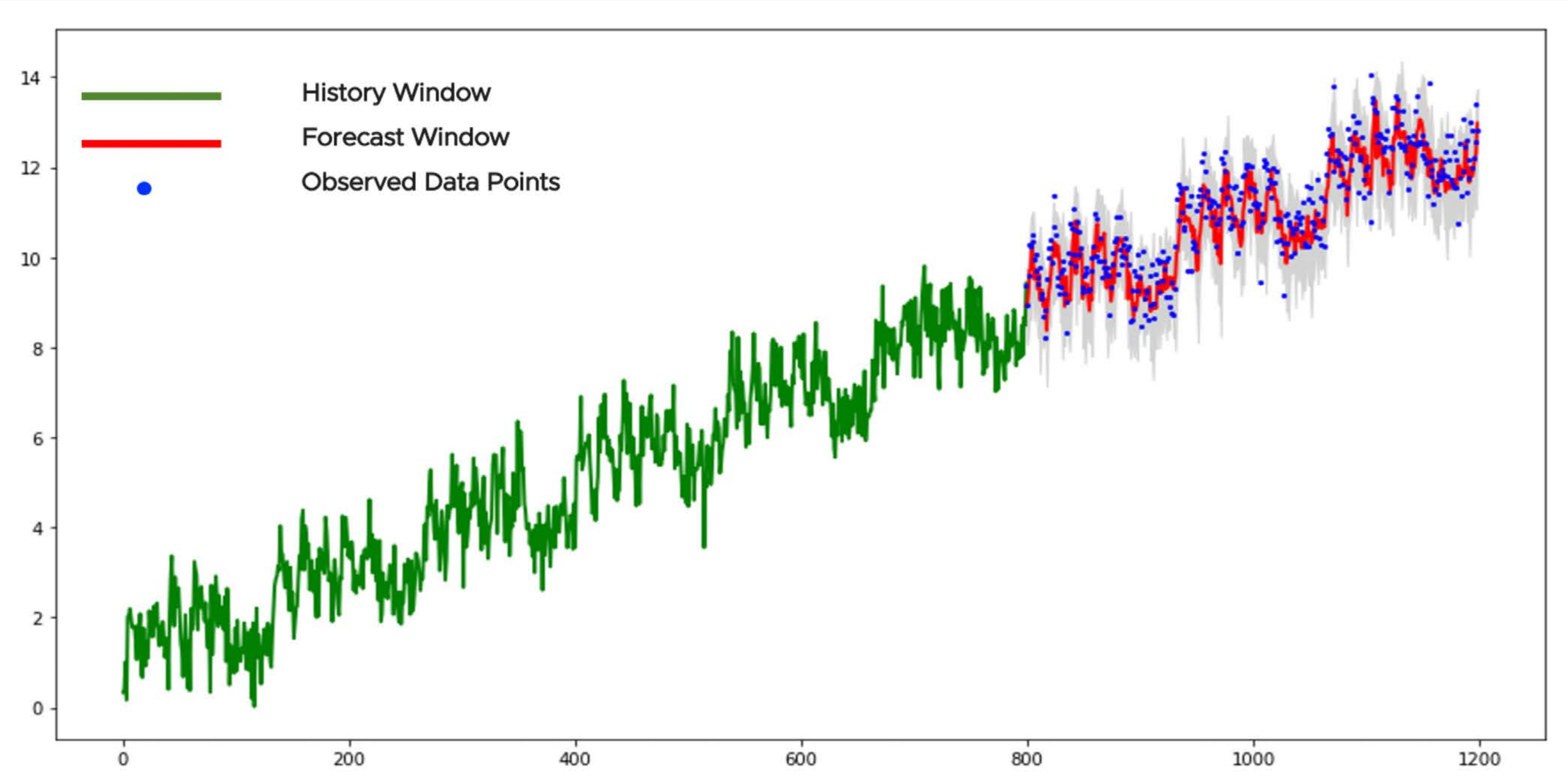






Large Synoptic Survey Telescope (LSST)

ARIMA models

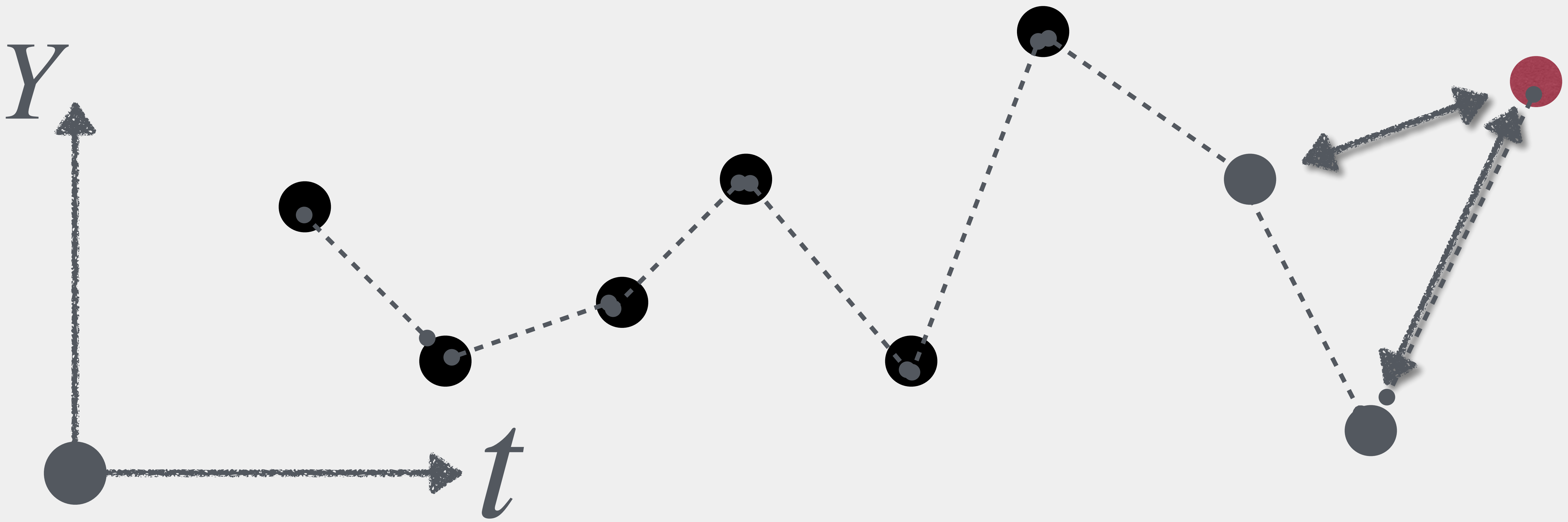


- Analysing and forecasting points $\hat{y}_t$ for a given time series y_t
- **Autoregressive (AR), Integrated (I), Moving Average (MA)**

What are ARIMA Models?

Autoregressive AR(p) :

- Modelling "autocorrelation" in time series.
- Each datapoint correlated with its own previous (or "lagged") values.
- For example : Daily average temperature

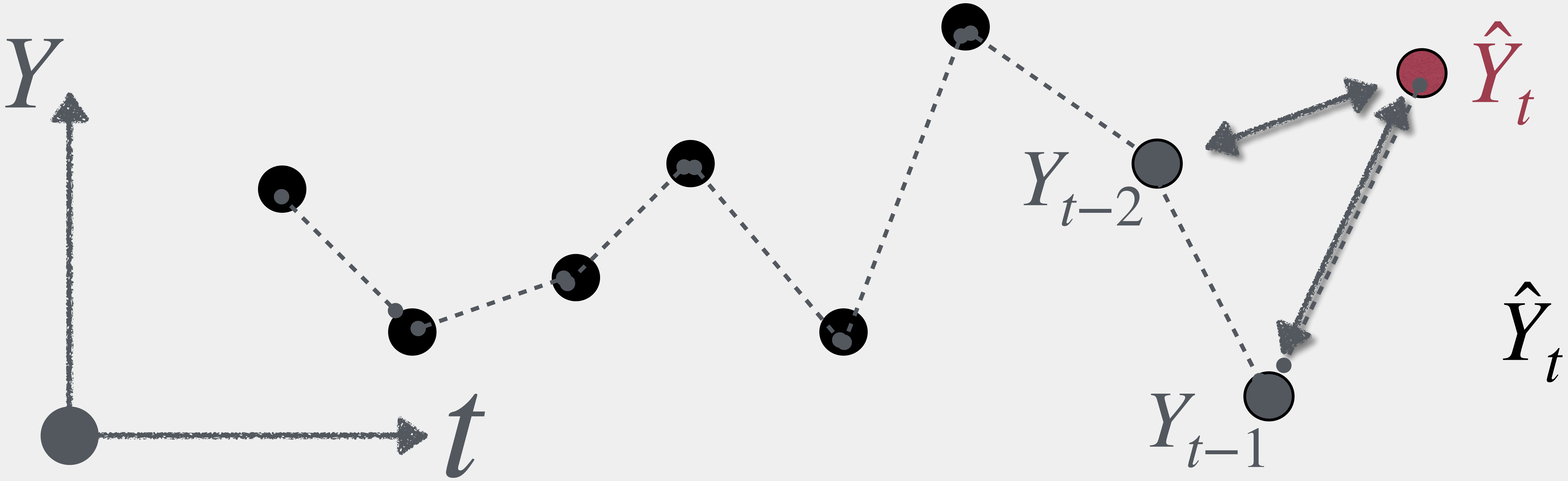


What are ARIMA Models?

Autoregressive AR(p) :

- Introduced in 1927, by Yule to model sunspot numbers.
- **p** denotes number of lagged terms.
- For example p=2 :

Udny Yule



$$\hat{Y}_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$$

What are ARIMA Models?

Moving Average MA(q) :

- Similar to **AR** models but present points correlated with lagged forecast errors (residuals).

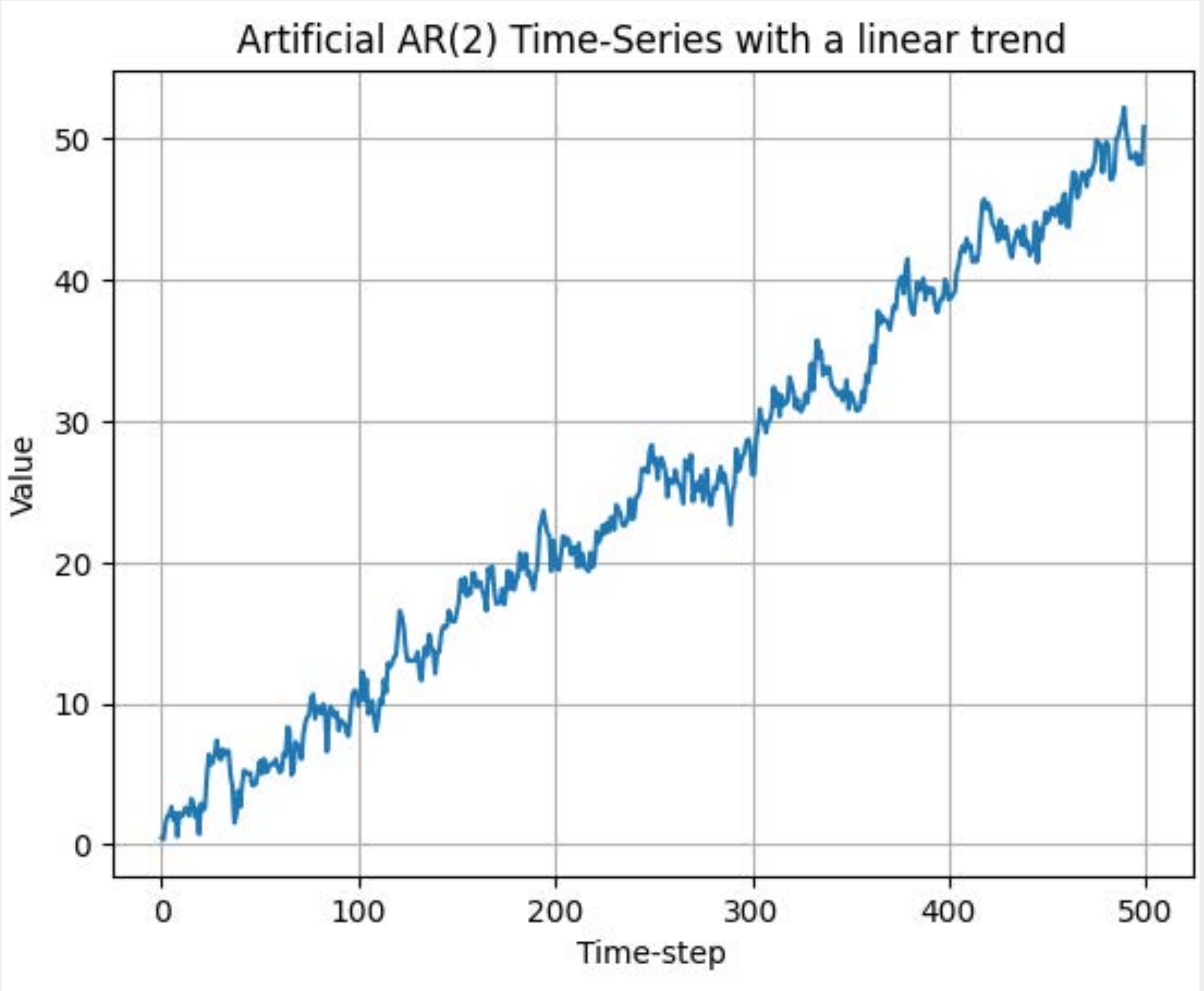
- q—> number of lagged forecast errors. For example, MA(2) model :

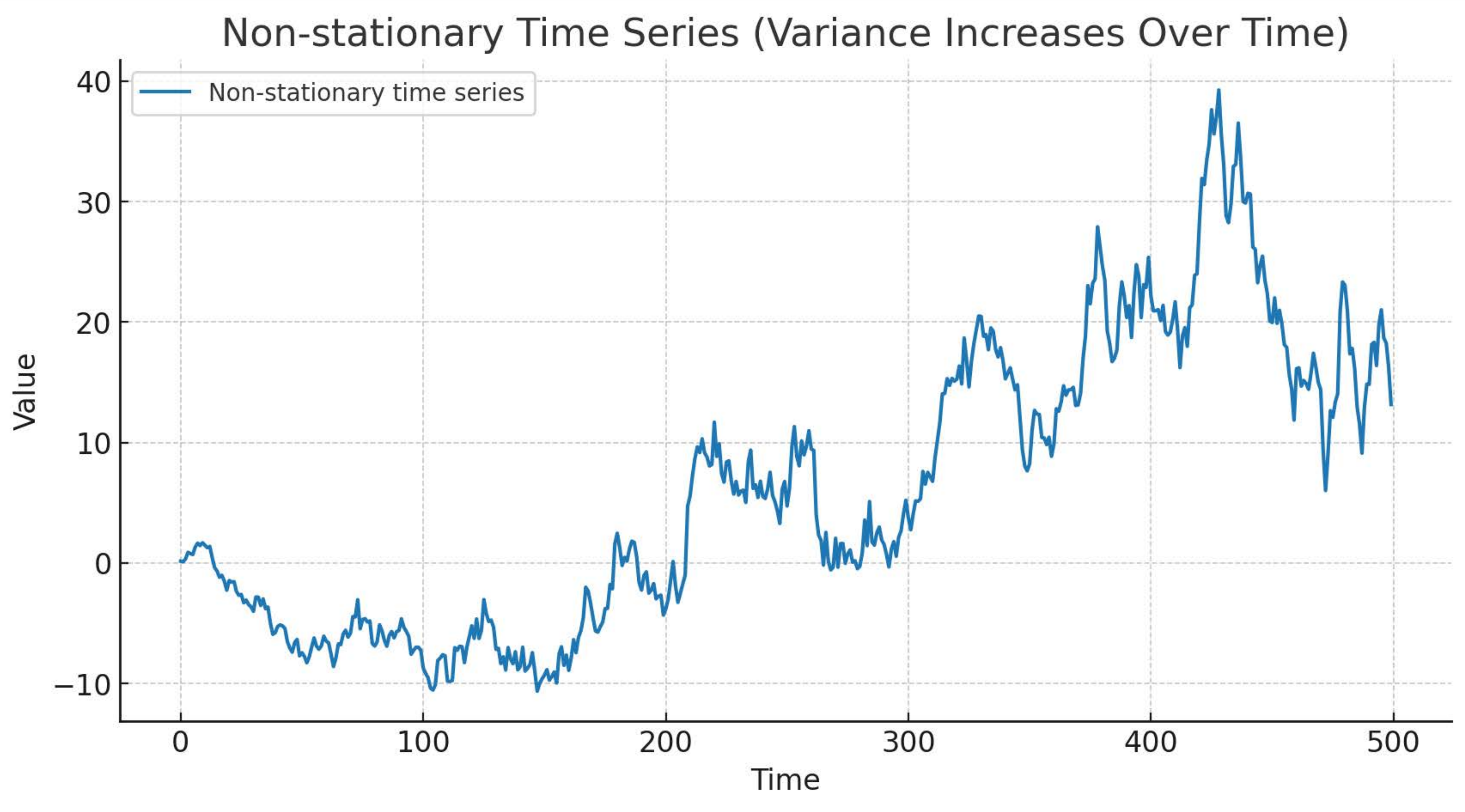
$$\hat{y}_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t$$

What are ARIMA Models?

Integrated I (d) :

- ARMA modelling requires a stationary time series i.e. constant mean and variance.
- Integrated (I) part of ARIMA takes care of this by de-trending the time series using finite differencing.
- d —> Order of differencing





What are ARIMA Models?

ARIMA (p, d, q):

- Combined into **ARIMA (p, d, q)** by Box and Jenkins in 1971.
- Used widely in economics, finance and weather/climate predictions.
- Not so common in Astronomy

$$\hat{y}_t = \mu + \phi_p y_{t-p} + \theta_q \epsilon_{t-q} + \epsilon_t$$

What are ARIMA Models?

ARIMA (p, d, q):

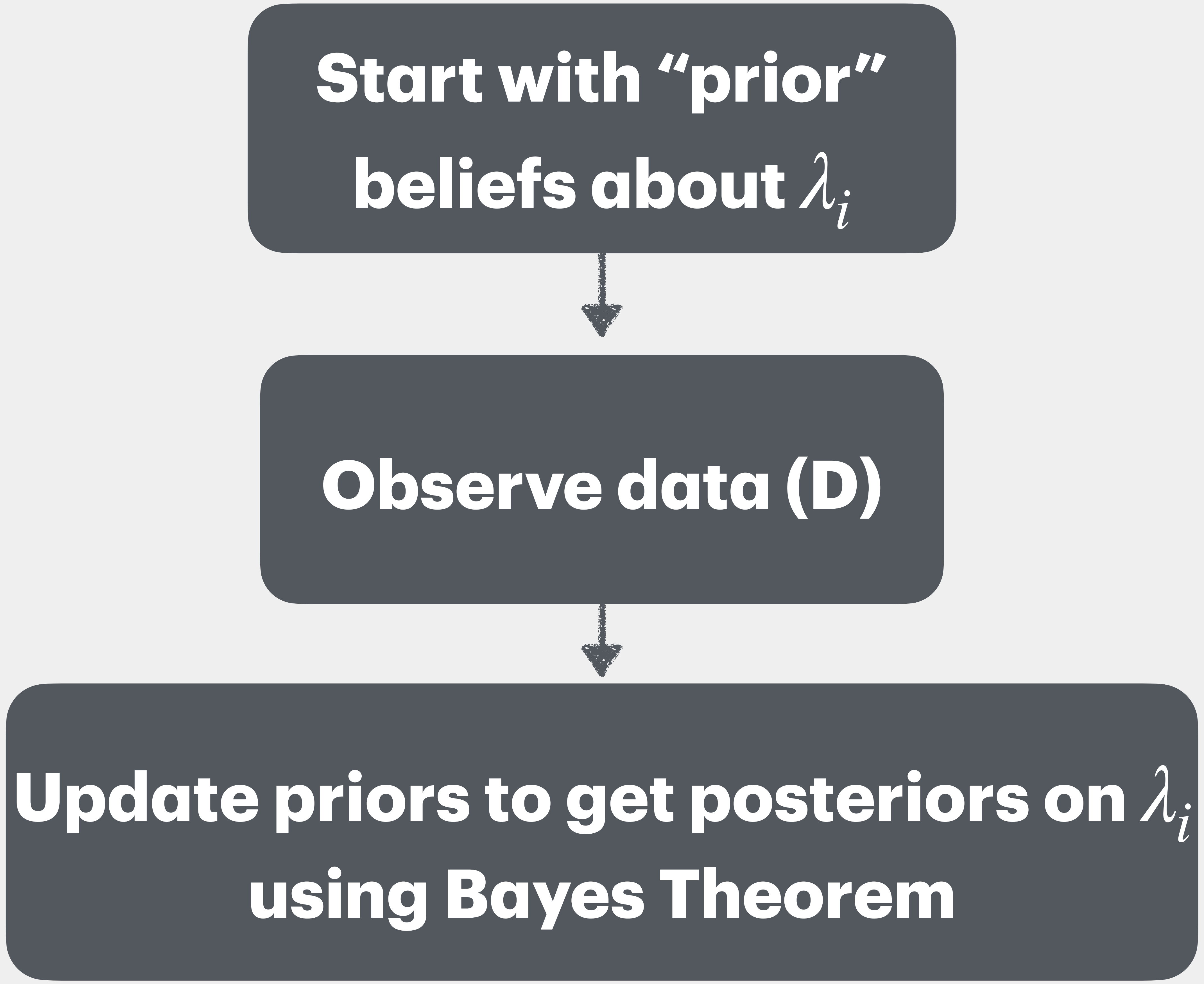
- **p, d** and **q** could be any positive integers.
- Difficult to select the right **(p, d, q)** order for fitting data.
- ARIMA models are over-parameterised, always a risk of overfitting.
- Need a method to choose the correct ARIMA Model for any given data.

$$\hat{y}_t = \mu + \phi_p y_{t-p} + \theta_q \epsilon_{t-q} + \epsilon_t$$

Bayesian Inference and Nested Sampling

A Primer on Bayesian Inference

- Infer the distribution of parameter values λ_i of a model **M** from data **D**.



Bayesian Inference and Nested Sampling

A Primer on Bayesian Inference :

$$P(\lambda_i; M \,|\, D) = \frac{L(D \,|\, \lambda_i; M)\pi(\lambda_i; M)}{Z}$$

Bayesian Inference and Nested Sampling

A Primer on Bayesian Inference :

Prior

$$P(\lambda_i; M \mid D) = \frac{L(D \mid \lambda_i; M)\pi(\lambda_i; M)}{Z}$$

Bayesian Inference and Nested Sampling

A Primer on Bayesian Inference :

Posterior

Prior

$$P(\lambda_i; M \mid D) = \frac{L(D \mid \lambda_i; M)\pi(\lambda_i; M)}{Z}$$

Bayesian Inference and Nested Sampling

A Primer on Bayesian Inference :

Likelihood

Prior

Posterior

$$P(\lambda_i; M \mid D) = \frac{L(D \mid \lambda_i; M)\pi(\lambda_i; M)}{Z}$$

Bayesian Inference and Nested Sampling

A Primer on Bayesian Inference :

Likelihood

Prior

Posterior

$$P(\lambda_i; M \,|\, D) = \frac{L(D \,|\, \lambda_i; M)\pi(\lambda_i; M)}{Z}$$

$$Z = \int L(D \,|\, \lambda_i)\pi(\lambda_i)d\lambda_i$$

Bayesian Inference and Nested Sampling

A Primer on Bayesian Inference :

Likelihood

Prior

Posterior

$$P(\lambda_i; M \mid D) = \frac{L(D \mid \lambda_i; M)\pi(\lambda_i; M)}{Z}$$

Evidence : $Z = \int L(D \mid \lambda_i)\pi(\lambda_i)d\lambda_i$

Bayesian Inference and Nested Sampling

A Primer on Bayesian Inference :

Posterior **Likelihood** **Prior**

$$P(\lambda_i; M \,|\, D) = \frac{L(D \,|\, \lambda_i; M)\pi(\lambda_i; M)}{Z}$$

Evidence : $Z = \int L(D \,|\, \lambda_i)\pi(\lambda_i)d\lambda_i$

Useful for model comparison!!

Bayesian Inference and Nested Sampling

Evidence for Model Comparison

$$Z_n = \int L(D \mid \lambda_i; M_n)\pi(\lambda_i; M_n)d\lambda_i = \mathbf{P(D \mid M_n)}$$

- Model with higher evidence statistically preferred by data.
- But, cumbersome to evaluate due to “curse of dimensionality”.
- Solution —> Nested Sampling!

Bayesian Inference and Nested Sampling

The Nested Sampling Algorithm

- Introduced by physicist John Skilling in 2003.
- Key idea is to define the “prior volume” - amount of prior mass contained inside an equal likelihood contour.

$$X(L) = \int_{L>L(\lambda)} \pi(\lambda) d\lambda$$

- Transform the multi-dimensional evidence integral to a simple one-dimensional integral:

$$Z(X) = \int_0^1 L(X) dX$$

ARIMA x Nested Sampling

The Idea

$$\hat{y}_t = \mu + \phi_p y_{t-p} + \theta_q \epsilon_{t-q} + \epsilon_t$$

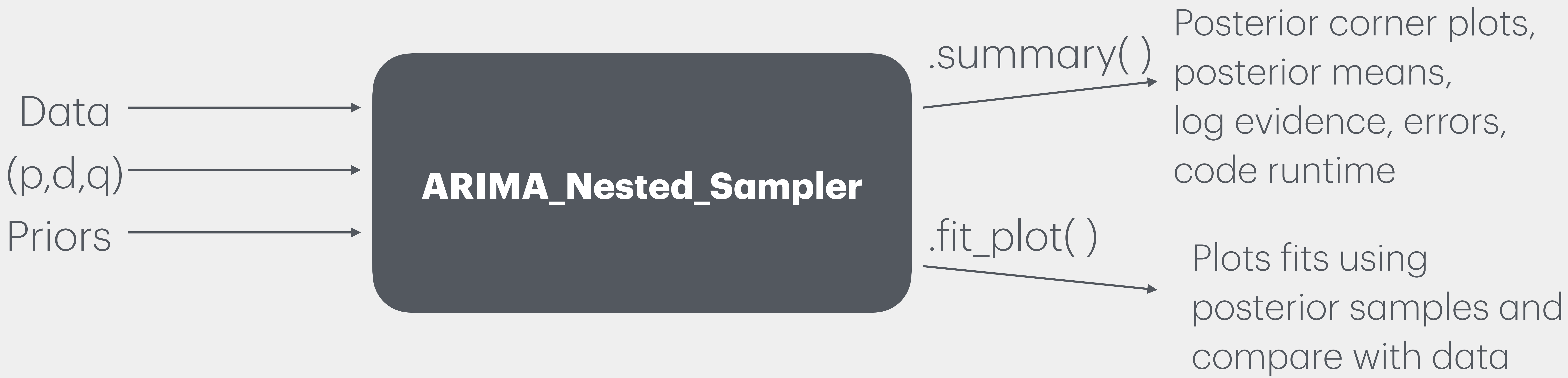
$$P(\lambda_i; M \,|\, D) = \frac{L(D \,|\, \lambda_i; M)\pi(\lambda_i; M)}{Z}$$

- Use the weights (ϕ_p, θ_q) and the standard deviation σ characterising ϵ_t as parameters λ_i for Bayesian Inference.
- Nested Sampling serves as an efficient tool : model selection + posterior distributions for parameters.
- Occam's penalty ensures overfitting is avoided.

ARIMA x Nested Sampling

The Code

- **BlackJAX** Nested Sampler.
- Leveraging the JAX ecosystem (runtime reduced from 3-4 minutes to just few seconds!)
- Main object : **ARIMA_Nested_Sampler** class

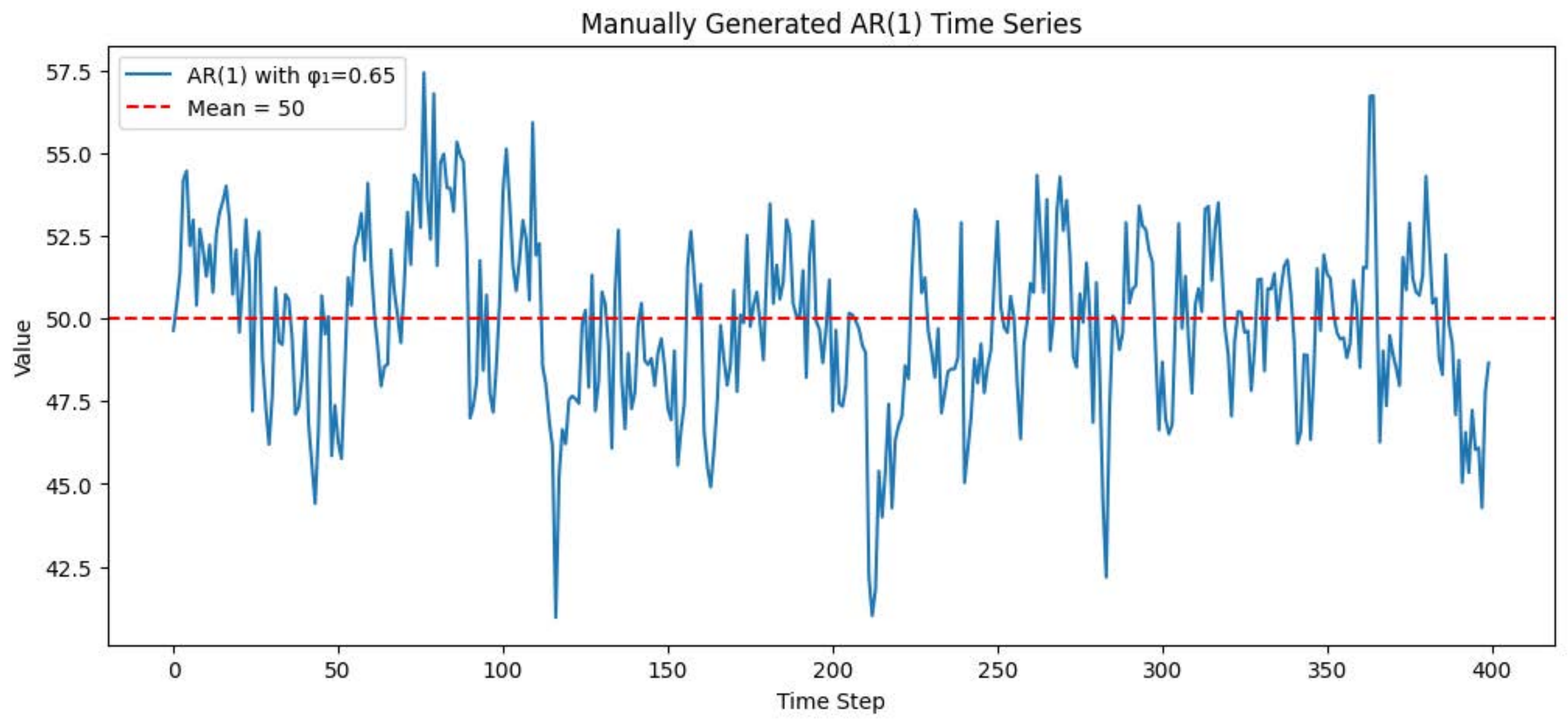


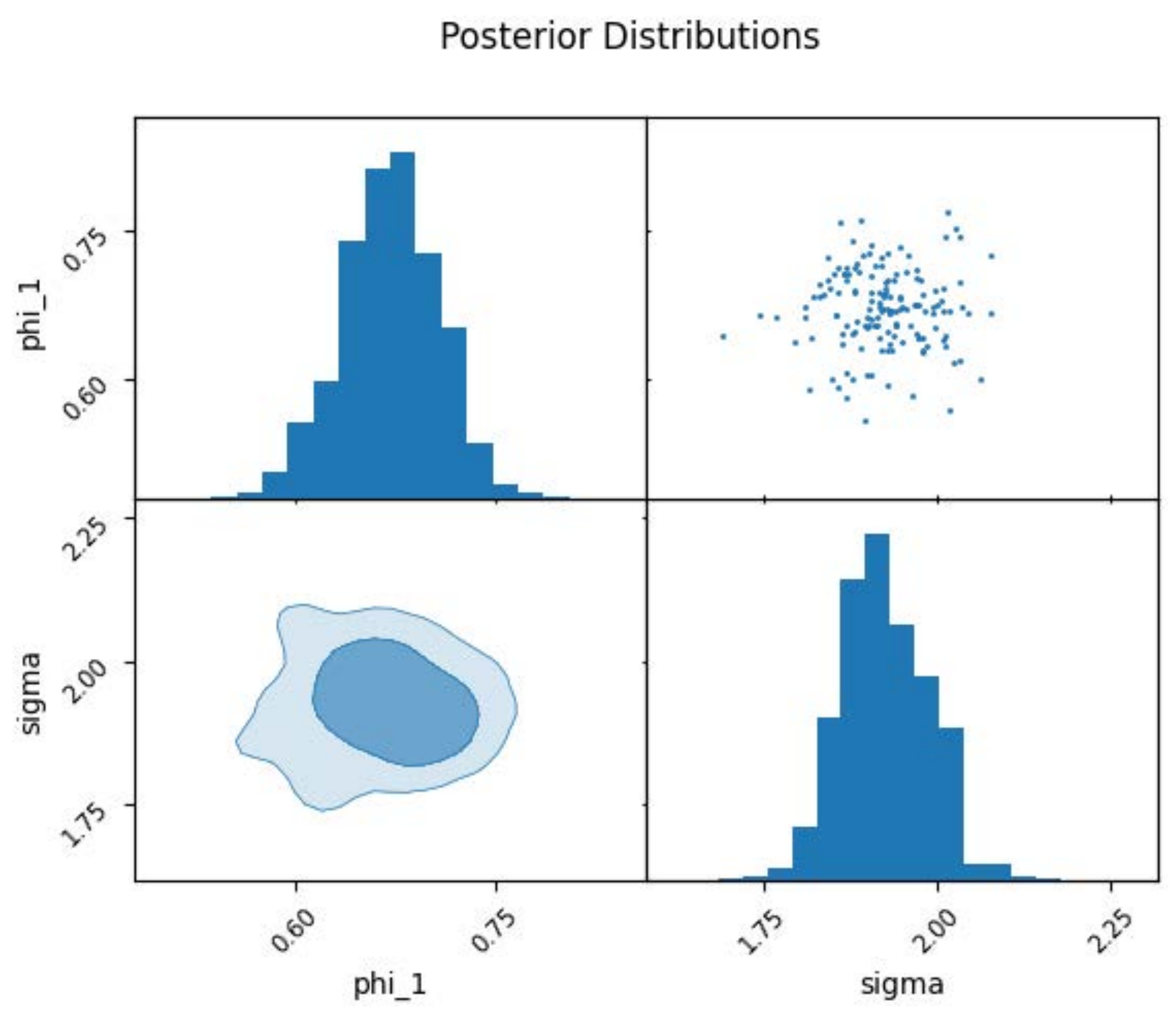
ARIMA x Nested Sampling

Model Comparison

ARIMA x Nested Sampling

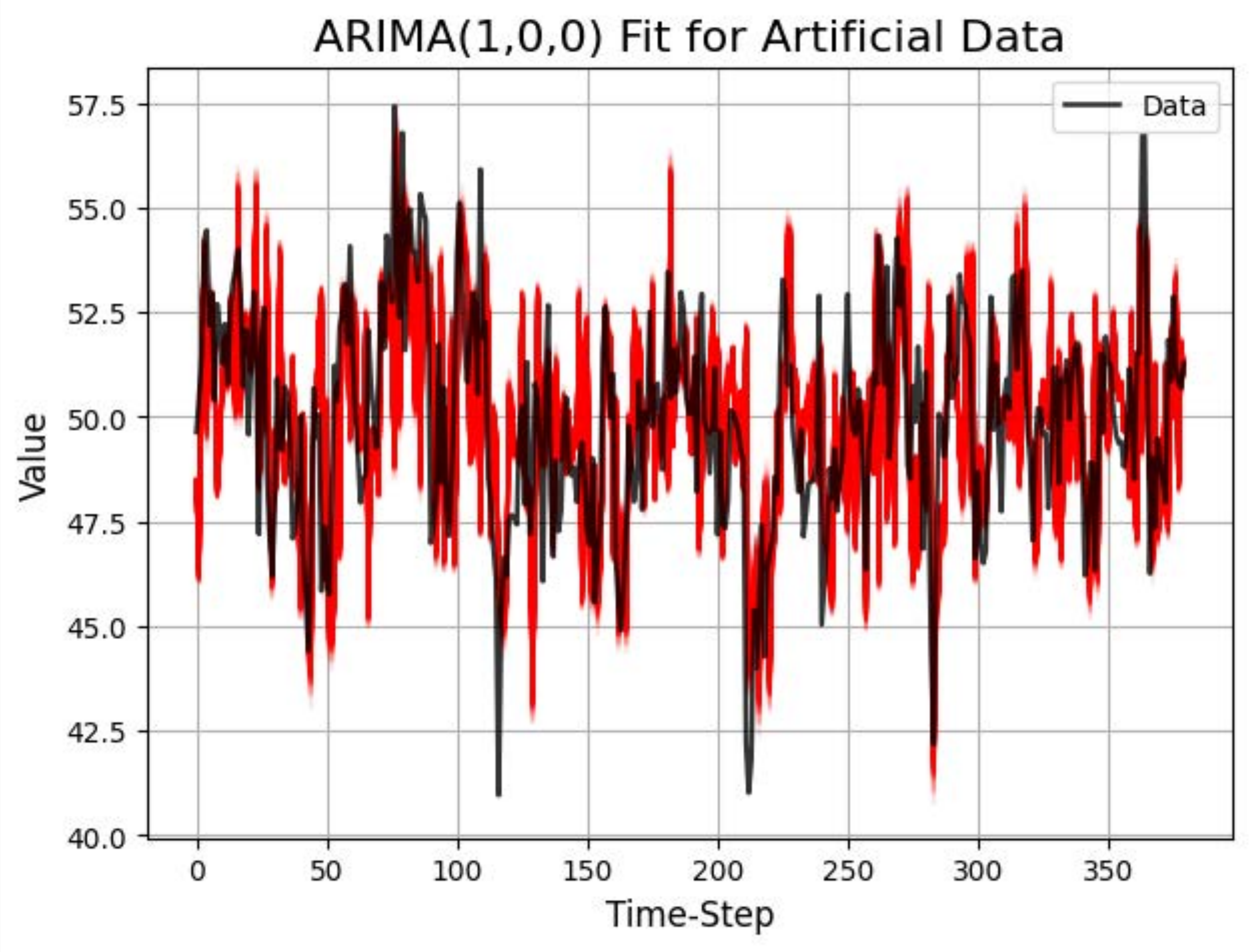
Testing on synthetic data





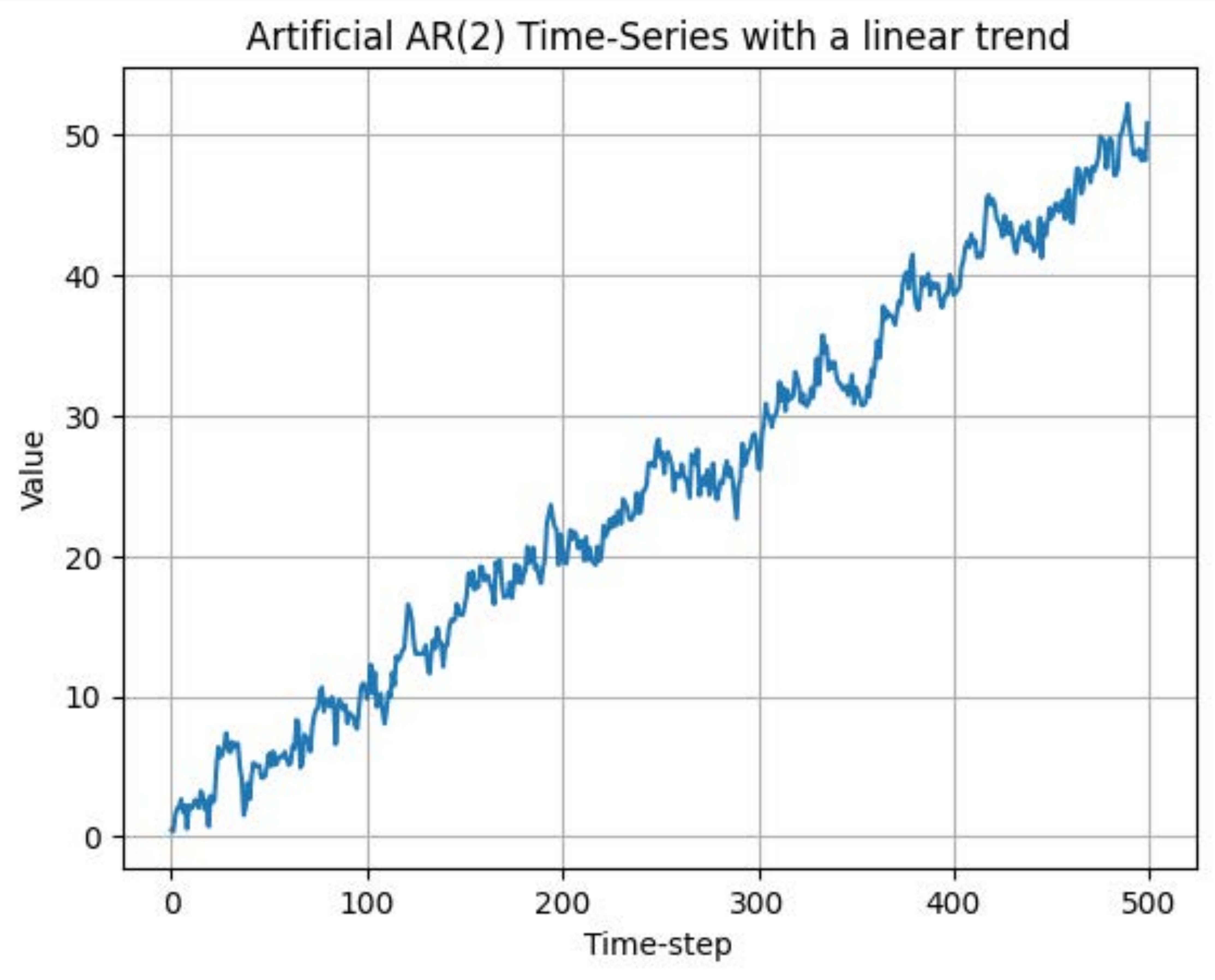
ARIMA x Nested Sampling

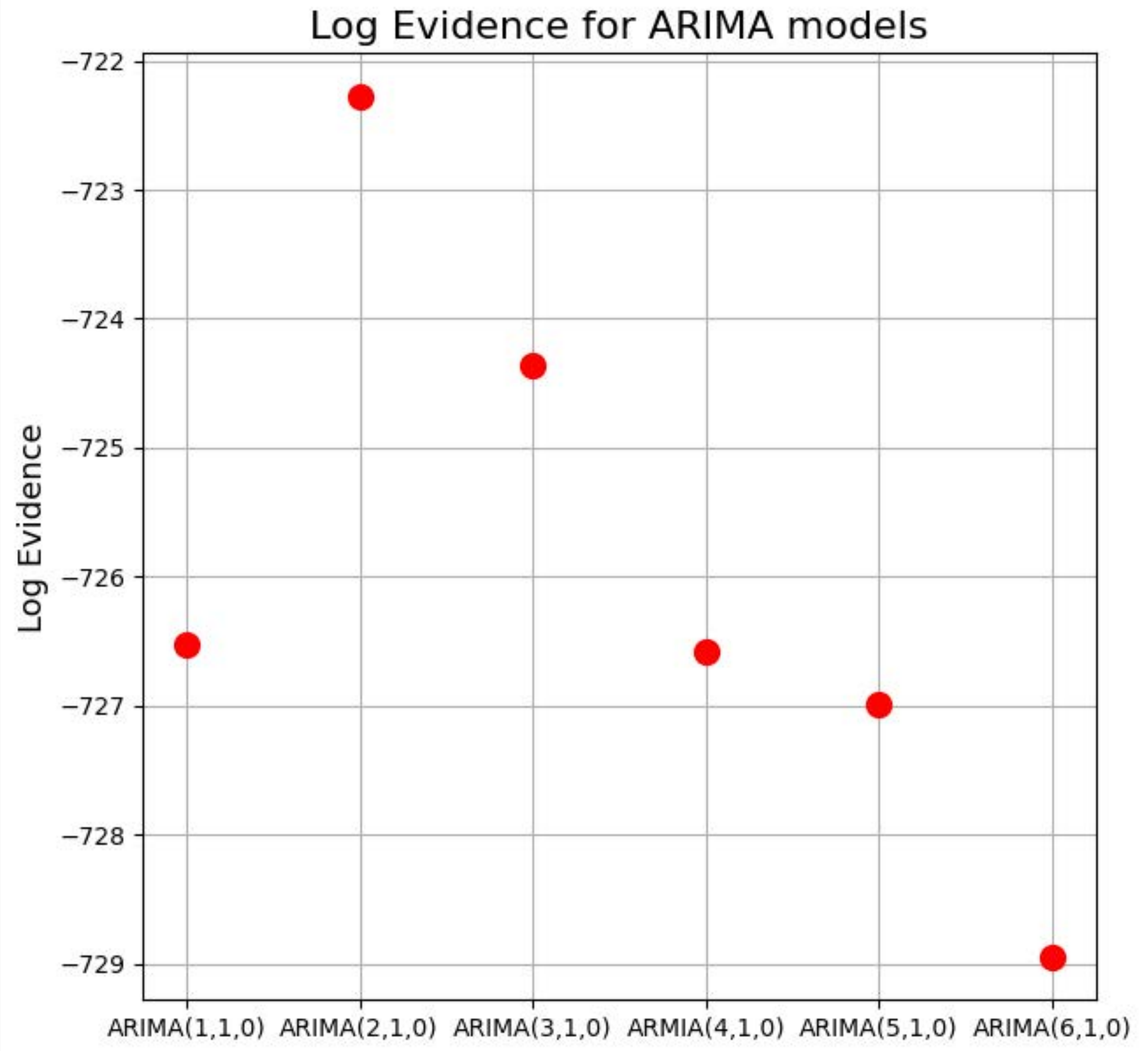
Testing on synthetic data



ARIMA x Nested Sampling

Testing on synthetic data

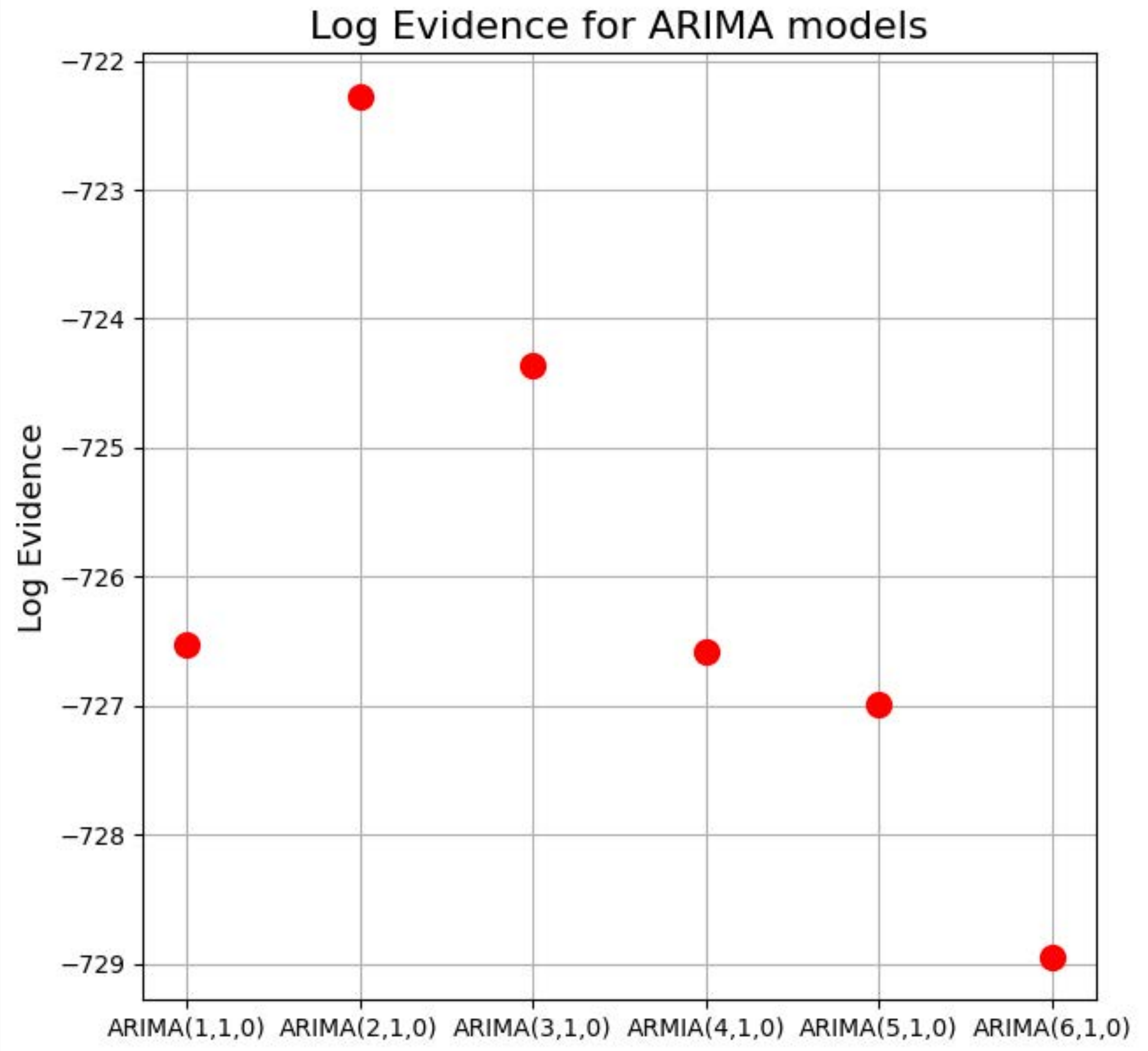




ARIMA x Nested Sampling

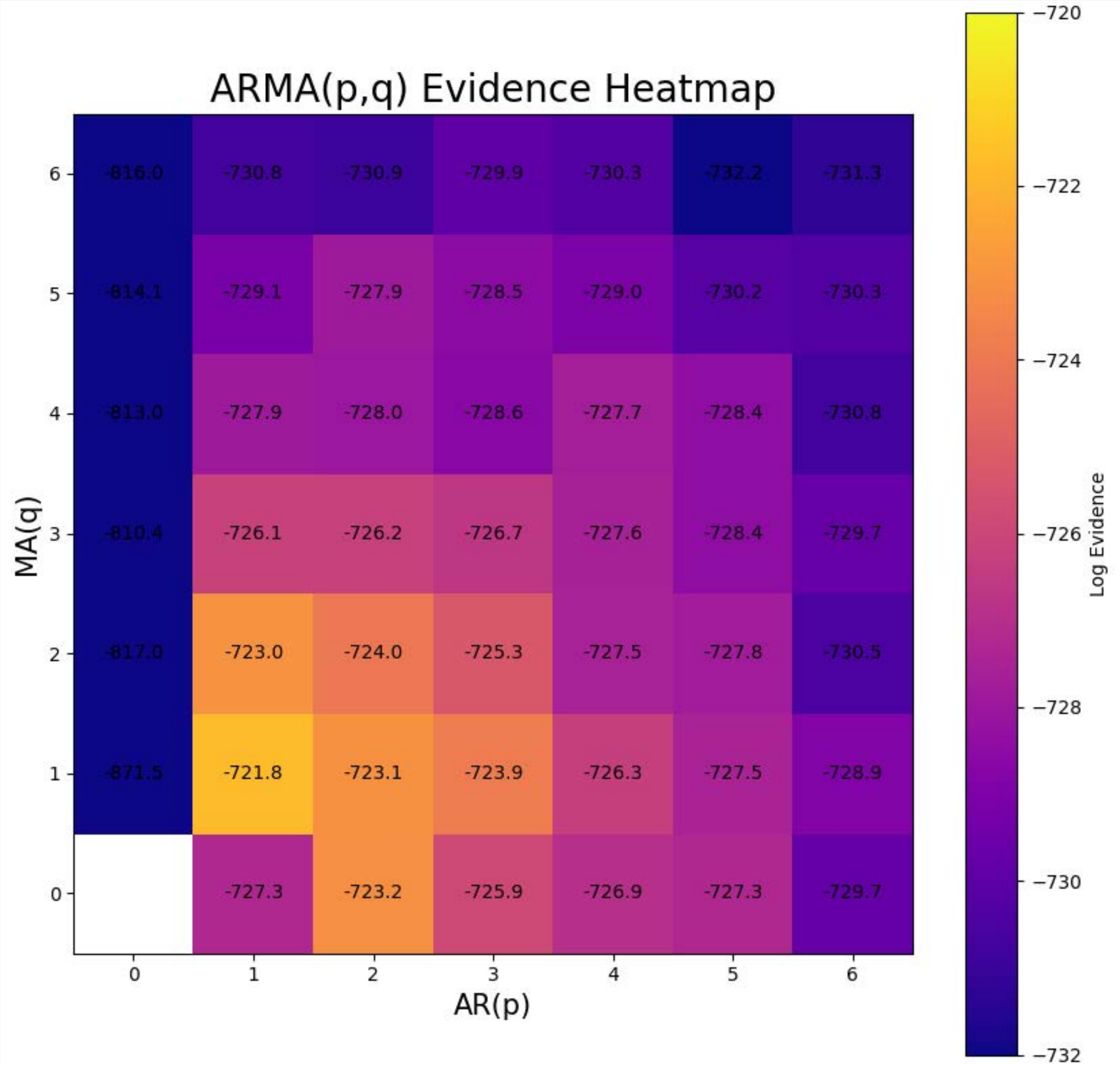
The Occam's Penalty in Action

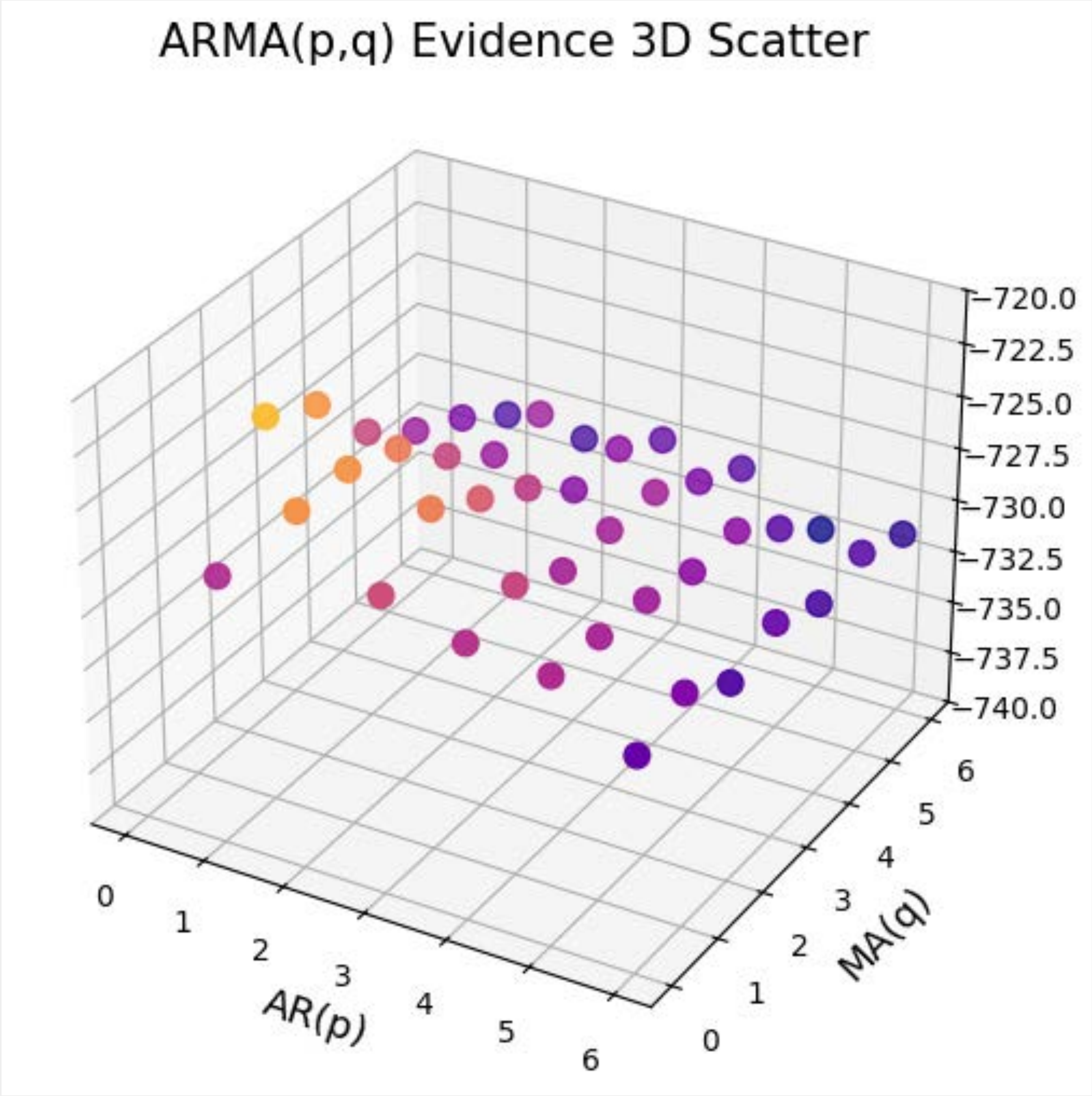
$$Z = \int L(\theta | D)\pi(\theta)d\theta$$



ARIMA x Nested Sampling

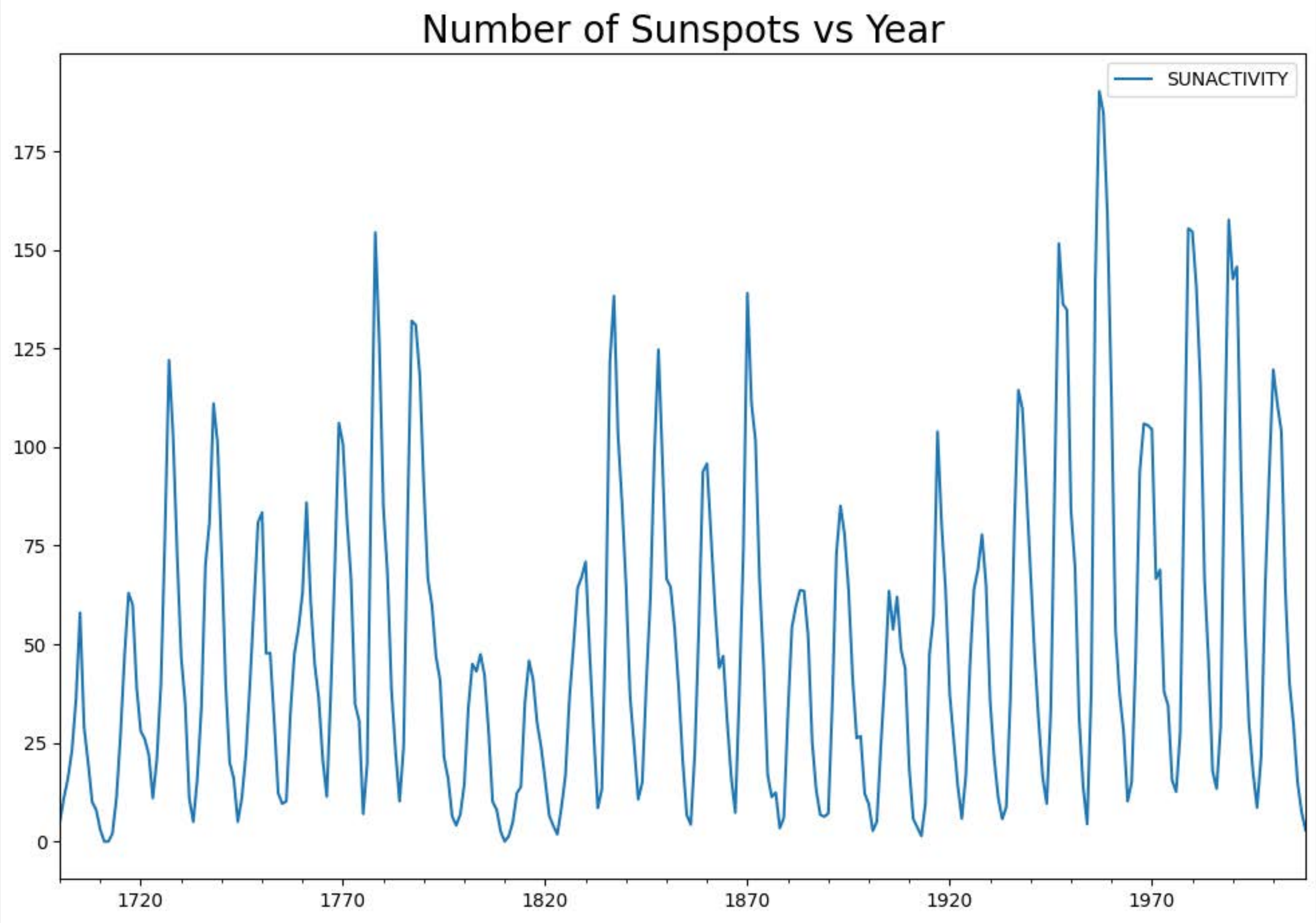
Testing on synthetic data





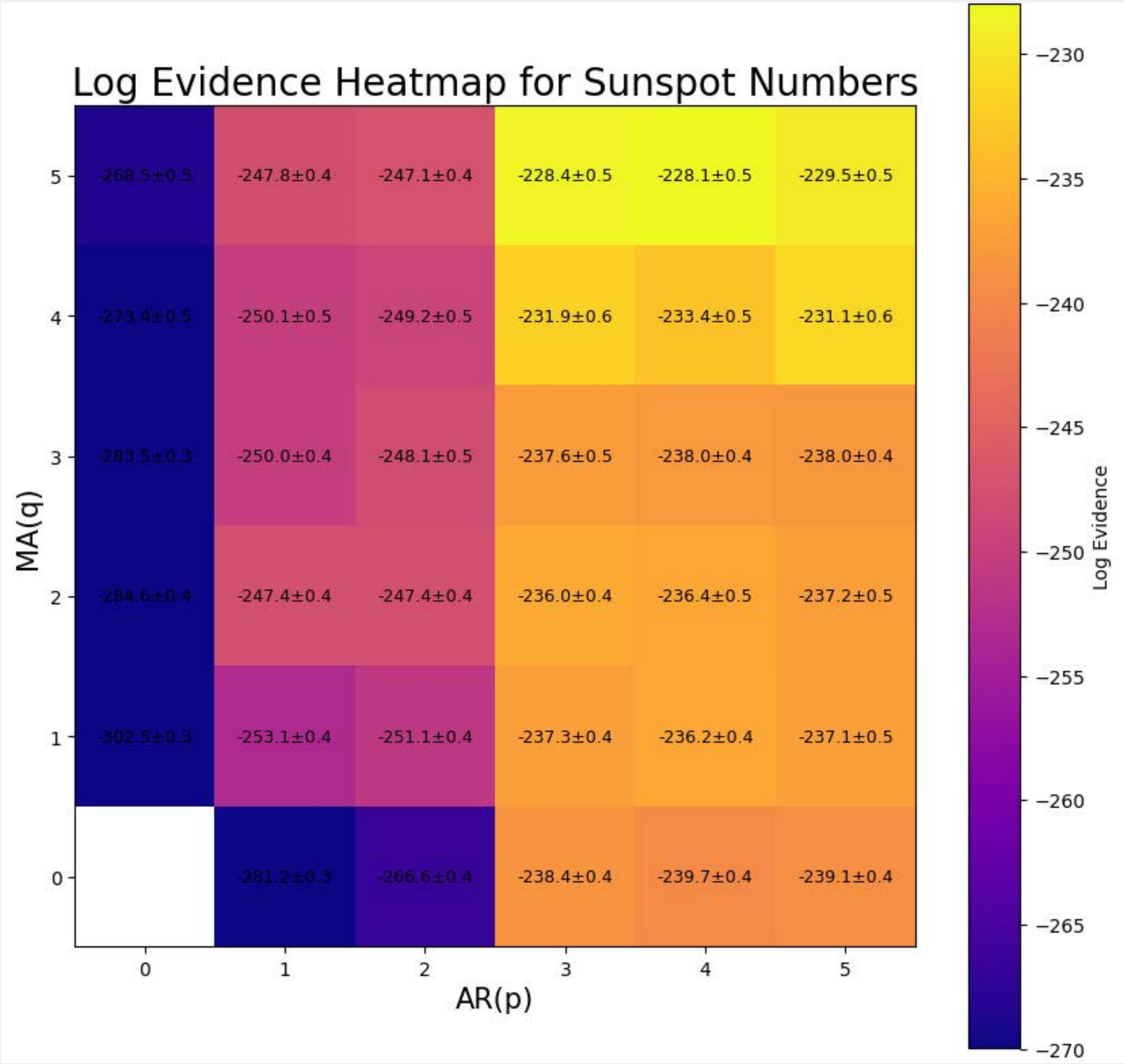
Astronomical Case Study

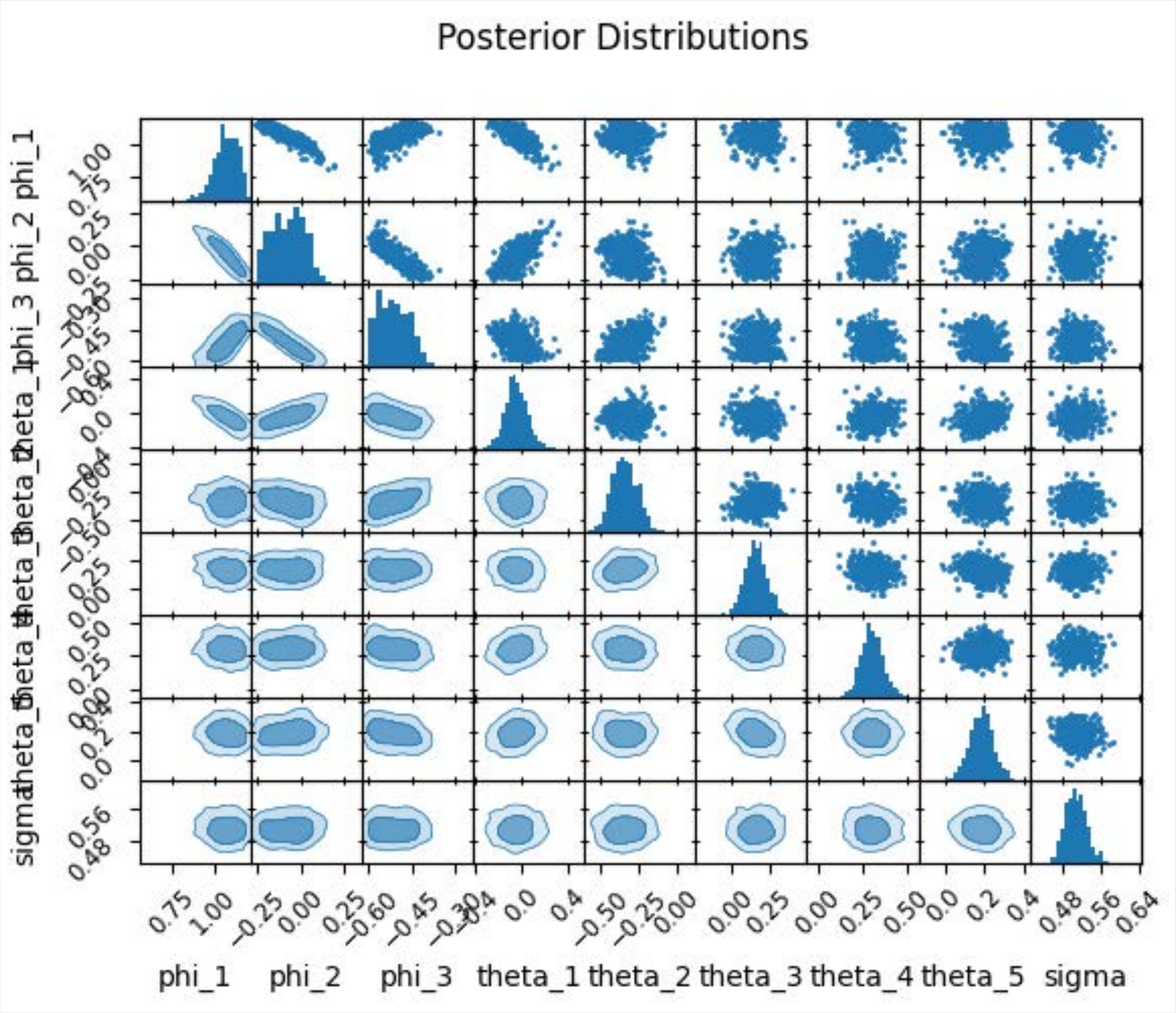
Sunspot Numbers



Astronomical Case Study

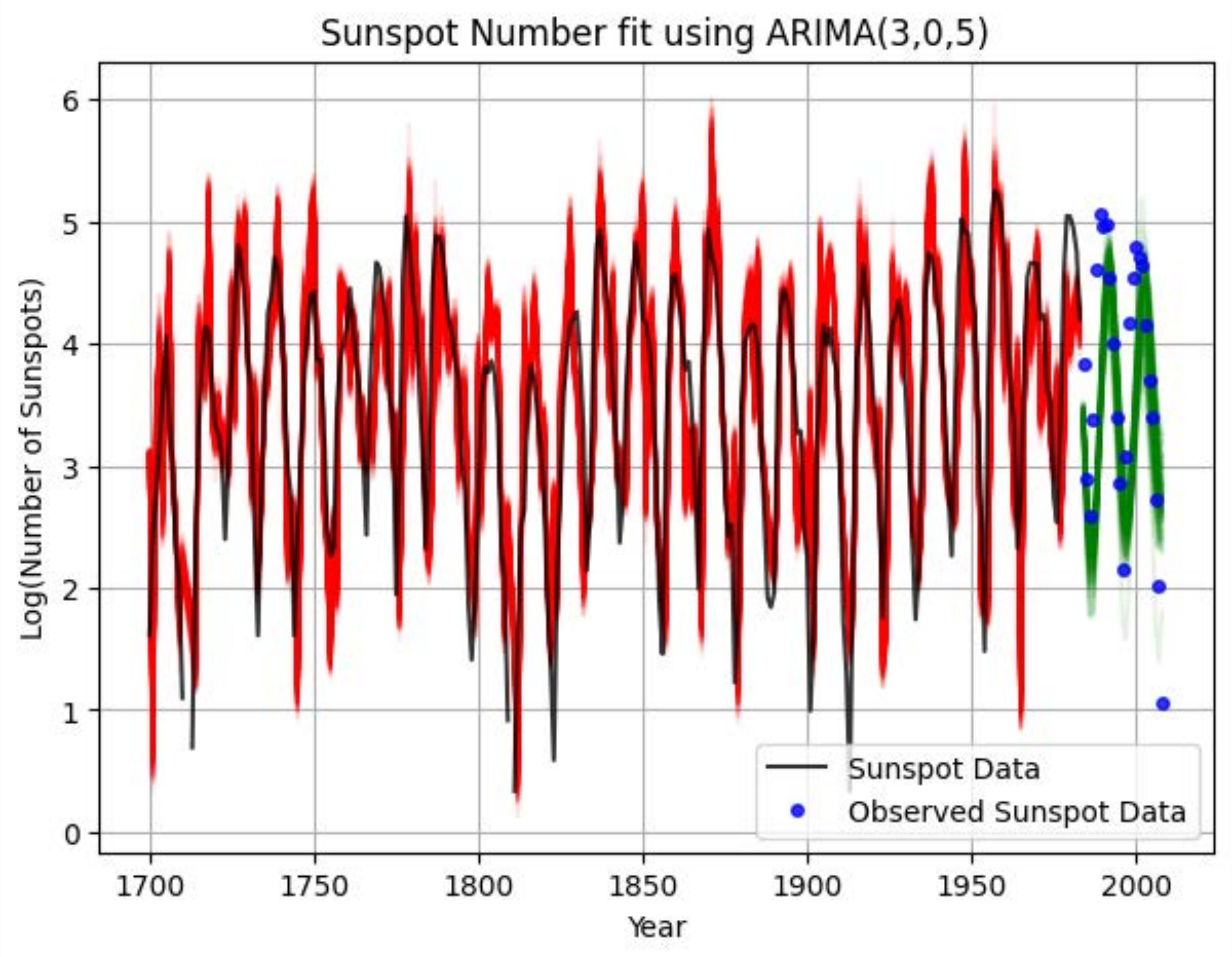
Sunspot Numbers





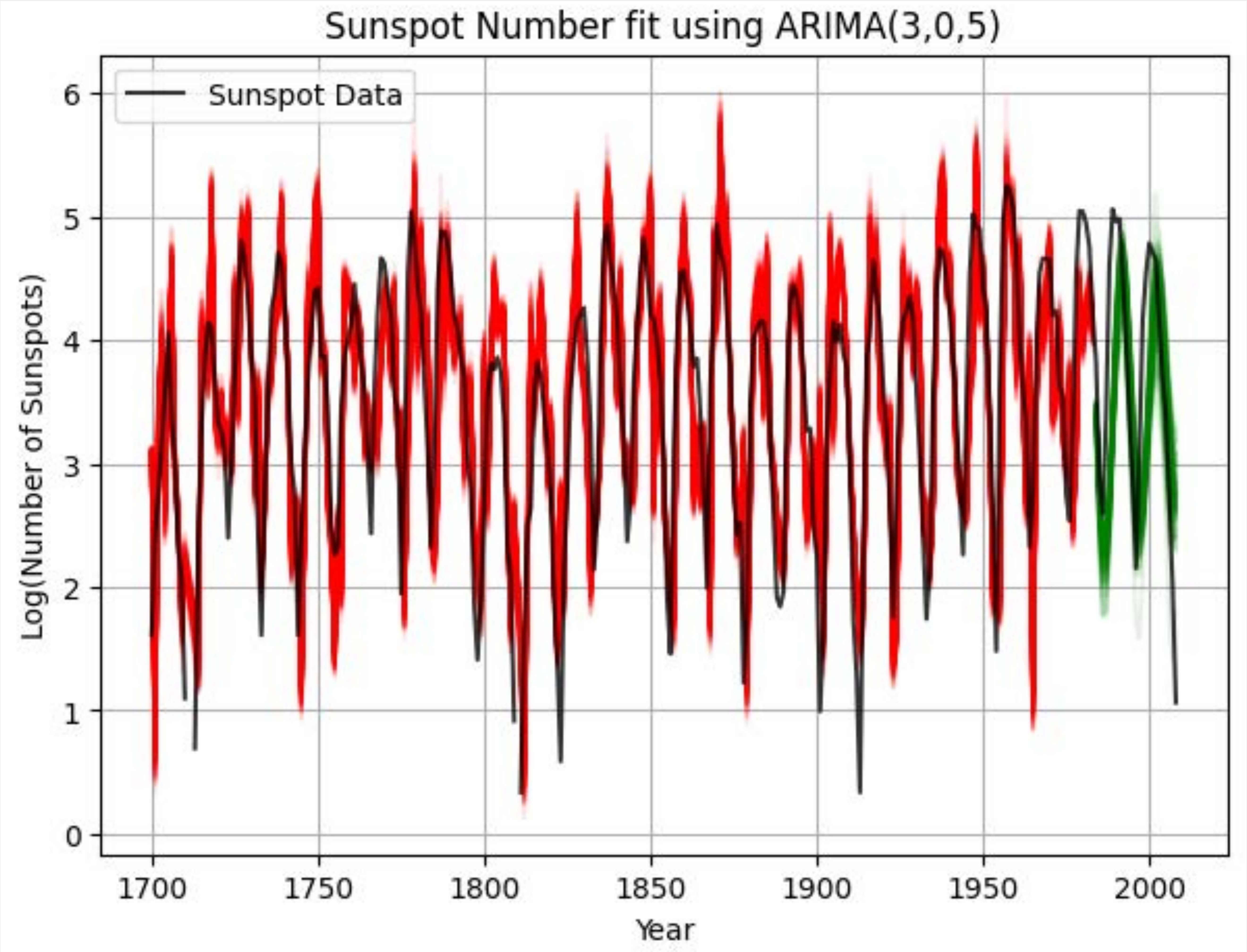
Astronomical Case Study

Sunspot Numbers



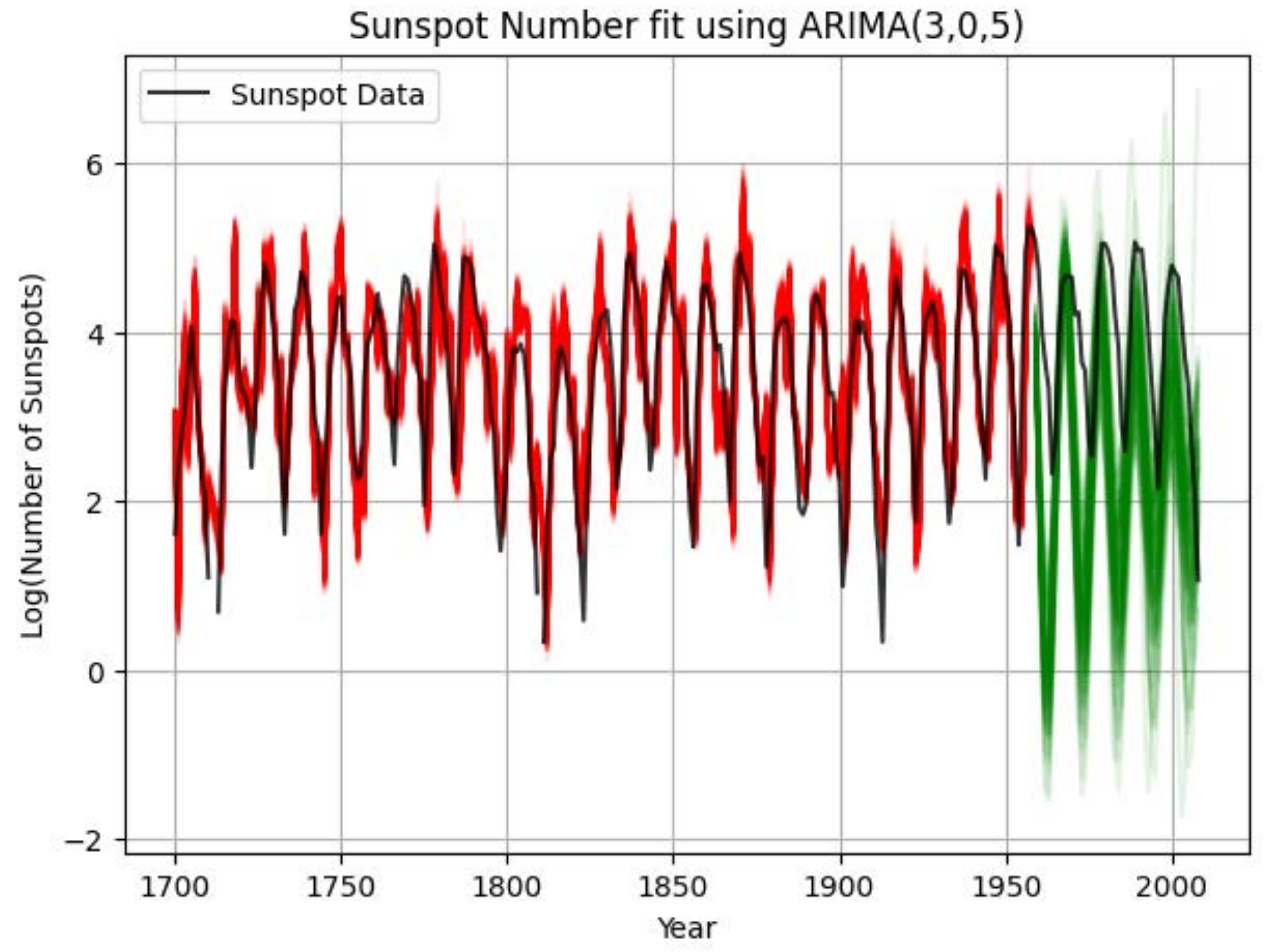
Astronomical Case Study

Sunspot Numbers



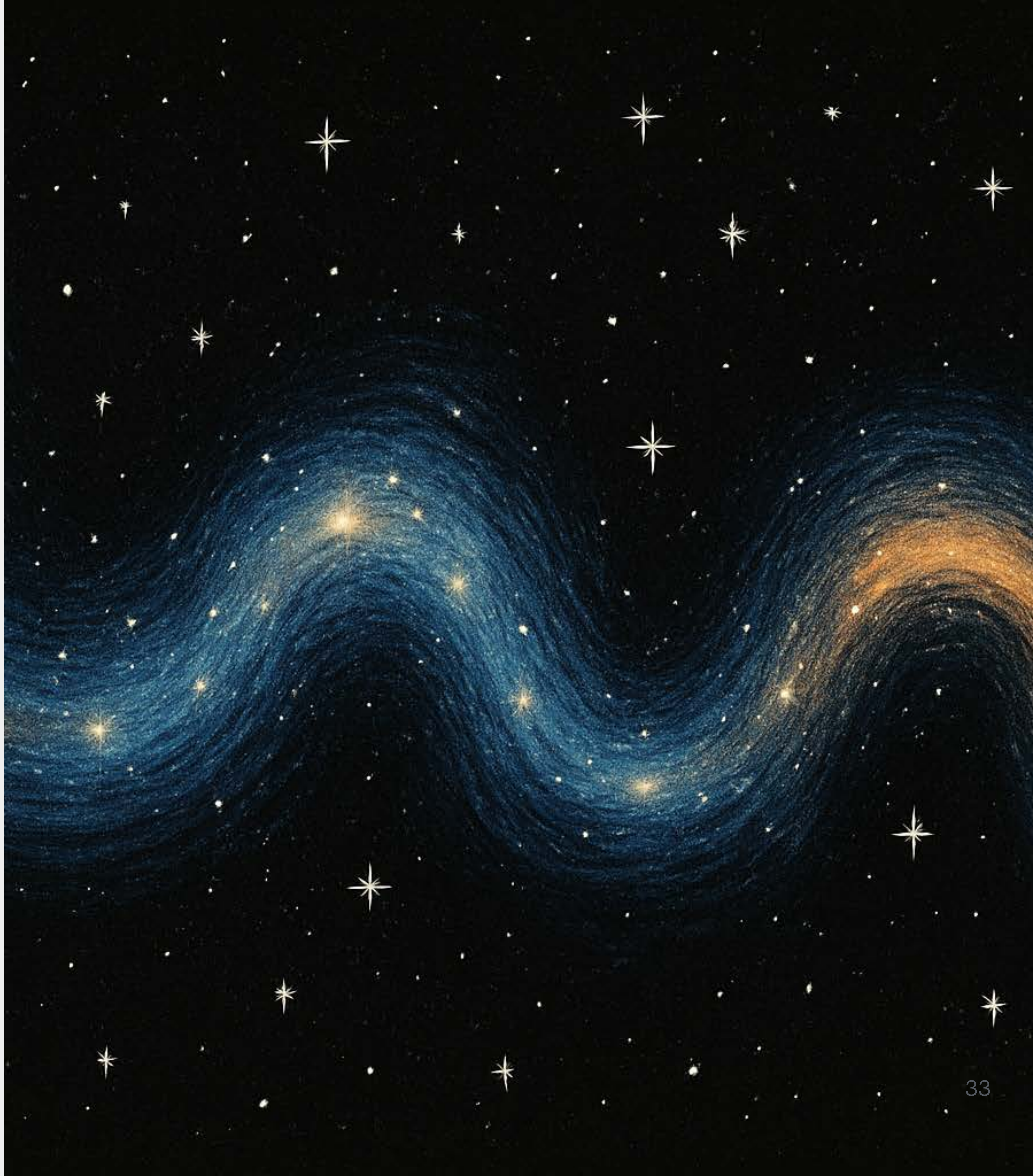
Astronomical Case Study

Sunspot Numbers



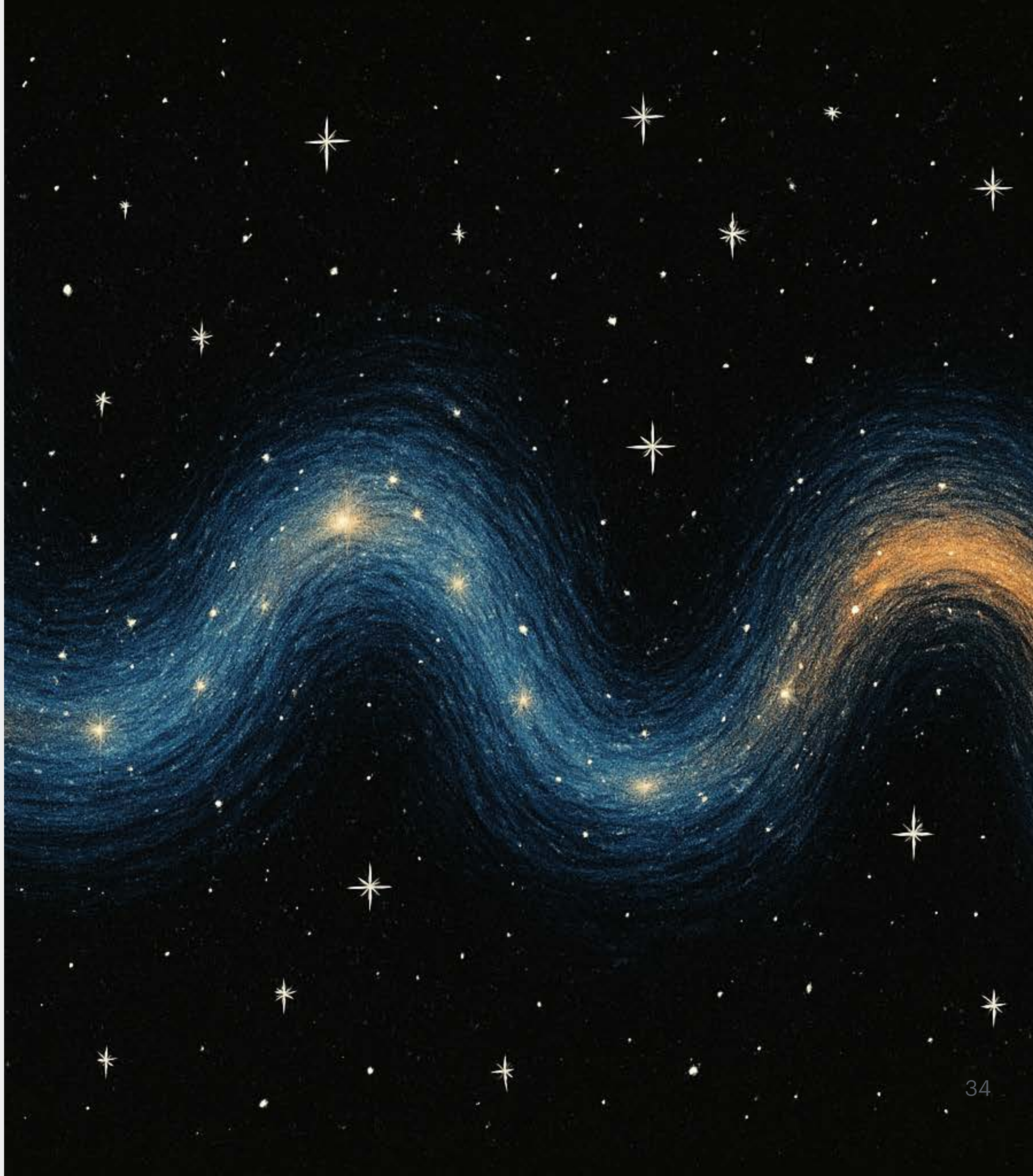
Limitations

- Requires data which is evenly spaced in time.
- Cannot capture long-term, seasonal trends.
- Lack of physical interpretations



Future Prospects

- Extending to other hybrid ARIMA models :Seasonal ARIMA, Continuous ARIMA, and so on.
- Implement on more datasets : AGN and quasar light curves, residual analysis, noise characterisation for gravitational wave data.
- Categorise astronomical datasets on the basis of preferred ARIMA models —> possible physical insights?



Thank You!