

Report on Data Analysis Master Class - Asteroseismology

Ema Salugová & S.M.Masruk Uddin

Contents

1 Introduction

Asteroseismology is a field of studying the frequency oscillations of the photosphere caused by the sound waves travelling in stellar interiors. Observations of flux or radial velocity variations as a results of these oscillations lead to the power spectrum density (PSD) of the time series. Asteroseismology can reveal information on stellar properties such as mass, radius and age. There are several modes that can be excited. Main two categories are radial and non-radial modes. We differentiate them by their spherical harmonic angular degree, l , azimuthal order, m , and a radial order, n . The essential asteroseismologic quantities are the frequency spacing $\delta\nu$, of the mode peaks and the frequency of maximum power, ν_{max} .

Markov Chains Monte Carlo MCMC

Markov Chain Monte Carlo (MCMC) methods have become indispensable tools in modern astronomy, enabling researchers to tackle complex problems involving large parameter spaces and high-dimensional data. At its core, MCMC combines the concepts of Markov chains and Monte Carlo integration to efficiently sample from probability distributions, often in the context of Bayesian inference. This introduction outlines the mathematical foundation of MCMC and highlights its applications in astronomical research.

Bayesian Inference: A Primer

Bayesian inference provides a framework for updating our knowledge about a system as new data becomes available. It is governed by Bayes' theorem [Rees]:

$$P(\theta|\mathbf{D}) = \frac{P(\mathbf{D}|\theta)P(\theta)}{P(\mathbf{D})}, \quad (1)$$

where:

- $P(\theta|\mathbf{D})$ is the posterior probability of the parameters θ given the data \mathbf{D} .
- $P(\mathbf{D}|\theta)$ is the likelihood of observing the data given the parameters.
- $P(\theta)$ is the prior probability, encapsulating prior knowledge about the parameters.

- $P(\mathbf{D})$ is the evidence, a normalization constant.

In astronomy, Bayesian methods are often employed to estimate parameters of stellar models, analyze light curves, or determine cosmological parameters. The posterior distribution, $P(\theta|\mathbf{D})$, is frequently intractable to compute directly due to high-dimensional integrals. This is where MCMC proves invaluable.

Markov Chains and Monte Carlo Sampling

A Markov chain is a stochastic process where the probability of transitioning to a new state depends only on the current state. Formally, if θ_n represents the current state, then the next state θ_{n+1} satisfies:

$$P(\theta_{n+1}|\theta_n, \theta_{n-1}, \dots, \theta_1) = P(\theta_{n+1}|\theta_n). \quad (2)$$

Monte Carlo methods use random sampling to estimate integrals or solve problems that are deterministic in principle. When combined, MCMC generates samples from the posterior distribution by constructing a Markov chain that converges to the desired distribution.

The MCMC Algorithm

The most commonly used MCMC algorithm is the Metropolis-Hastings algorithm. It operates as follows [2015arXiv150401896R]:

1. Initialize the chain at some state θ_0 .
2. For each iteration n :
 - (a) Propose a new state θ^* sampled from a proposal distribution $q(\theta^*|\theta_n)$.
 - (b) Compute the acceptance ratio:

$$r = \frac{P(\mathbf{D}|\theta^*)P(\theta^*)q(\theta_n|\theta^*)}{P(\mathbf{D}|\theta_n)P(\theta_n)q(\theta^*|\theta_n)}. \quad (3)$$

- (c) Accept the proposal with probability $\alpha = \min(1, r)$:
 - If accepted, set $\theta_{n+1} = \theta^*$.
 - Otherwise, set $\theta_{n+1} = \theta_n$.
3. Repeat until convergence.

Log-Likelihood Function

The log-likelihood function quantifies how well a model with a given set of parameters explains the observed data. It measures the agreement between:

- A **theoretical model** defined by parameters.
- The **observed data**.

Higher (less negative) log-likelihood values indicate a better fit between the model and the data, while lower (more negative) values suggest a poor fit.

Mathematical Definition

Given:

- A set of observed data $\mathbf{D} = \{D_1, D_2, \dots, D_N\}$,
- A model $M(\boldsymbol{\theta})$ characterized by parameters $\boldsymbol{\theta}$,
- The likelihood $P(\mathbf{D}|\boldsymbol{\theta})$, which expresses the probability of observing the data given the parameters,

the **log-likelihood** is defined as:

$$\mathcal{L}(\boldsymbol{\theta}) = \ln P(\mathbf{D}|\boldsymbol{\theta}). \quad (4)$$

For independent data points, the likelihood is a product of individual probabilities, and the log-likelihood becomes:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^N \ln P(D_i|\boldsymbol{\theta}). \quad (5)$$

Purpose of the Log-Likelihood Function

The log-likelihood function serves the following purposes:

1. **Measures goodness-of-fit:** It evaluates how well the model explains the observed data.
2. **Guides parameter estimation:** It identifies the parameter set $\boldsymbol{\theta}$ that maximizes the likelihood (i.e., minimizes the mismatch between the model and data).
3. Is essential for **Bayesian inference**, where it contributes to the posterior probability of the parameters.

In simpler terms, it answers the question: *How likely is it that the observed data came from this model with these parameters?*

2 Asteroseismic analysis

We analysed observations of star (effective temperature of $T_{eff} = 4896K$ and log of surface gravity of $logg = 2.53$) taken by the Kepler mission with ID 001723700. The full power density spectrum can be seen in Figure 1.

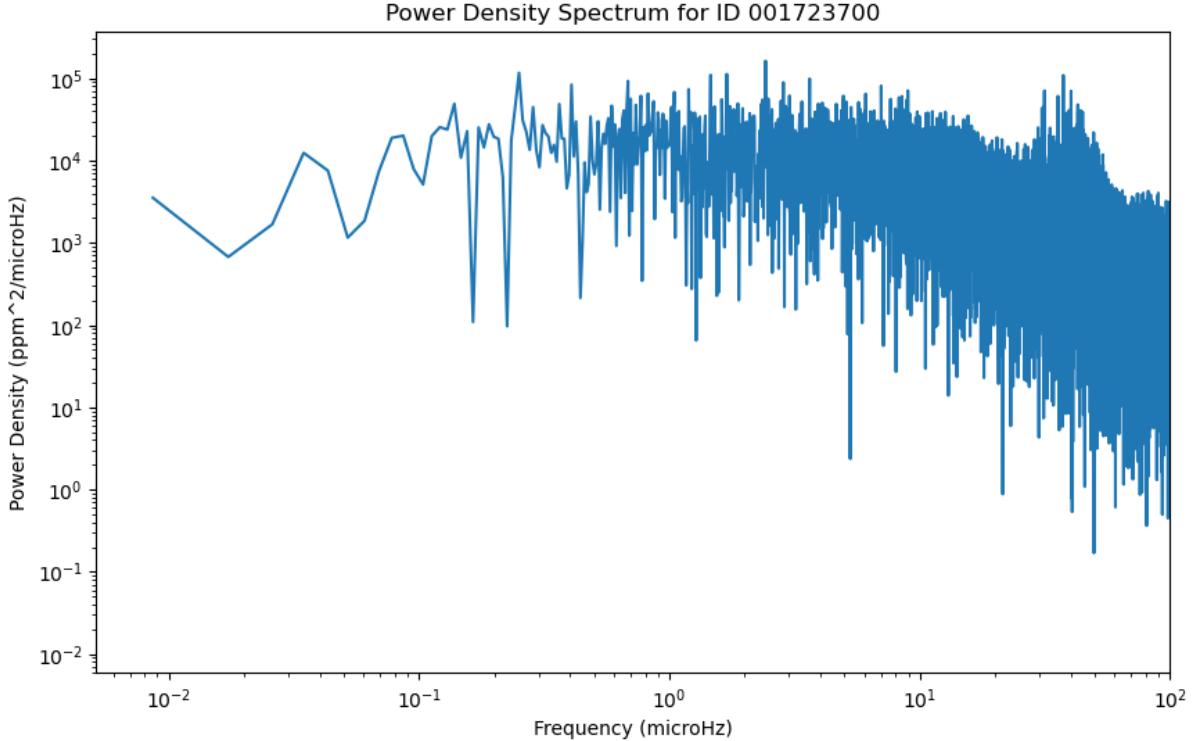


Figure 1: Power spectrum density of a star observed by the Kepler mission

2.1 Oscillation spectrum

The power spectrum density is a super-composition of stellar background, noise and oscillation spectrum. The region on the power spectrum density where the oscillations modes dominate over the noise is called the power excess. This region can be modelled by Gaussian envelope centered on the frequency of the maximum power, ν_{max} .

Stellar background (oscillations coming from the granulation on the surface of the Sun) can be approximated by a power law resulting in a local mean background, $B(\nu)$, around ν_{max} . These all were put into following model:

Power Spectrum Model for the background fit

The power spectrum model combines a background term, a power excess (Gaussian envelope), and constant photon noise:

$$P_{\text{model}}(\nu) = P_{\text{background}}(\nu) + P_{\text{Gaussian}}(\nu) + P_{\text{photon}} \quad (6)$$

where:

- **Stellar Background Term (power law):**

$$P_{\text{background}}(\nu) = B_{\nu_{\text{max}}} \left(\frac{\nu}{\nu_{\text{max}}} \right)^{\beta} \quad (7)$$

with:

- $B_{\nu_{\max}}$: Scaling factor for the background power which is background function at frequency of maximum power ν_{\max} .
- β : Spectral slope.
- ν_{\max} : Frequency of maximum power.

- **Excess Power (Gaussian Peak):**

$$P_{\text{Gaussian}}(\nu) = A \exp \left(-\frac{1}{2} \left(\frac{\nu - \nu_{\max}}{\sigma} \right)^2 \right) \quad (8)$$

with:

- A : Amplitude of the Gaussian peak.
- ν_{\max} : Frequency of maximum power = Central frequency of the peak.
- σ : Standard deviation (width) of the peak.

- **Photon Noise:**

$$P_{\text{photon}}.$$

Thus:

$$P_{\text{model}}(\nu) = B_{\nu_{\max}} \left(\frac{\nu}{\nu_{\max}} \right)^{\beta} + A \exp \left(-\frac{1}{2} \left(\frac{\nu - \nu_{\max}}{\sigma} \right)^2 \right) + P_{\text{photon}}. \quad (9)$$

The model (equation 9) was fitted to the observations using MCMC with algorithm as mentioned in the Introduction section. Six parameters used in a fit are displayed in Table 1. Prior probability ranges were adjusted in the process of finding the best fit with help of corner plots (see the corner plot for MCMC fit of PSD in Appendix A (Figure 7)). The ranges were adjusted in a way to encapsulated the converged normal posterior distribution. The fit plotted with respect to its component and observed data can be seen in Figure 2. We found the frequency of the maximum power to be $\nu_{\max} = 38.15^{+0.30}_{-0.34} \mu\text{Hz}$.

Param. Name	Code Name	Initial Guess	Prior Distr. Range	The best fit value
$B_{\max}(\nu_{\max})$	<code>b_nu_max</code>	1573.04*	[0, 1e6]	$606.64^{+30.68}_{-29.39}$
β	<code>beta</code>	-1.51*	[-3, -1]	$-2.24^{+0.05}_{-0.05}$
A	<code>A</code>	1e4	[0, 1e6]	$3480.68^{+120.22}_{-103.57}$
ν_{\max}	<code>nu_max</code>	39.23*	[35, 45]	$38.15^{+0.30}_{-0.34}$
σ	<code>sigma</code>	5	[0, 10]	$8.54^{+0.29}_{-0.23}$
σ_{photon}	<code>P_photon</code>	1e2	[0, 1e3]	$456.80^{+4.52}_{-3.87}$

Table 1: Parameters for the MCMC fitting of background and power excess. The initial guesses where found by simple analysis of the PSD zoomed in on excess power (except *, where the values were given in the text data file). The second to last column represents the uniform prior probability of parameters. The last column is the best fit value from the MCMC fit.

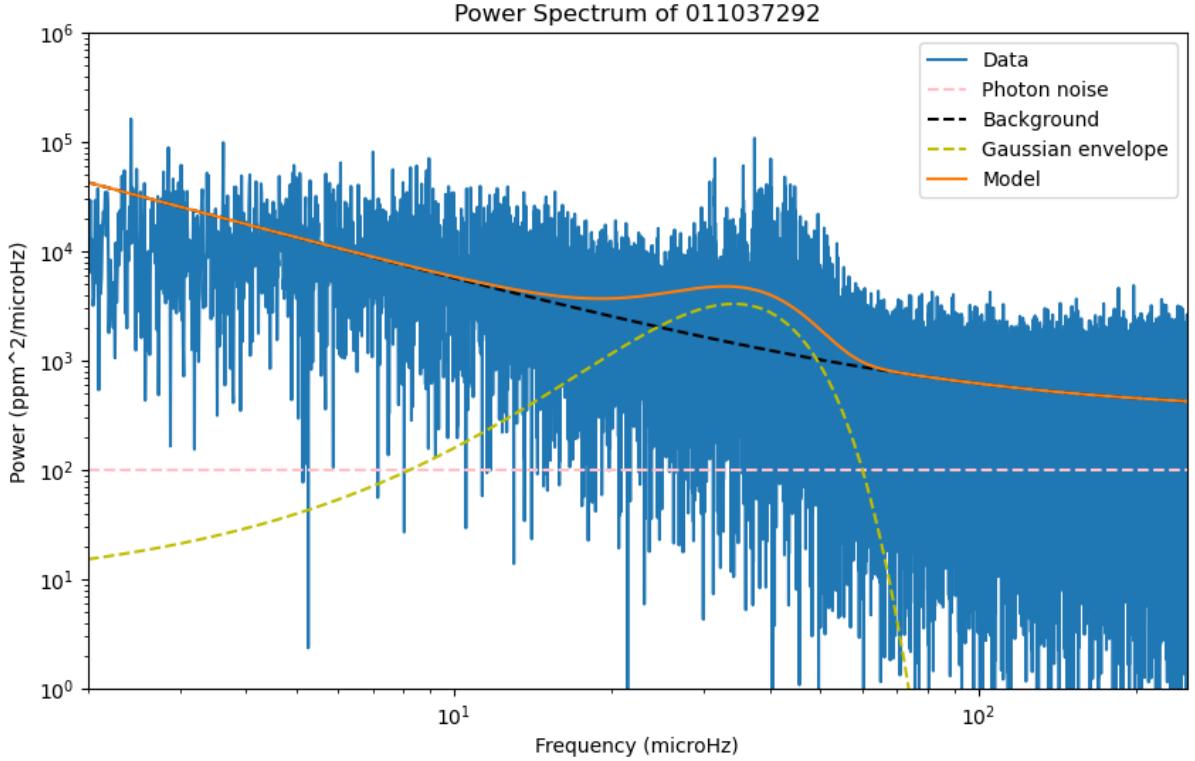


Figure 2: Fit of the global power density spectrum. The pink dashed line is a photon noise, the yellow dashed line is approximation of the local background and the dark dashed line corresponds to the oscillation excess power. The sum of these three components gives the fit of the PSD in orange solid line.

Log-Likelihood in the Provided Code

In the provided code, the log-likelihood is defined as:

$$\mathcal{L}(\boldsymbol{\theta}) = - \sum_i \left[\ln M(\nu_i, \boldsymbol{\theta}) + \frac{P_{\text{obs},i}}{M(\nu_i, \boldsymbol{\theta})} \right],$$

where:

- $M(\nu_i, \boldsymbol{\theta})$: The model power spectrum at frequency ν_i with parameters $\boldsymbol{\theta}$.
- $P_{\text{obs},i}$: The observed power density at ν_i .

This expression penalizes discrepancies between the observed and modeled power spectrum:

- The term $\ln M(\nu_i, \boldsymbol{\theta})$ reflects the logarithmic likelihood of the model.
- The term $\frac{P_{\text{obs},i}}{M(\nu_i, \boldsymbol{\theta})}$ penalizes mismatches between the observed and modeled values.

2.2 Radial pressure modes

In order to isolate the oscillations spectrum, we subtracted the background and noise from the PSD (see the results zoomed in our region of interest around ν_{max} in Figure 3). We performed fit of pure radial pressure modes using the Lorentzian model:

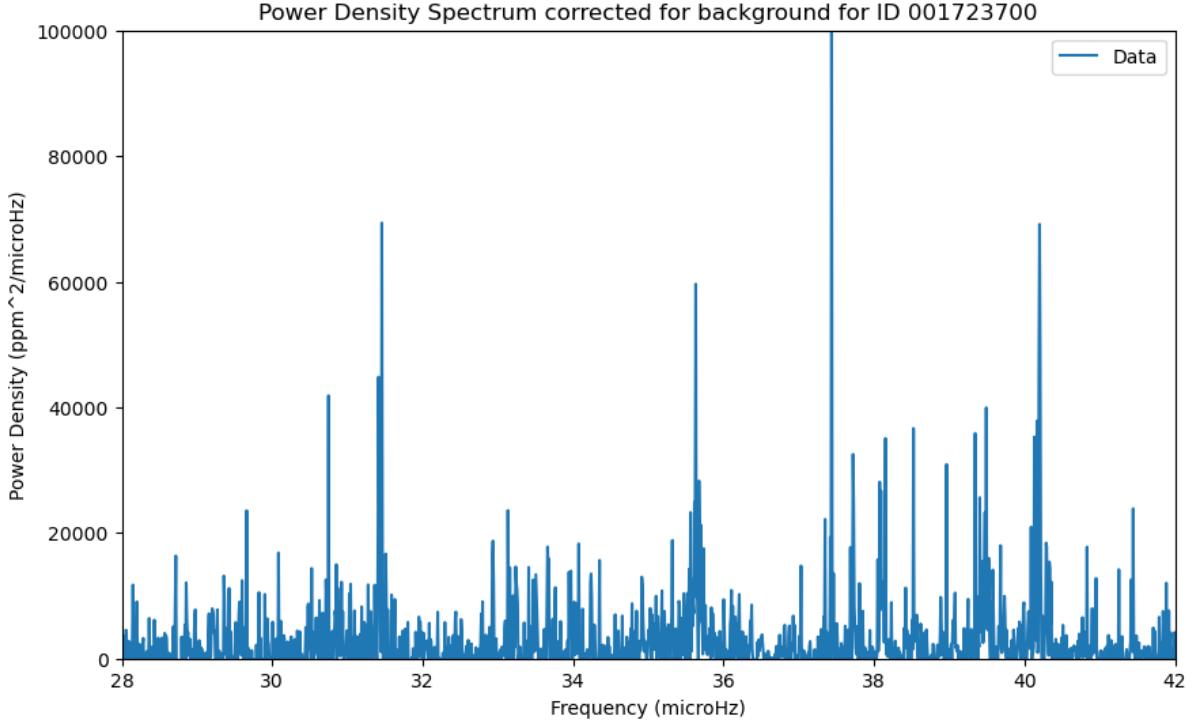


Figure 3: Power spectrum density after a subtraction of background zoomed in our region of further interest.

Lorentzian Model

The Lorentzian model represents the sum of two Lorentzian functions at different frequency centers (we dropped m in each equations as we considered it to be 0):

$$L_{\text{model}}(\nu) = \sum_n \sum_{l_0}^{l_{\max}} L_{n,l}(\nu) \quad (10)$$

where each Lorentzian is defined as:

$$L_{n,l}(\nu) = \frac{H_{n,l}}{1 + 4 \left(\frac{\nu - \nu_{\text{center},n,l}}{\Gamma_{n,l}} \right)^2} \quad (11)$$

where

- $H_{n,l}$ is height of the Lorentzian peak.
- $\Gamma_{n,l}$ is line-width of the Lorentzian peak (half width at half-maximum).
- $\nu_{\text{center},n,l}$ is a central frequency of the peak.

Central frequency of the peak is defined as

$$\nu_{\text{center},n,l} = \left(n + \frac{l}{2} + \epsilon + d_{0l} + \frac{\alpha}{2}(n - n_{\max})^2 \right) \Delta\nu \quad (12)$$

where:

- $n_{\max} = \frac{\nu_{\max}}{\Delta\nu} - \epsilon$ is a central radial order.
- $\Delta\nu$ is a large frequency spacing.
- ϵ is a phase constant.
- α is the curvature term on $\Delta\nu$.

The radial order n have to be considered around the value n_{\max} in equation 10 and only degrees $l < l_{\max} = 3$ are present in oscillation spectra obtained by photometric observations.

At first, we determine the n_{\max} with ϵ , $\Delta\nu$ and ν_{\max} to be $n_{\max} = 7.54$. As this is not integer, we took our radial order to be $n = 8$ (we also tried $n = 7$ but the fit was unsuccessful). We fit for angular degrees $l = 0, 1, 2$. However, the fit was not successful with $l = 1$. This mode $L_{8,1}$ could not be resolved and hence it was excluded from our final fit and was not used in further analysis as well. Parameters used in a fit are displayed in Table 2, 5 of the parameters were free while others were kept constant (ϵ , $\Delta\nu$ and ν_{\max} were assumed to be constant). Prior probability ranges were adjusted in the process of finding the best fit with help of corner plots (see the corner plot for MCMC Lorentzian fit of pressure modes in Appendix B or Figure 8). The ranges were adjusted in a way to encapsulated the converged normal posterior distribution. The Lorentzian fit plotted with respect to the observed pressure modes can be seen in Figure 4.

Param. Name	Code Name	Initial Guess	Prior Distr. Range	The best fit value
$H_{8,0}$	H1	5e4	[1e4, 6e4]	25613.70
$H_{8,2}$	H2	2e4	[1e2, 4e4]	8719.84
$\Gamma_{8,0}$	Gamma1	0.25	[0.01, 1]	0.14
$\Gamma_{8,2}$	Gamma2	0.25	[0.01, 1]	0.47
d_{0l}	d_0l	-0.15	[-1, -0.01]	-0.16
ϵ	eps	0.961	constant	—
α	alpha	0.0098	constant	—
$\Delta\nu$	delta_nu	4.483	constant	—

Table 2: Parameters for the MCMC fitting of Lorentzian fit to pressure modes. The initial guesses were found by simple analysis pressure modes. The second to last column represents the uniform prior probability of parameters. The last column is the best fit value from the MCMC fit. The constant values were given in data file. To see the uncertainty on the fitted parameters, see corner plot in Appendix B

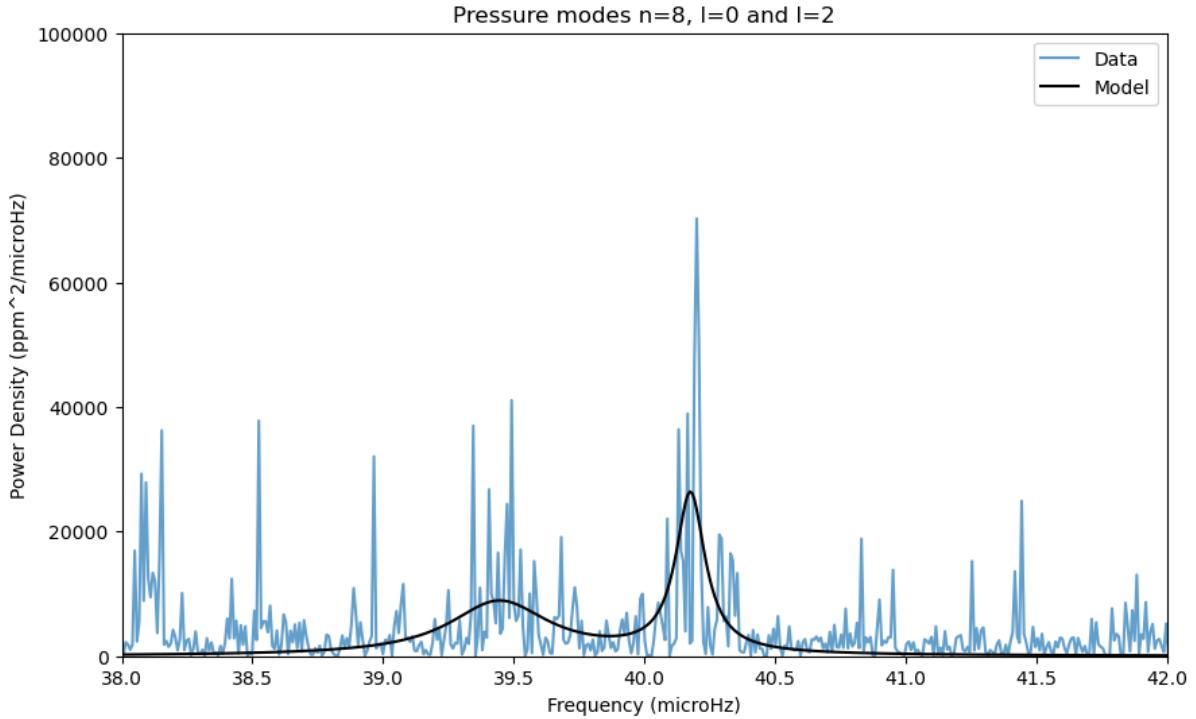


Figure 4: Fit of the pressure mode with radial order $n=8$ and angular degree $l=0$ and $l=2$. Pressure mode $L_{8,0}$ is the higher peak.

To get the value of the $\Delta\nu$, we fitted another radial orders ($n = 7, 8$), each for $l = 0, 2$. In this case, we have assumed only α to be constant (taken from the data file) and we had 10 free parameters. This doubled the free parameters, hence, the parameters, initial guesses, priors and results can be found in the appendix C. We obtained $\Delta\nu = 4.57 \pm 0.03 \mu\text{Hz}$. The Lorentzian fit plotted with respect to the observed pressure modes can be seen in Figure 5.

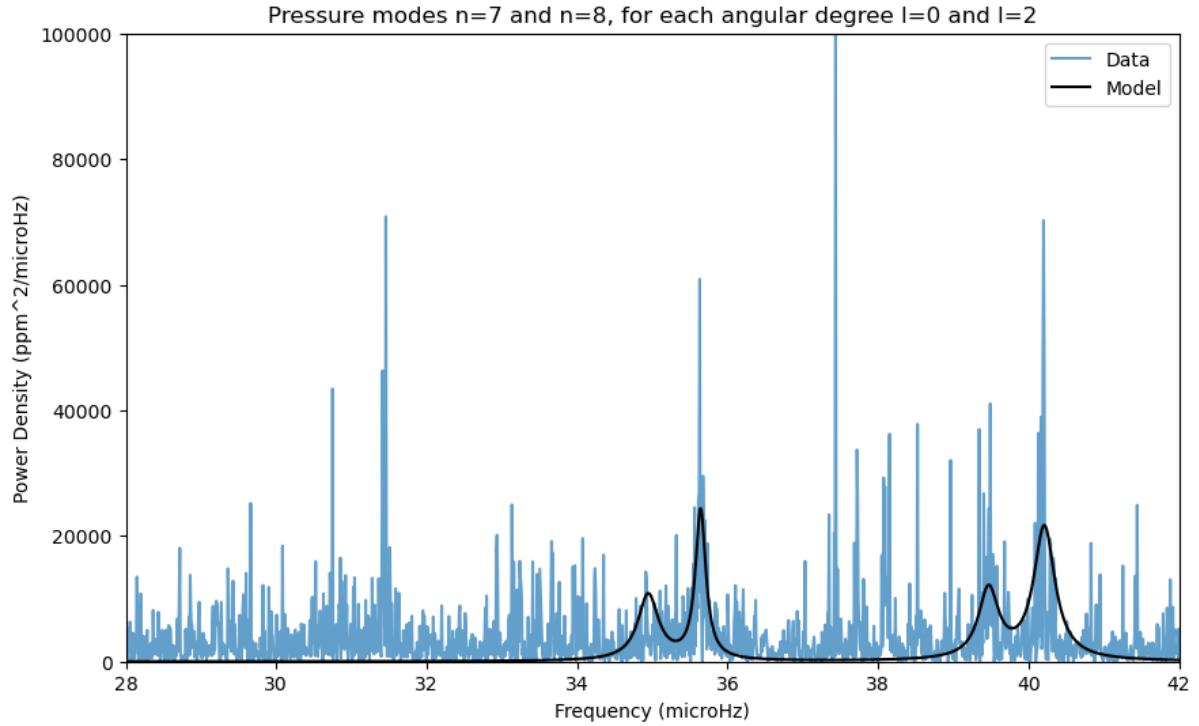


Figure 5: Fit of the pressure mode with radial order $n=7$ and $n=8$ and angular degree $l=0$ and $l=2$. Pressure mode $L_{8,0}$ is the higher peak on the right. Pressure mode $L_{7,0}$ is the higher peak on the left radial mode.

Then, we fitted the three pressure modes of radial degree $n = 6, 7, 8$ and for each radial degree, two angular degrees $l = 0, 2$. Here, we kept constant $\nu_{max} = 38.1443\mu Hz$ and we have 22 free parameters which increased the computational time significantly. We found $\alpha = 0.03 \pm 0.004$ and $\Delta\nu = 4.48 \pm 0.01\mu Hz$. The final fit can be found in Figure 6. The corner plot is under appendix D (or Figure 10).

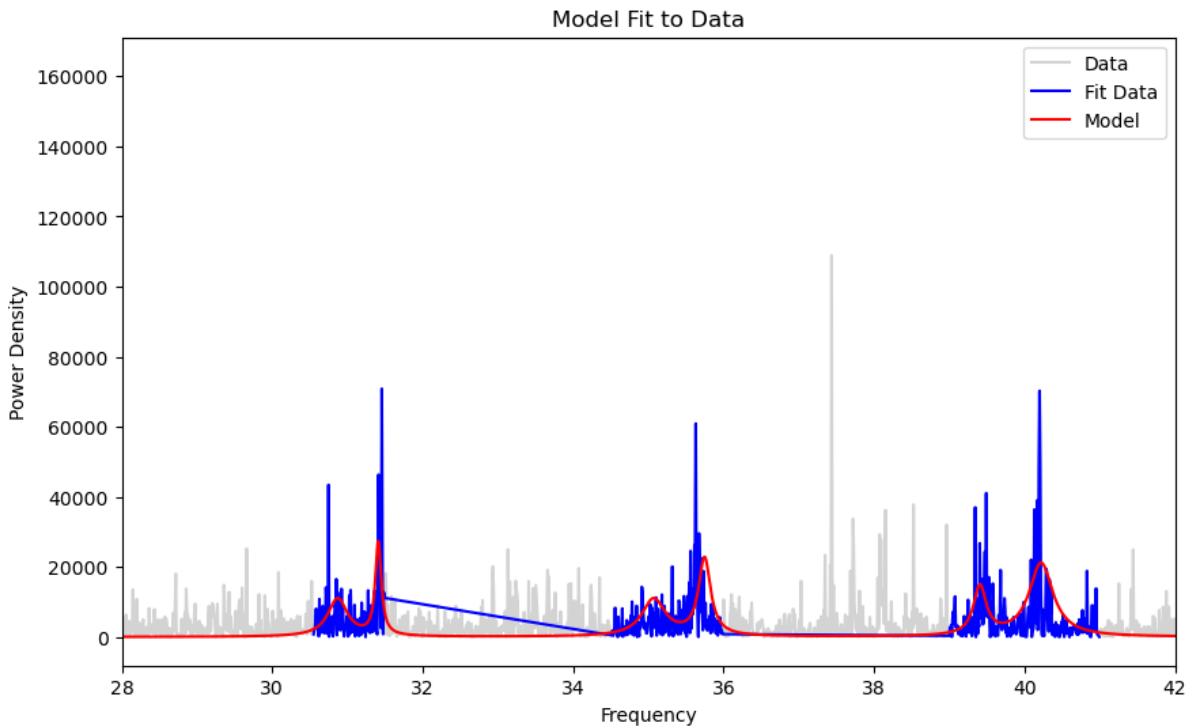


Figure 6: Fit of the pressure modes with radial order $n=6,7,8$ and angular degree $l=0$ and $l=2$ (in red). In blue are the ranges of interest containing the pressure modes, in gray the data. Radial order are from left to right increasing (6, 7, 8). In the radial orders the higher peaks are $l=0$, lower height $l=2$. The frequency position difference of successive maximum peaks is $\Delta\nu$

The formulas we have used here to fit the parameters are taken from the note on Seismic Analysis [**Benoit**]; provide in the Data Analysis Master class.

A Corner plot for one radial order

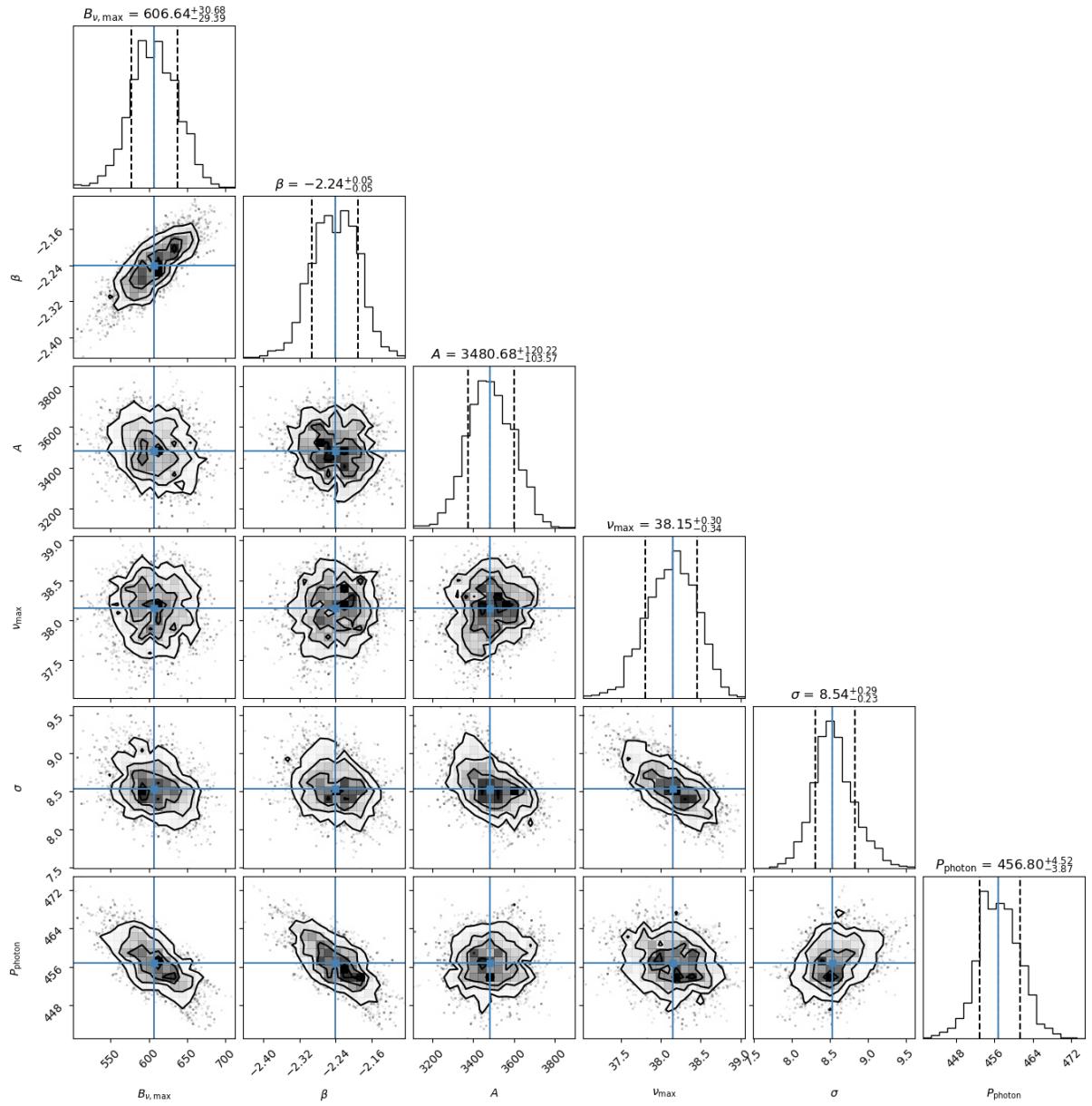


Figure 7: Corner plot displaying the posterior distributions of fitted parameters

B Corner plot for singel radial order

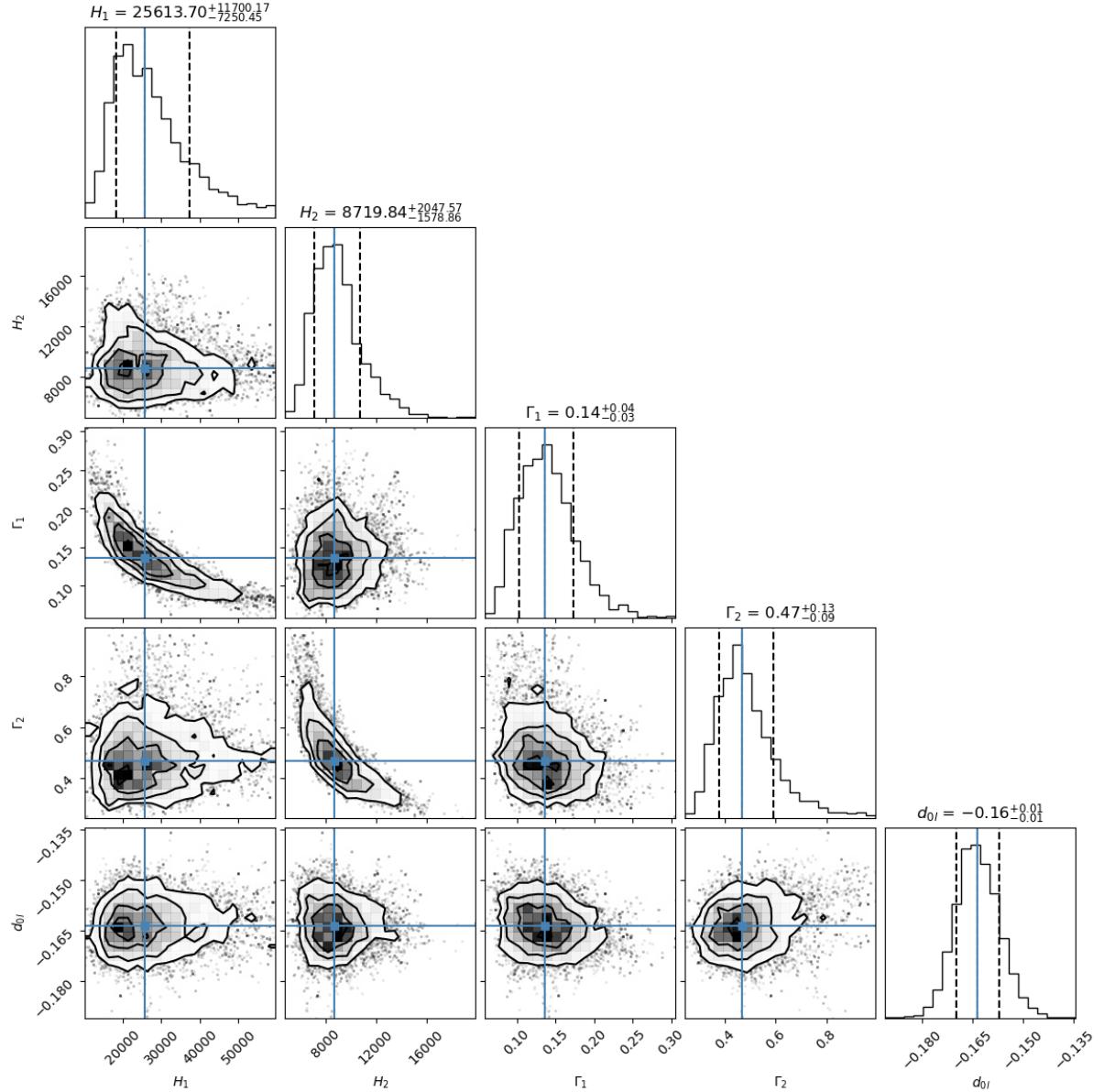


Figure 8: Corner plot displaying the posterior distributions of fitted parameters. Here H_1 , Γ_1 corresponds to $H_{n,l} = H_{8,0}$, $\Gamma_{8,0}$ and H_2 , Γ_2 corresponds to $H_{n,l} = H_{8,2}$, $\Gamma_{8,2}$

C Corner plot for two radial orders

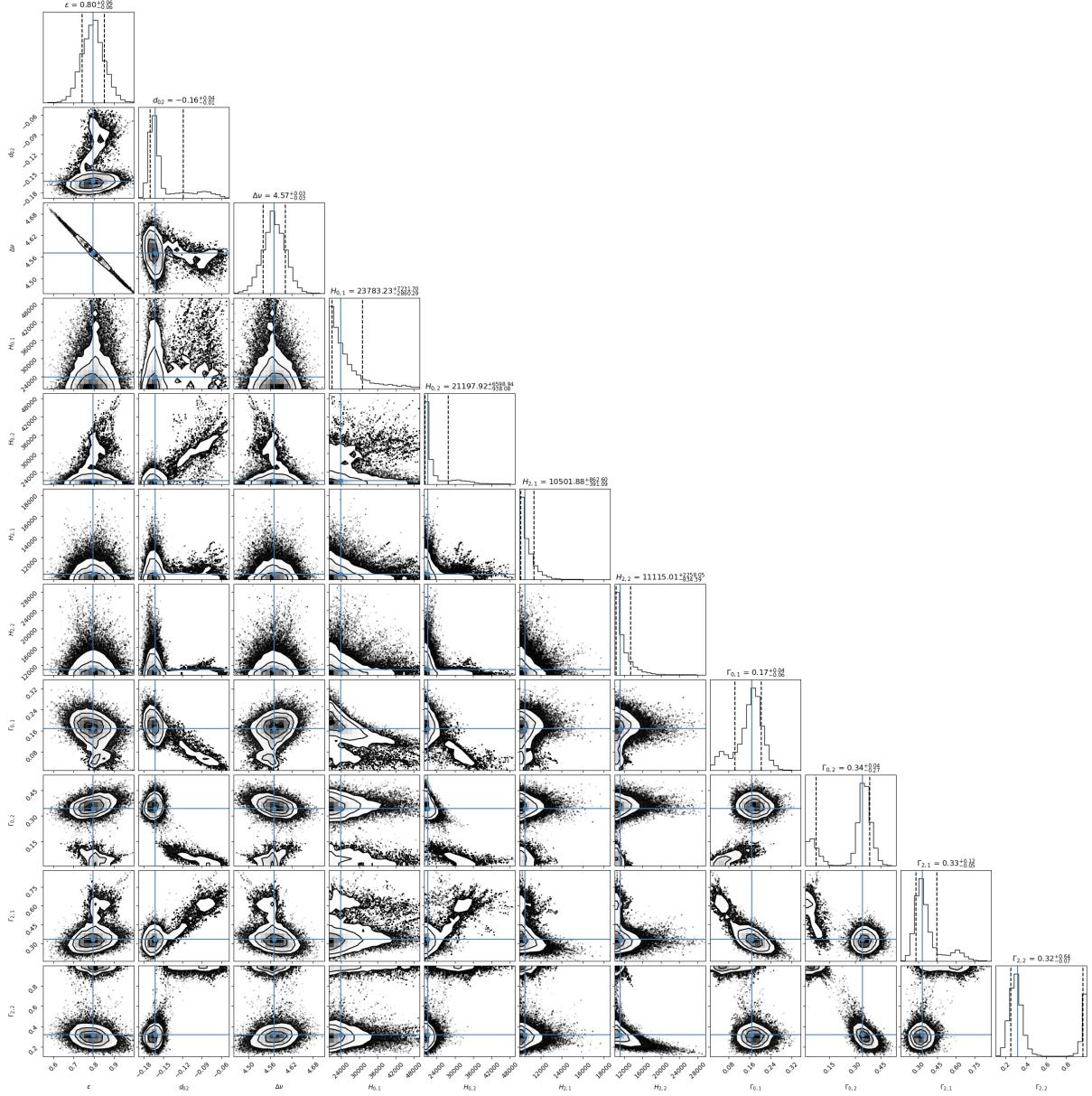


Figure 9: Corner plot displaying the posterior distributions of fitted parameters. Here the first index 0,1 are for $n=7,8$ and the second index 0,1 are $l=0,2$.

D Corner plot for three radial orders

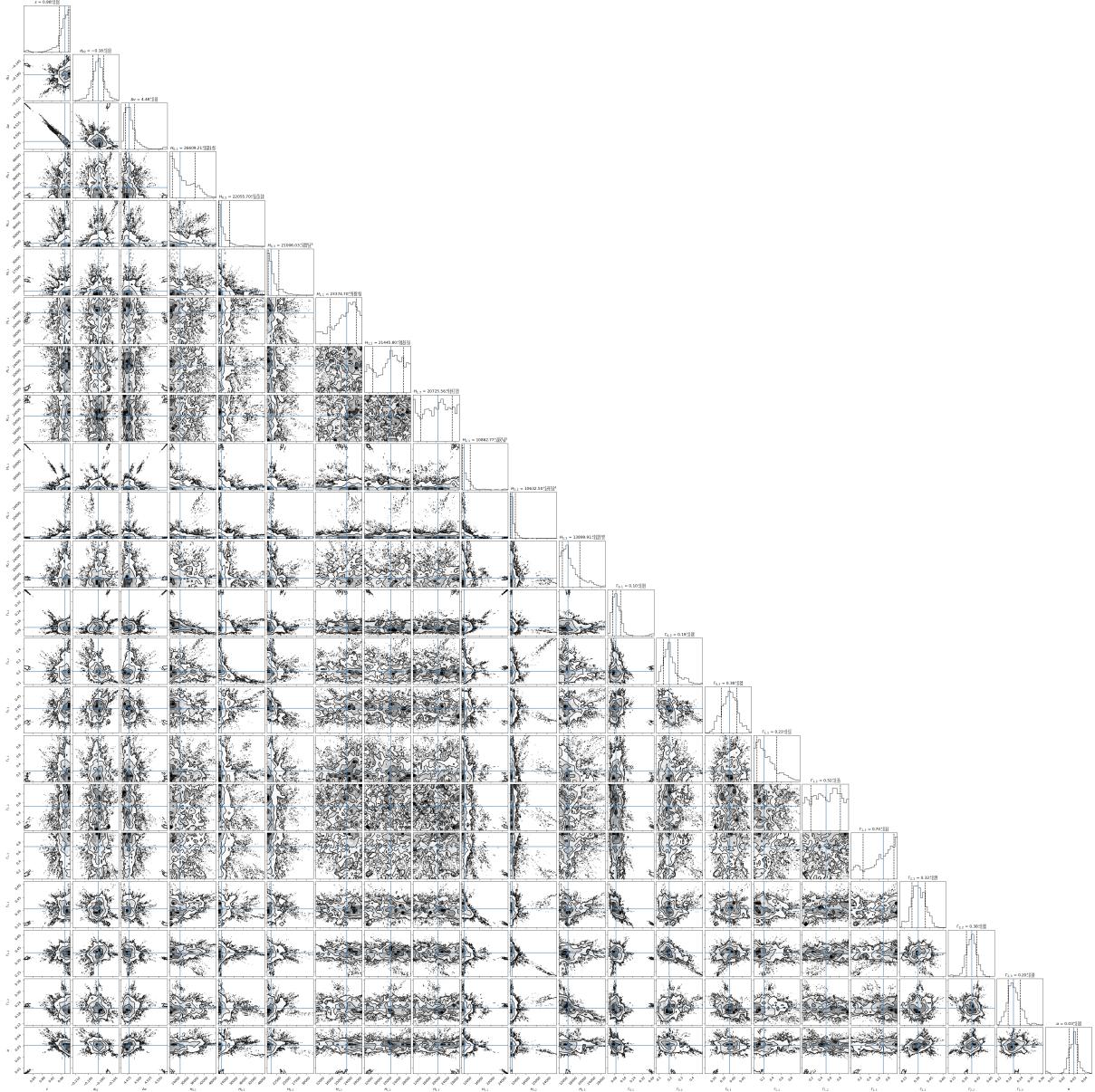


Figure 10: Corner plot displaying the posterior distributions of fitted parameters. Here the first index 0,1,2 are for $n=6,7,8$ and the second index 0,1 are $l=0,2$.