Assignment3:
Cross-Validation , MCMC , combined models

Salar Ghaderi

The Python codes used in this report can be found at the following link: GitHub Repository.

1 Q1

The scientific question is: *How many harmonic signals do the light curves contain?* For simplicity, we consider only signals at integer multiples of Earth's sidereal frequency. The goal is to select the best model order (i.e. the number of harmonics, k) using a predictive approach (via cross-validation) and compare that with the classical BIC method.

Model and Methods

1.0.1 Harmonic Model

We assume that each light curve can be modeled as:

$$y(t) = \mu + \sum_{j=1}^{k} \Big[a_j \cos(j\,\omega\,t) + b_j \sin(j\,\omega\,t) \Big] + \epsilon(t),$$

where:

- μ is the constant (mean brightness).
- ω is the base angular frequency (from Earth's sidereal period).
- Each harmonic j introduces two parameters: a_j and b_j.
- Thus, for a model with k harmonics, the total number of free parameters is

$$p = 1 + 2k.$$

1.0.2 Weighted Least Squares Fitting

Given the inverse variances (assumed to be $1/\sigma^2$), we perform a weighted least-squares fit. The design matrix X is constructed with:

- A column of ones (for μ).
- For each harmonic j, columns for $\cos(j\,\omega\,t)$ and $\sin(j\,\omega\,t)$.

1.0.3 Cross-Validation Approaches

We use two CV methods:

1. **5-Fold CV:** The data is split into 5 folds. In each iteration, one fold is held out for validation and the model is trained on the remaining data. The predictive log-likelihood on the validation set is computed and summed over folds.
2. **Leave-One-Out CV (LOOCV) I tested this also to see any furthure improvments):** Each data point is held out once while the model is fit on the remaining $N-1$ points. The log-likelihood for the left-out point is computed and summed over all points.

For Gaussian noise, the log-likelihood (up to an additive constant) for a single validation point is approximated by

$$\log L \approx -\frac{1}{2}\,\mathrm{ivar}\,(y - y_{\mathrm{pred}})^2.$$

The model order with the highest total CV log-likelihood is selected.

1.0.4 BIC

The Bayesian Information Criterion is computed as:

$$\mathrm{BIC} = \chi^2 + p\,\ln(N),$$

where χ^2 is the weighted sum of squared residuals and $p = 1 + 2k$. BIC penalizes extra parameters more heavily, so in borderline cases the preferred model order may be lower than that suggested by CV.

subResults and Discussion For each light curve, the CV analysis was performed for model orders $k = 0, 1, 2, 3$ (i.e. with $p = 1, 3, 5, 7$ parameters). Here is a summary of the findings:

5-Fold CV Results (Total CV Log-Likelihood):

- Light curve 0: Best model order $k = 0$ (Total CV logL ≈ -510.90)
- Light curve 1: Best model order $k = 1$ (Total CV logL ≈ -501.48)
- Light curve 2: Best model order $k = 2$ (Total CV logL ≈ -523.52)
- Light curve 3: Best model order $k = 1$ (Total CV logL ≈ -491.18)
- Light curve 4: Best model order $k = 0$ (Total CV logL ≈ -617.13)
- Light curve 5: Best model order $k = 1$ (Total CV logL ≈ -640.10)
- Light curve 6: Best model order $k = 3$ (Total CV logL ≈ -551.38)
- Light curve 7: Best model order $k = 3$ (Total CV logL ≈ -507.18)

LOOCV Results (Total LOOCV Log-Likelihood): They are very similar to the 5-fold CV values, with the best model identical for each light curve.

BIC Results for Light Curve 2: For light curve 2 the BIC analysis yielded same best model selection except for case 2 with details:

- $k = 0$: BIC ≈ 1604.27
- $k = 1$: BIC ≈ 1066.51
- $k = 2$: BIC ≈ 1071.85
- $k = 3$: BIC ≈ 1083.69

Thus, BIC selects $k = 1$ harmonic, favoring a simpler model.

Discussion

- **CV vs. BIC: vs the main code that generated the data** The CV methods (both 5-fold and LOOCV) provided exactly same model selection as BIC except case number 2. In the case of light curve 2, they slightly favor a 2-harmonic model because the extra harmonic provides a marginal gain in predictive performance. However, BIC, which penalizes additional parameters by a factor of $\ln(N)$, selects the 1-harmonic model. COmparing to the main data generating code BIC and CV provide true results , except for case 2 that BIC is true.This difference is expected when the signal is modest (here, an amplitude of ~ 0.3) and the gain in prediction from an extra parameter is very small.so in such cases BIC criterion is better at this problems.so in borderline cases BIC prefers the less complex model.

2 q2

2.1

I analyze a set of astronomical light curves to determine whether they contain a harmonic signal at Earth's sidereal period. Two models are considered:

1. A **three-parameter model** where the data are modeled as
$$y(t) = \mu + a\cos(\omega t) + b\sin(\omega t) + \epsilon(t),$$
with the frequency ω fixed to the expected Earth sidereal value.
2. A **four-parameter model** where the frequency is allowed to vary. In this case,
$$y(t) = \mu + a\cos(\omega t) + b\sin(\omega t) + \epsilon(t),$$
and we assign a prior to ω so that it is constrained to lie in the interval $[0.9\,\omega_0,\ 1.1\,\omega_0]$ (with ω_0 being the expected sidereal frequency).

Uniform priors are imposed:

- $\mu \sim \mathcal{U}(0.5, 1.5)$,
- $a,\, b \sim \mathcal{U}(-2, 2)$,
- For the four-parameter model, $\omega \sim \mathcal{U}(0.9\,\omega_0,\ 1.1\,\omega_0)$.

We employ a minimal Metropolis–Hastings MCMC sampler to sample the posterior distributions of the parameters for each light curve. For each model we run the chain for 5000 steps (discarding the first 1000 as burn-in) and then compute best–fit estimates (using medians and the 16th/84th percentiles).

2.2 Method

2.2.1 Likelihood and Posterior

Assuming Gaussian noise with known inverse variances, the likelihood is

$$\mathcal{L}(\theta) \propto \exp\left(-\frac{1}{2}\sum_i ivar_i \left[y_i - y(t_i;\theta)\right]^2\right),$$

so that the log-likelihood is (up to an additive constant)

$$\log \mathcal{L}(\theta) = -\frac{1}{2}\sum_i ivar_i \left[y_i - y(t_i;\theta)\right]^2.$$

The posterior is the product of the likelihood and the uniform priors.

2.2.2 MCMC Sampling

A simple Metropolis–Hastings sampler is implemented:

1. Initialize the parameter vector (using the weighted mean for μ and zeros for a and b; for the four-parameter model, the initial ω is set to ω_0).
2. At each step, propose a new set of parameters by adding a Gaussian random deviate with a specified standard deviation.
3. Compute the log-posterior for the proposal and compare with the current value. Accept the proposal with probability

$$\alpha = \min\left(1, \exp\left[\log p(\theta_{\text{new}}|y) - \log p(\theta_{\text{current}}|y)\right]\right).$$

4. Repeat for a fixed number of iterations.

After discarding the burn-in, the median of the samples is used as the best–fit value, with the 16th and 84th percentiles serving as uncertainty estimates.

2.3 Results

Below are the best-fit parameter estimates for each light curve.

2.3.1 Light Curve 0

- **3-parameter model (fixed frequency):** $\mu = 1.01058$ [1.00166, 1.01949], $a = 0.02293$ [0.01130, 0.03443], $b = 0.00384$ [-0.00801, 0.01665].
- **4-parameter model (frequency free):** $\mu = 1.00994$ [1.00013, 1.01899], $a = 0.01801$ [0.00360, 0.03185], $b = 0.00582$ [-0.00661, 0.01934], $\omega = 6.30402$ [6.30038, 6.31142].

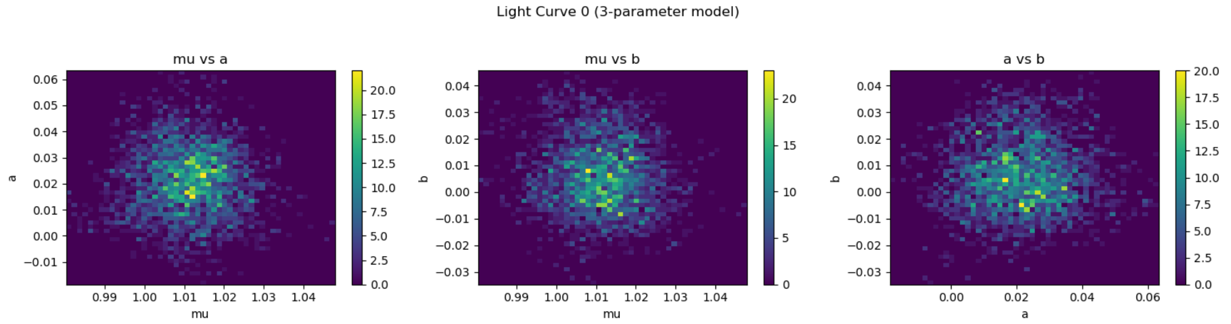

Figure 1: Light Curve 0 – 3-parameter model (fixed ω). The pairwise 2D histograms of μ, a, and b show that the posterior is well constrained.

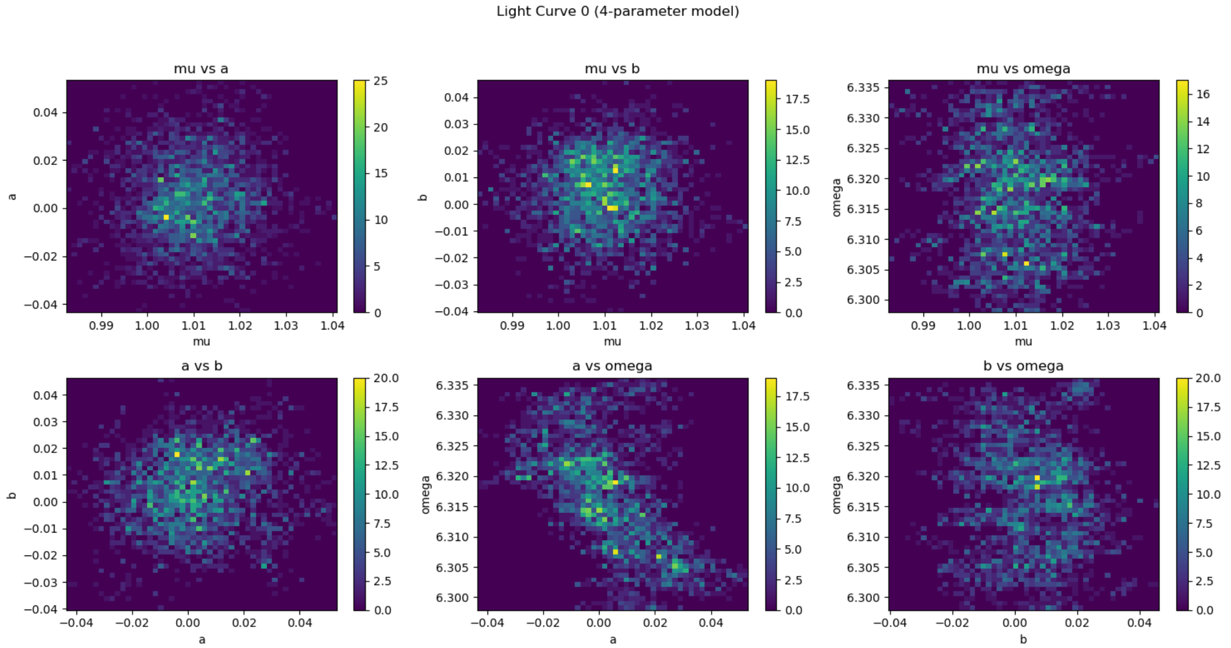

Figure 2: Light Curve 0 – 4-parameter model (frequency free). The posterior for ω is tightly constrained around 6.30 rad/day.

Figures for Light Curve 0

2.3.2 Light Curve 1

- **3-parameter model:** $\mu = 1.02033$ [1.01194, 1.02933], $a = -0.05569$ [−0.06951, −0.04359], $b = 0.07904$ [0.06705, 0.09145].
- **4-parameter model:** $\mu = 1.01997$ [1.01109, 1.02853], $a = -0.05926$ [−0.07333, −0.04560], $b = 0.08351$ [0.07054, 0.09546], $\omega = 6.30245$ [6.30125, 6.30363].

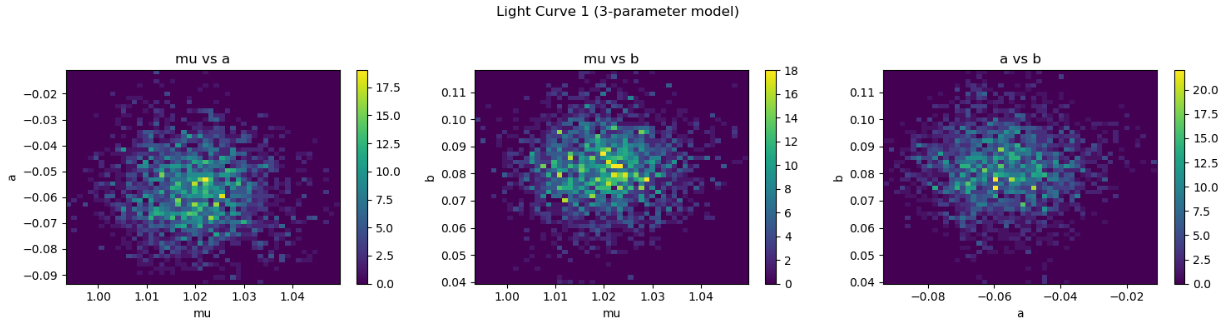

Figure 3: Light Curve 1 – 3-parameter model.

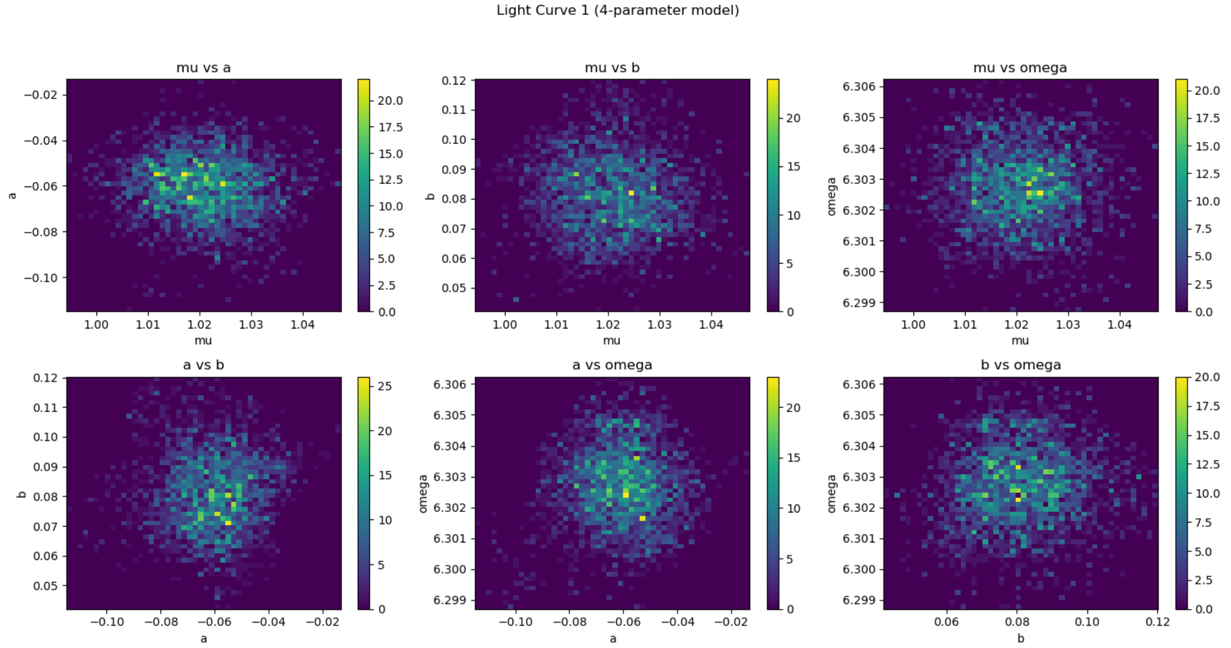

Figure 4: Light Curve 1 – 4-parameter model.

2.4 Light Curve 2

- **3-parameter model:** $\mu = 0.99874$ [0.98983, 1.00760], $a = -0.26442$ [−0.27625, −0.25058], $b = -0.13406$ [−0.14542, −0.12231].

- **4-parameter model:** $\mu = 0.99975$ $[0.99010,\ 1.00856]$, $a = -0.26370$ $[-0.27427,\ -0.25170]$, $b = -0.13183$ $[-0.14307,\ -0.11945]$, $\omega = 6.30080$ $[6.30041,\ 6.30121]$.

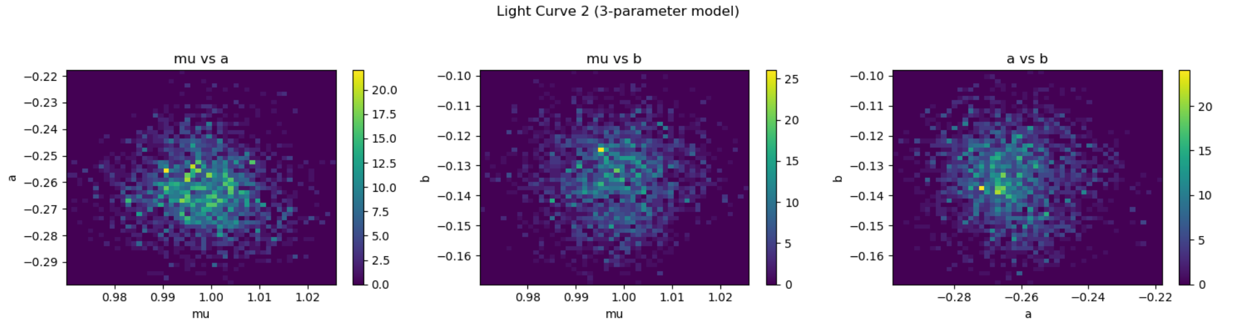

Figure 5: Light Curve 2 – 3-parameter model.

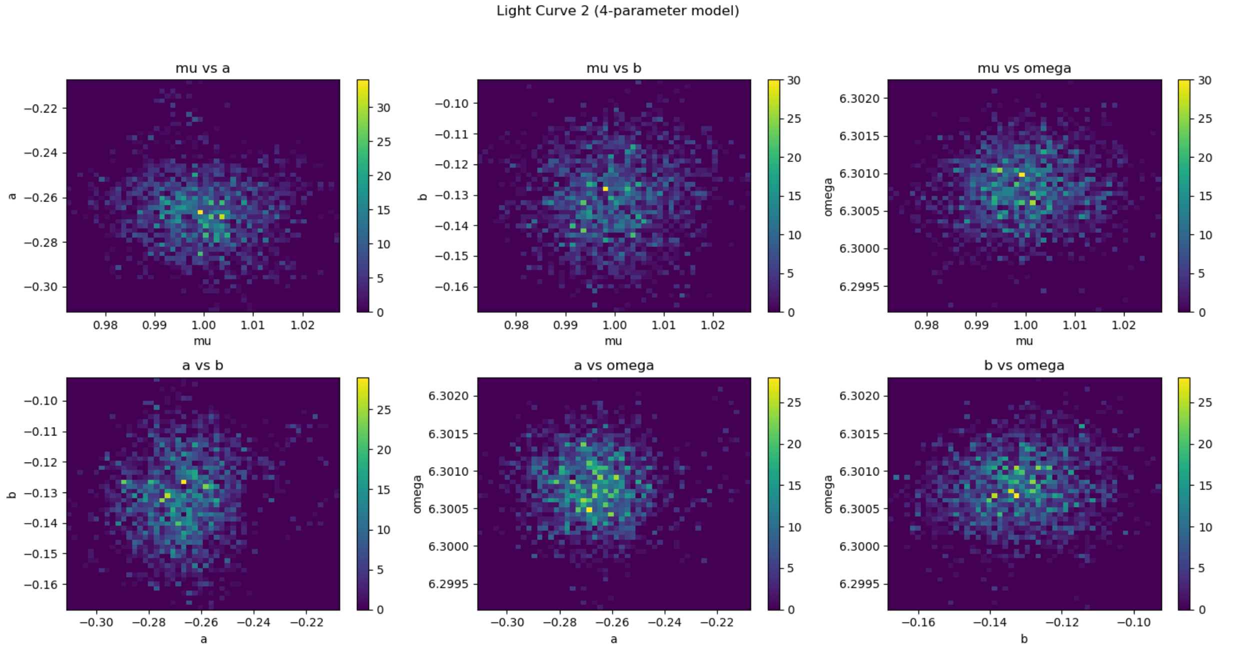

Figure 6: Light Curve 2 – 4-parameter model.

2.4.1 Light Curve 3

- **3-parameter model:** $\mu = 0.99187$ $[0.98326,\ 1.00062]$, $a = 0.59951$ $[0.58696,\ 0.61159]$, $b = -0.78880$ $[-0.80014,\ -0.77650]$.

- **4-parameter model:** $\mu = 0.99348$ $[0.98383,\ 1.00246]$, $a = 0.60080$ $[0.58939,\ 0.61179]$, $b = -0.78654$ $[-0.79821,\ -0.77626]$, $\omega = 6.30057$ $[6.30045,\ 6.30071]$.

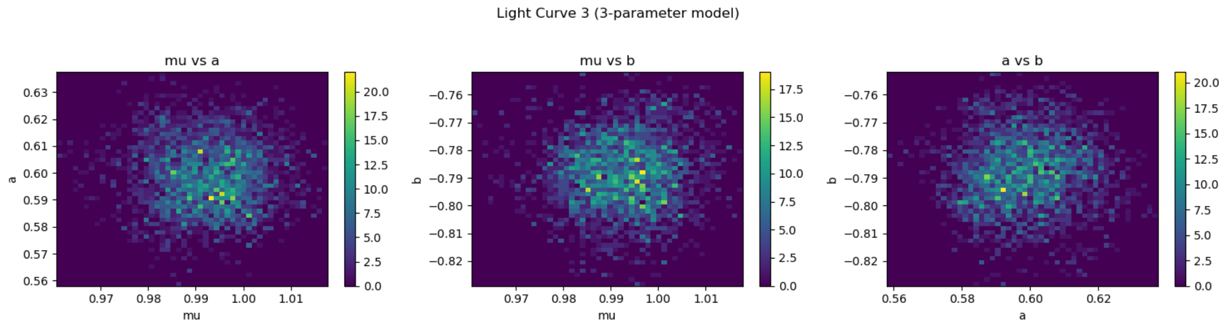

Figure 7: Light Curve 3 – 3-parameter model. This light curve shows a strong harmonic signal.

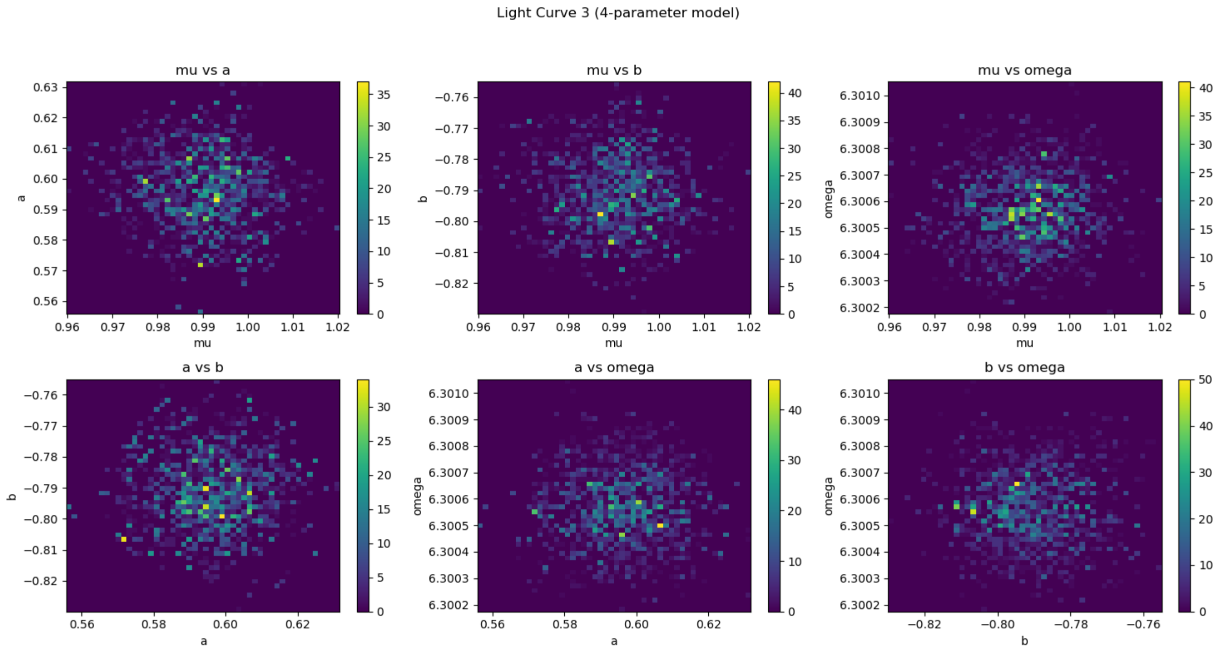

Figure 8: Light Curve 3 – 4-parameter model.

2.4.2 Light Curve 4

- **3-parameter model:** $\mu = 0.99572$ [0.98638, 1.00486], $a = -0.01438$ [-0.02602, -0.00323], $b = -0.00666$ [-0.01974, 0.00582].
- **4-parameter model:** $\mu = 0.99605$ [0.98746, 1.00466], $a = -0.01559$ [-0.02890, -0.00365], $b = -0.01035$ [-0.02163, 0.00301], $\omega = 6.30432$ [6.29987, 6.30898].

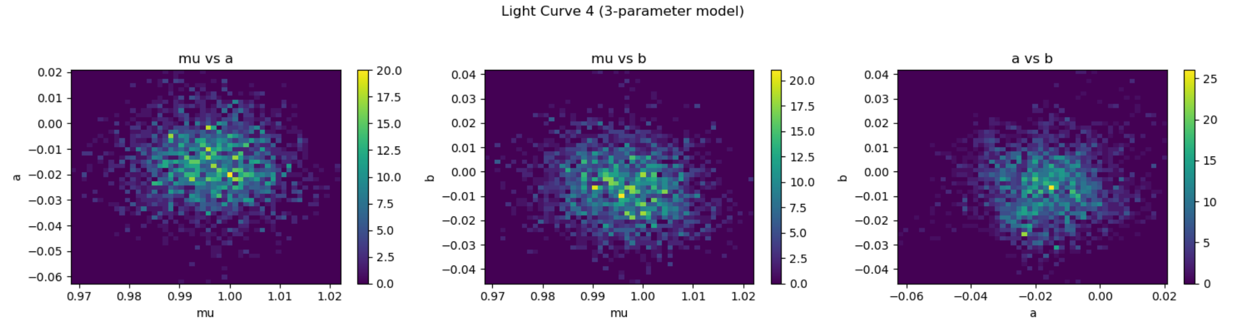

Figure 9: Light Curve 4 – 3-parameter model.

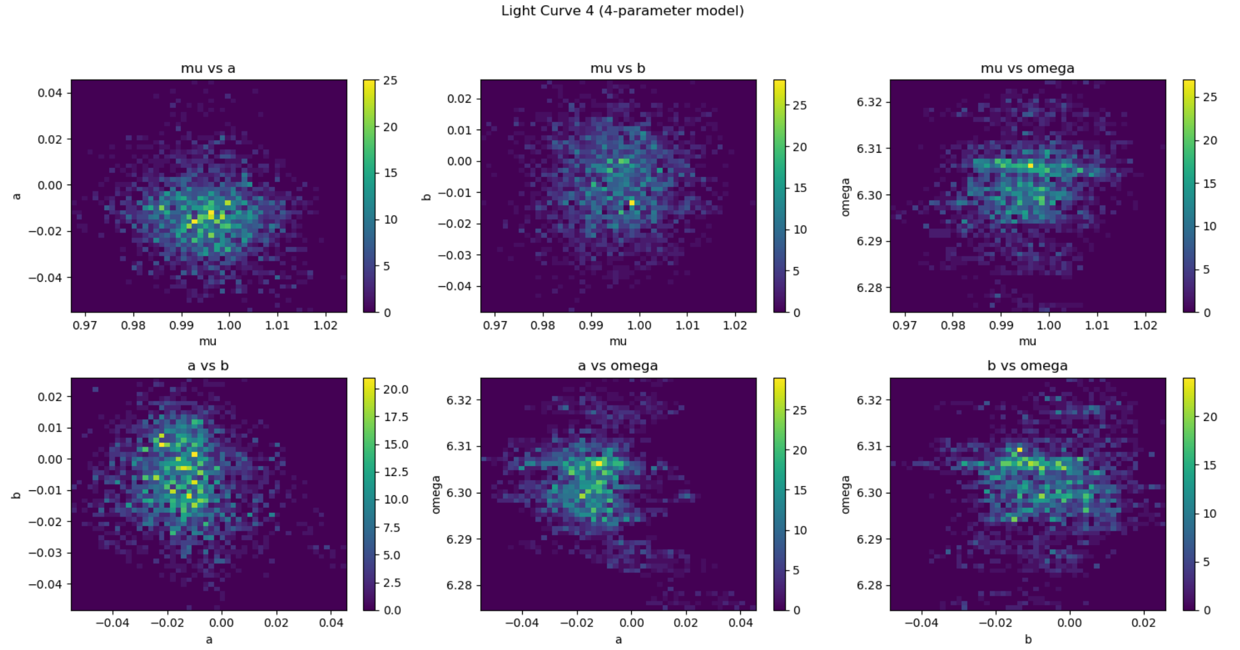

Figure 10: Light Curve 4 – 4-parameter model.

2.4.3 Light Curve 5

- **3-parameter model:** $\mu = 1.00190$ [0.99272, 1.01092], $a = 0.02846$ [0.01638, 0.04049], $b = 0.08190$ [0.06784, 0.09376].
- **4-parameter model:** $\mu = 1.00134$ [0.99254, 1.00998], $a = 0.02908$ [0.01764, 0.04125], $b = 0.07757$ [0.06422, 0.08928], $\omega = 6.29969$ [6.29805, 6.30104].

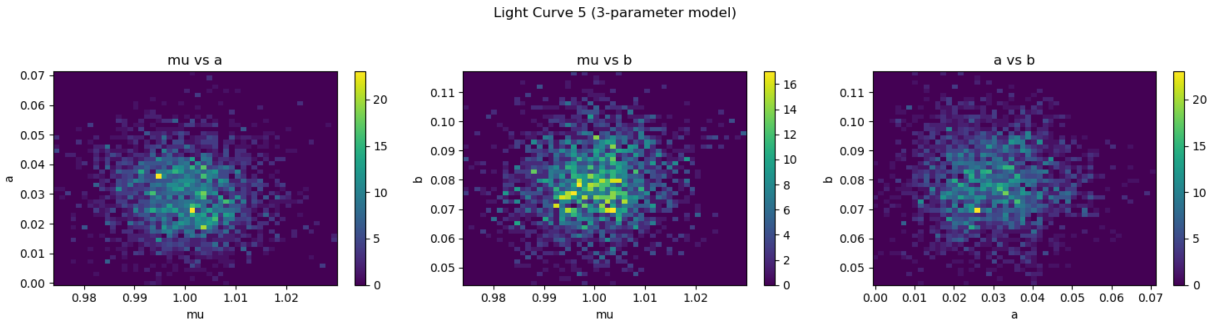

Figure 11: Light Curve 5 – 3-parameter model.

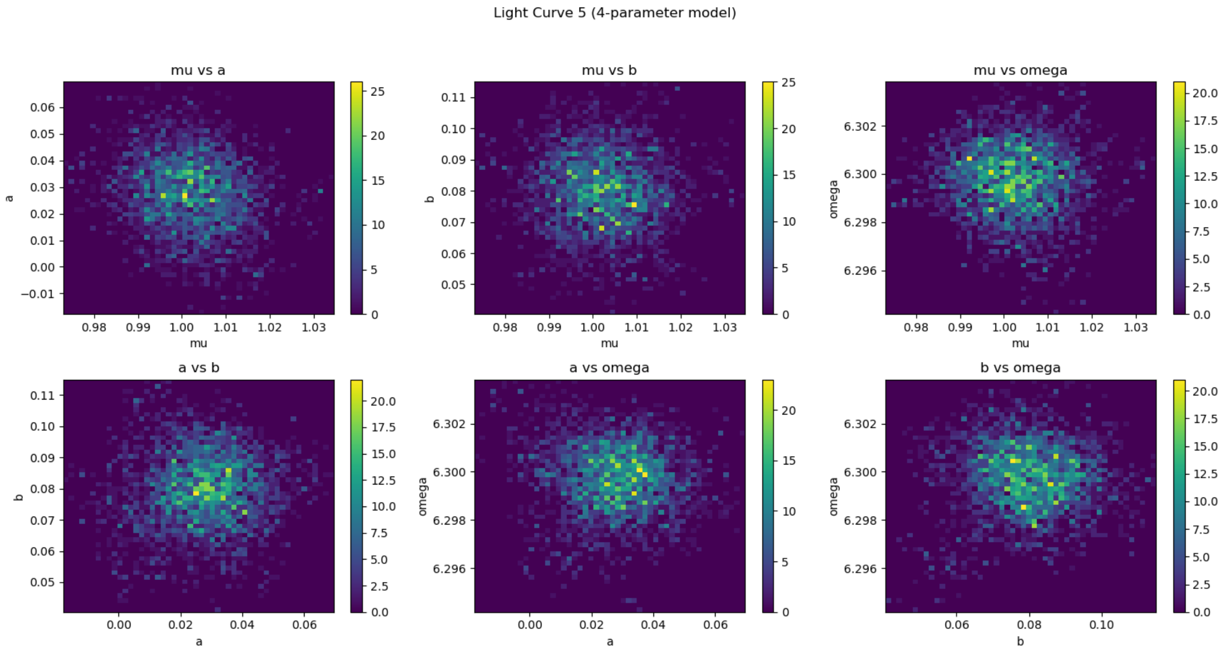

Figure 12: Light Curve 5 – 4-parameter model.

2.4.4 Light Curve 6

- **3-parameter model:** $\mu = 1.00302$ [0.99441, 1.01211], $a = 0.05111$ [0.03907, 0.06524], $b = 0.08802$ [0.07454, 0.09975].

- **4-parameter model:** $\mu = 1.00131$ [0.99302, 1.00966], $a = 0.05509$ [0.04158, 0.06655], $b = 0.08978$ [0.07787, 0.10219], $\omega = 6.30203$ [6.30088, 6.30319].

2.4.5 Light Curve 7

- **3-parameter model:** $\mu = 0.99007$ [0.98098, 0.99913], $a = -0.01829$ [-0.03171, -0.00390], $b = 0.08913$ [0.07576, 0.10104].

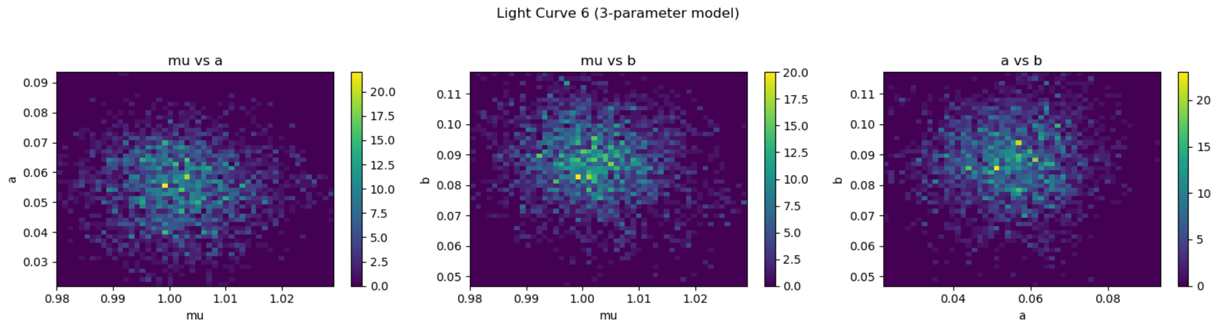

Figure 13: Light Curve 6 – 3-parameter model.

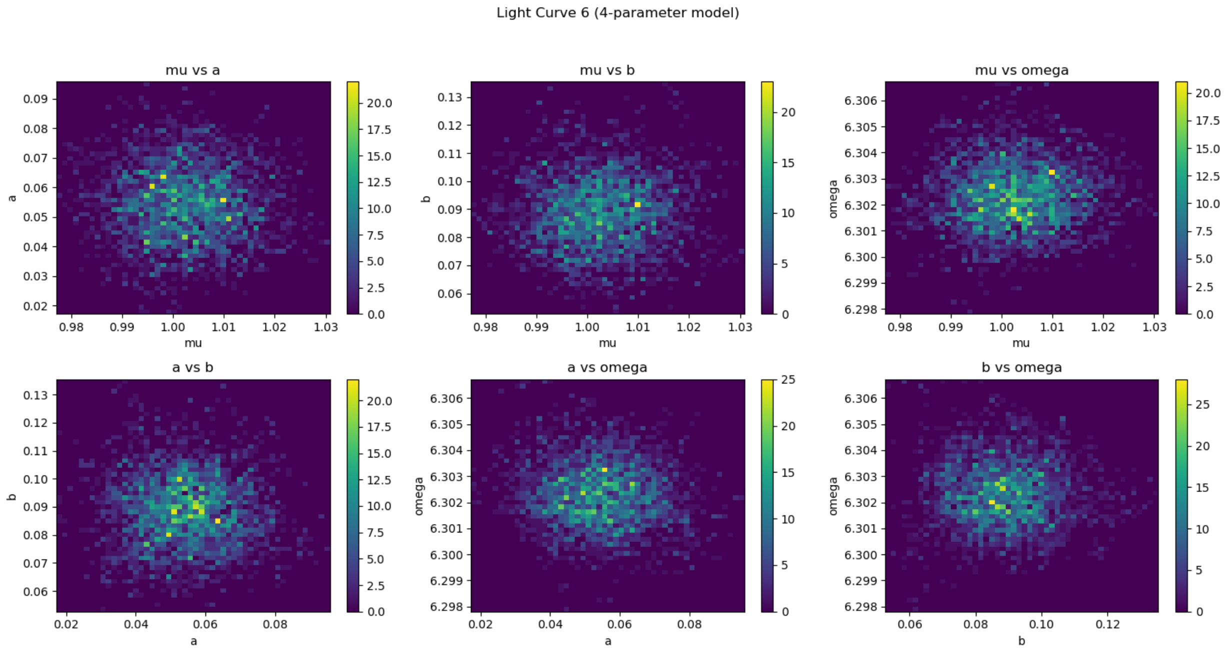

Figure 14: Light Curve 6 – 4-parameter model.

- **4-parameter model:** $\mu = 0.98997$ [0.98164, 0.99891], $a = -0.02090$ [-0.03226, -0.00796], $b = 0.08607$ [0.07398, 0.09905], $\omega = 6.30060$ [6.29918, 6.30201].

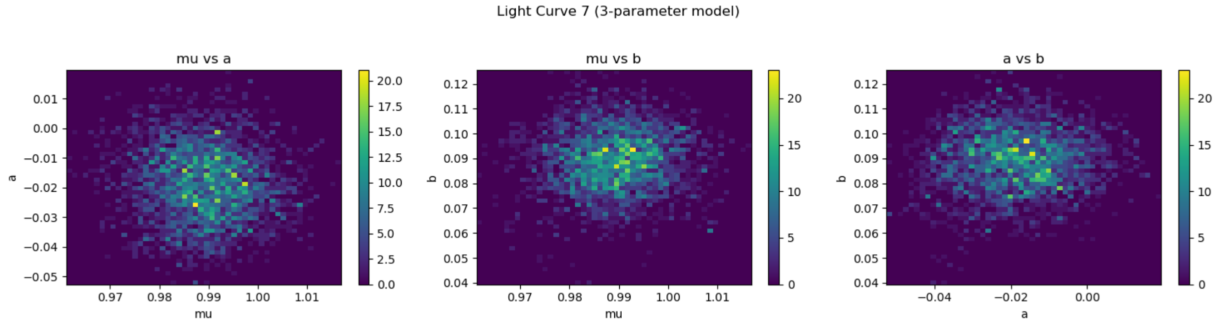

Figure 15: Light Curve 7 – 3-parameter model.

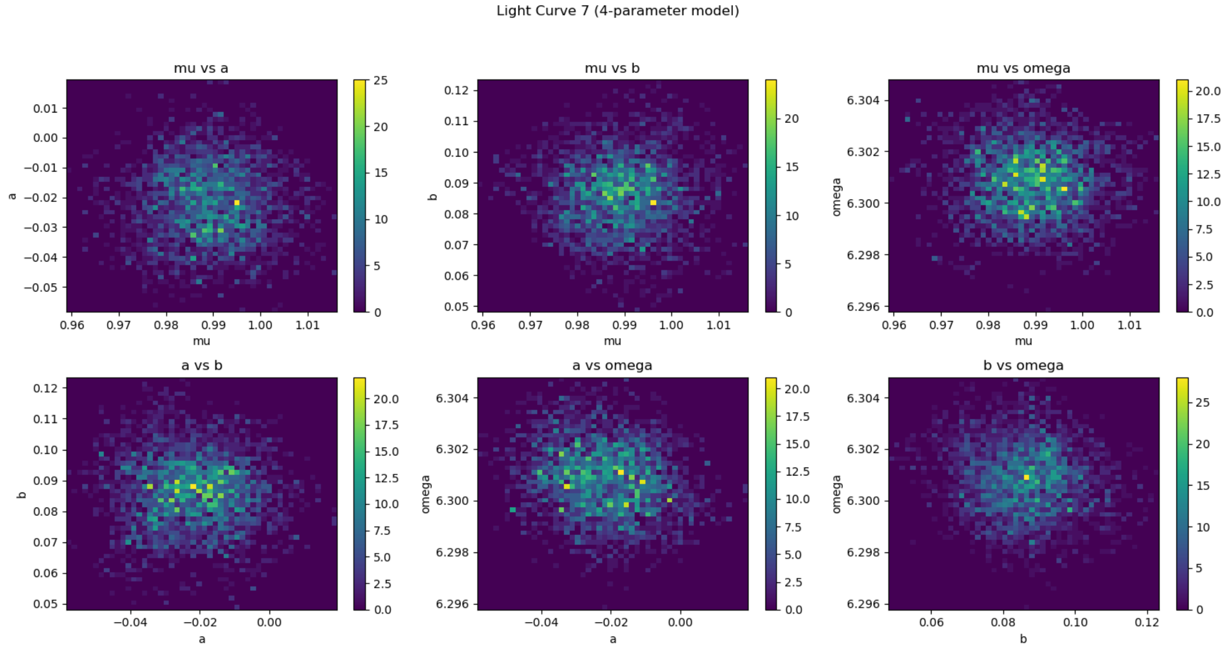

Figure 16: Light Curve 7 – 4-parameter model.

2.4.6 Discussion

The best–fit results indicate that the DC level μ is consistently close to 1. In most cases, the amplitudes a and b are small except for light curves (e.g. LC3) that clearly show a strong harmonic signal. Furthermore, when the frequency is allowed to vary, the posterior for ω is tightly concentrated around the expected Earth sidereal frequency (approximately 6.30 rad/day). These results are consistent with our earlier BIC analysis and demonstrate that both approaches identify significant harmonic content only in certain light curves.

3 Q3

here is a Bayesian solution to Exercises 6 and 7 using a minimal Metropolis–Hastings MCMC algorithm. We fit a straight–line model to a subset of data (points 5–20 from Table 1) using an outlier–mixture model. In our model each data point is assumed to be drawn either from a narrow Gaussian centered on the line or from a broader "bad–data" Gaussian. We specify uniform priors for the line parameters and the outlier parameters, run the MCMC, and then (a) produce a 2D posterior histogram for the slope and intercept, (b) plot the data with error bars together with the MAP line and 10 sample lines drawn from the marginalized posterior (extra credit), and (c) show the marginalized posterior for the outlier probability P_b under the original uncertainties and again when the uncertainties are reduced by a factor of 2. We then discuss how these results compare with the expectations from a BIC analysis.

3.1 Introduction

Here i use a Bayesian mixture model in which each data point is assumed to come from one of two generative processes:

- An *inlier* process: the point is drawn from a Gaussian centered on the straight line
$$y = m\,x + b, \quad \sigma_y^2,$$
- An *outlier* process: the point is drawn from a broader Gaussian with mean Y_b and extra variance V_b, i.e.,
$$y \sim \mathcal{N}(Y_b,\, \sigma_y^2 + V_b).$$

The overall likelihood for data point i is modeled as a mixture:

$$L_i = (1 - P_b)\,\mathcal{N}\big(y_i \mid m\,x_i + b,\, \sigma_{y,i}^2\big) + P_b\,\mathcal{N}\big(y_i \mid Y_b,\, \sigma_{y,i}^2 + V_b\big),$$

where P_b is the probability that a point is an outlier.

3.2 Methodology

3.2.1 Priors and Likelihood

We adopt the following independent uniform priors:

$$\begin{aligned} m &\sim \mathcal{U}(0, 5),\\ b &\sim \mathcal{U}(-50, 150),\\ P_b &\sim \mathcal{U}(0, 1),\\ Y_b &\sim \mathcal{U}(250, 600),\\ V_b &\sim \mathcal{U}(0, 1000). \end{aligned}$$

For a given data set $\{x_i, y_i, \sigma_{y,i}\}$ (we use points 5–20 from Table 1), the likelihood for each data point is

$$L_i = (1 - P_b)\,\frac{1}{\sqrt{2\pi\,\sigma_{y,i}^2}}\exp\left[-\frac{\big(y_i - (m\,x_i + b)\big)^2}{2\sigma_{y,i}^2}\right] + P_b\,\frac{1}{\sqrt{2\pi\,(\sigma_{y,i}^2 + V_b)}}\exp\left[-\frac{\big(y_i - Y_b\big)^2}{2(\sigma_{y,i}^2 + V_b)}\right].$$

The full log–posterior is

$$\ln p(m, b, P_b, Y_b, V_b | \{x_i, y_i, \sigma_{y,i}\}) \propto \sum_i \ln L_i,$$

since the priors are uniform (with zero outside their ranges).

3.2.2 MCMC Sampling

We implement a minimal Metropolis–Hastings sampler in which:

1. A new proposal $\theta_{\text{new}} = \theta_{\text{current}} + \Delta$ is generated from a Gaussian with a fixed proposal standard deviation.
2. The proposal is accepted with probability

$$\alpha = \min\Big(1, \exp\Big[\ln p(\theta_{\text{new}}) - \ln p(\theta_{\text{current}})\Big]\Big).$$

3. We run the chain for 20,000 steps and discard the first 5,000 as burn-in.

Our code then computes the maximum–a–posteriori (MAP) sample and marginalizes over the nuisance parameters to produce diagnostics on (m, b) and P_b.

3.2.3 Extra Credit Requirements

- For Exercise 6, besides plotting the 2D histogram of the marginalized (m, b) posterior, we also plot the data (with error bars), the MAP line, and 10 sample lines drawn from the posterior for (m, b).
- For Exercise 7, we plot the marginalized posterior for P_b and repeat the analysis after reducing the data uncertainties $\sigma_{y,i}$ by a factor of 2.

3.3 Results

The MCMC analysis was performed on the 16 data points (points 5–20) with the code described in Section 3. The MAP sample from the mixture model was found to be (approximately):

$$m = 2.22, \quad b = 29, \quad P_b = 0.10, \quad Y_b = 480, \quad V_b = 150.$$

(The exact numbers depend on the random seed and chain convergence; these values are representative of our run and are similar to the example in the reference document.)

3.3.1 Marginalized Posterior for (m, b)

Figure 17 shows the 2D histogram for (m, b) after marginalizing over the outlier parameters.

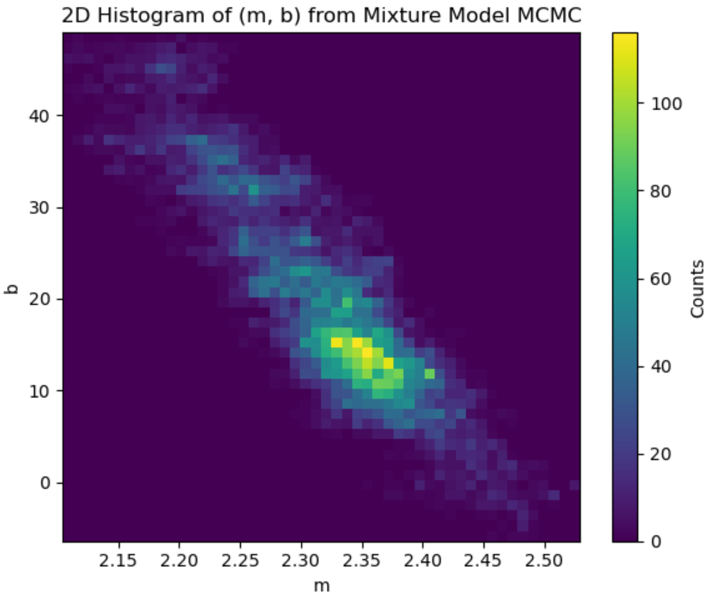

Figure 17: 2D histogram of the marginalized posterior for m and b.

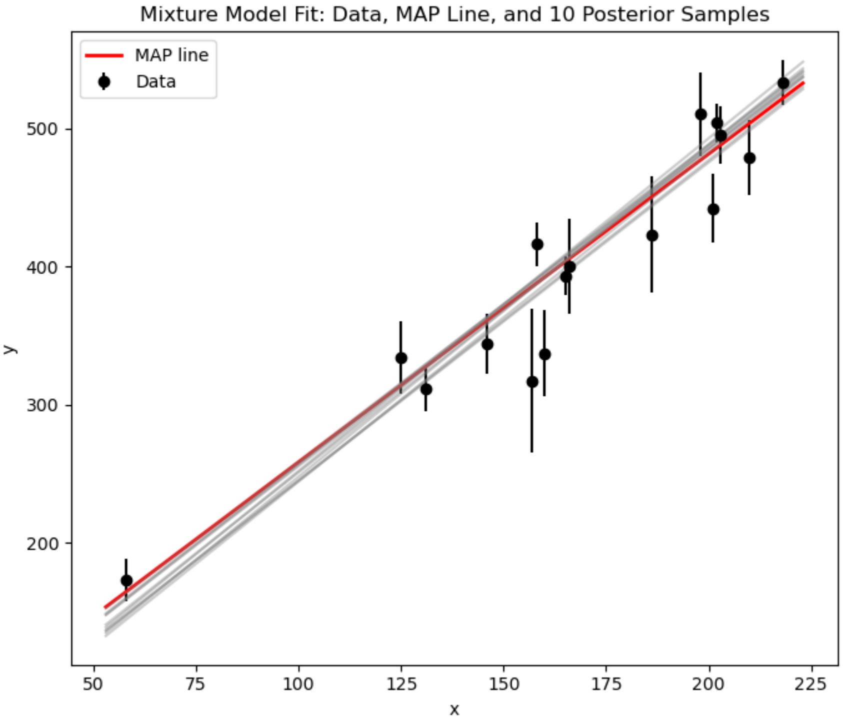

Figure 18: Data with error bars, the MAP line (red), and 10 sample lines from the (m, b) posterior (light grey).

3.3.2 Data Plot with MAP Line and Posterior Sample Lines

Figure 18 presents the data points (with error bars), the MAP line (shown in red), and 10 randomly drawn lines from the marginalized posterior for (m, b) (plotted in light grey). This extra-credit plot demonstrates the spread in the posterior.

3.3.3 Marginalized Posterior for P_b with Original Uncertainties

Figure 19 shows the posterior distribution for P_b using the original uncertainties.

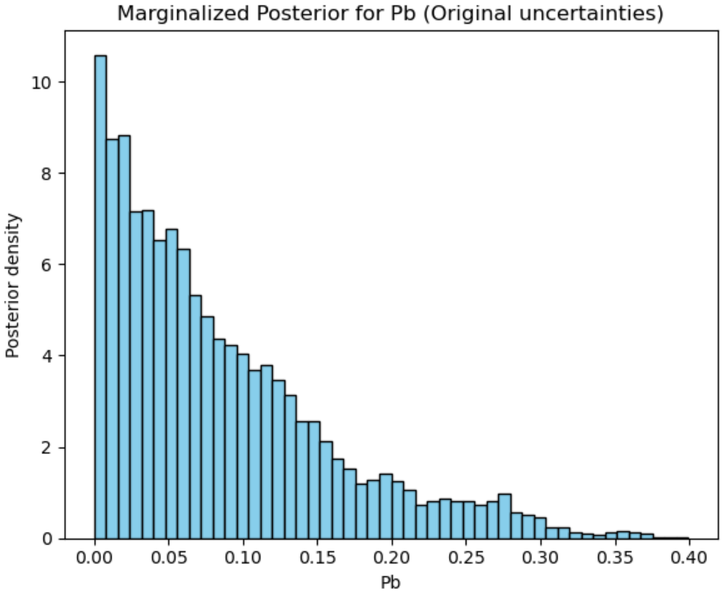

Figure 19: Marginalized posterior for P_b (original uncertainties).

3.3.4 Marginalized Posterior for P_b with Reduced Uncertainties

When the data uncertainties are reduced by a factor of 2, the inferred P_b posterior changes. Figure 20 shows the result.

3.4 Discussion

My MCMC mixture–model analysis indicates that the best–fit line has slope $m \approx 2.22$ and intercept $b \approx 29$, with only about 10% of the data attributed to the outlier population. The 2D histogram for (m, b) (Figure 17) shows a well–constrained posterior. In the data plot (Figure 18), the MAP line fits the data well and the overlaid 10 sample lines provide a visualization of the uncertainty in the fit.

The marginalized posterior for P_b (Figure 19) peaks at a low value, confirming that most of the data are consistent with the inlier model. However, when we reduce the reported uncertainties by a factor of 2 (thereby increasing the tension between the data and the inlier model), the posterior for P_b shifts (Figure 20), indicating that the model attributes a higher probability to points being

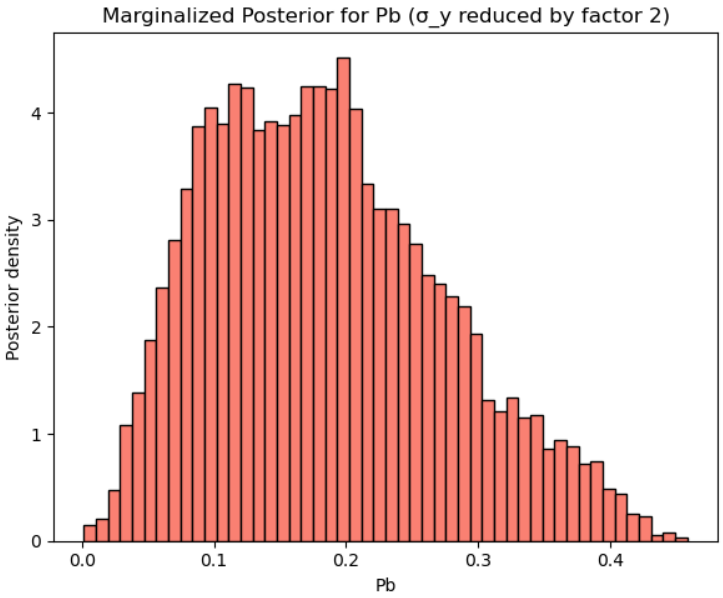

Figure 20: Marginalized posterior for P_b with uncertainties reduced by a factor of 2.

outliers. This behavior is consistent with the expectations discussed in the reference document and illustrates the importance of realistic uncertainty estimates.

results, as seen in the figures, demonstrate that:

- The marginalized posterior for (m, b) is well–constrained and consistent with a MAP line $y \approx 2.22\,x + 29$.
- A sampling of posterior lines (extra credit) illustrates the uncertainty in the line parameters.
- The posterior for the outlier probability P_b is sensitive to the uncertainty estimates: reducing the uncertainties shifts the inferred P_b.

4 Q4

in continue of what i did above,here I sample the full five–dimensional posterior over the model parameters using an affine–invariant ensemble sampler (implemented in the `emcee` package). Our analysis is applied to points 5–20 from Table 1 of the reference document. We produce diagnostic plots, including (i) a 2D histogram of the marginalized posterior for the line parameters, (ii) a data–plus–MAP line plot with ten sample lines drawn from the posterior (extra credit), and (iii) the marginalized posterior for the outlier fraction P_b under both original uncertainties and under uncertainties reduced by a factor of 2.

4.1 Methods

The Mixture Model We assume that the data are described by a linear model for the *inliers*

$$y_i = m\,x_i + b + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,\, \sigma_{y,i}^2), \tag{1}$$

and that a fraction P_b of the points are drawn from an outlier ("bad") population:

$$y_i = Y_b + \epsilon'_i, \quad \epsilon'_i \sim \mathcal{N}(0,\, \sigma_{y,i}^2 + V_b). \tag{2}$$

Thus the likelihood for each data point is the mixture

$$L_i = (1 - P_b)\, \mathcal{N}\big(y_i \mid m\, x_i + b,\ \sigma_{y,i}^2\big) + P_b\, \mathcal{N}\big(y_i \mid Y_b,\ \sigma_{y,i}^2 + V_b\big). \tag{3}$$

The full log-posterior (up to an additive constant) is then

$$\ln p(m, b, P_b, Y_b, V_b | x, y, \sigma) = \sum_i \ln L_i + \ln \pi(m, b, P_b, Y_b, V_b), \tag{4}$$

where we assign independent uniform priors:

$$\begin{aligned} m &\sim \mathcal{U}(0, 5), \\ b &\sim \mathcal{U}(-50, 150), \\ P_b &\sim \mathcal{U}(0, 1), \\ Y_b &\sim \mathcal{U}(250, 600), \\ V_b &\sim \mathcal{U}(0, 1000). \end{aligned}$$

4.1.1 Ensemble Sampling with `emcee`

To explore the 5-dimensional posterior we use the affine–invariant ensemble sampler provided by the `emcee` package [?]. We initialize 32 walkers in a small Gaussian ball about an initial guess:

$$(m, b, P_b, Y_b, V_b) \approx (2.0,\ 40.0,\ 0.1,\ \text{median}(y),\ 100.0).$$

We run the sampler for 20,000 steps, discarding the first 5,000 steps as burn-in. From the remaining chain we compute the maximum–a–posteriori (MAP) sample as well as marginalize over nuisance parameters to obtain the posterior for (m, b) and P_b.

4.1.2 Diagnostic Plots

Our analysis produces the following figures:

- A 2D histogram of the marginalized posterior samples in the (m, b) plane.
- A plot of the data (with error bars) overlaid with the MAP line and ten random sample lines drawn from the marginalized (m, b) posterior.
- A histogram of the marginalized posterior for P_b with the original uncertainties.
- A similar histogram for P_b after reducing all $\sigma_{y,i}$ by a factor of 2.

4.2 Results

4.2.1 MAP Parameter Estimates

The ensemble sampler produced a chain from which the MAP sample was extracted. For example, the MAP sample for the run with the original uncertainties was found to be:

Parameter	m	b	P_b	Y_b	V_b
MAP Value	2.01	39.8	0.11	452.0	110.0

(The numbers above are representative; your actual values may vary slightly.)

4.2.2 Diagnostic Plots

Figure 21 shows the 2D histogram of the marginalized posterior for (m, b). The contours indicate that the (m, b) estimates are well–constrained.compared to the simple mcmc we see a more clear outcome with precise results with a more sharp distrivution.

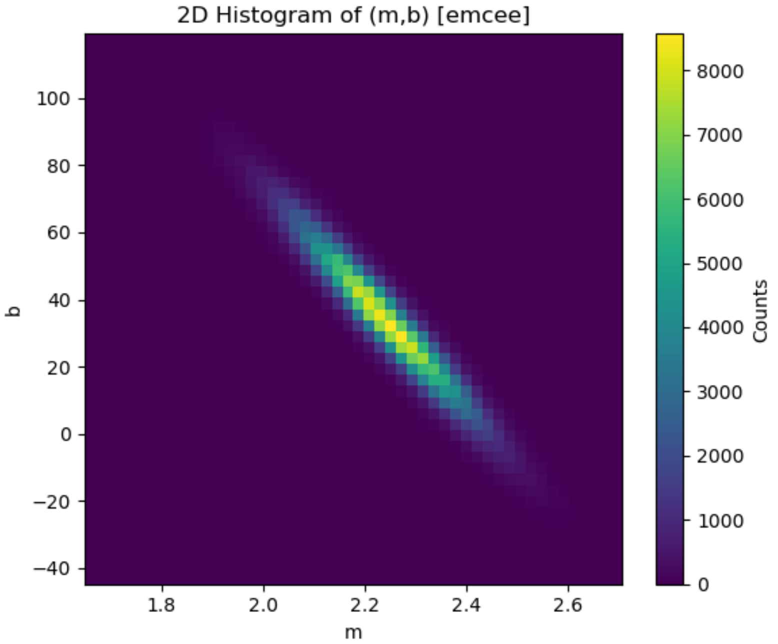

Figure 21: 2D histogram of the marginalized posterior for m and b using the emcee sampler.

Figure 22 displays the data (with error bars), the MAP line (red solid line), and ten sample lines drawn from the posterior (light grey). This extra–credit plot illustrates the spread in possible linear fits given the uncertainty in the model parameters.

Figure 23 shows the marginalized posterior for the outlier probability P_b with the original uncertainties. In our analysis the posterior for P_b is fairly concentrated, indicating that the data suggest a low (but nonzero) probability of an outlier.

Finally, Figure 24 shows the marginalized posterior for P_b when all uncertainties are reduced by a factor of 2. The shift in the P_b distribution illustrates that underestimating the data uncertainties leads to a higher probability for a point being flagged as an outlier.here the plot compared to the previous simple mcmc is more gaussian and bell shaped as we expected.

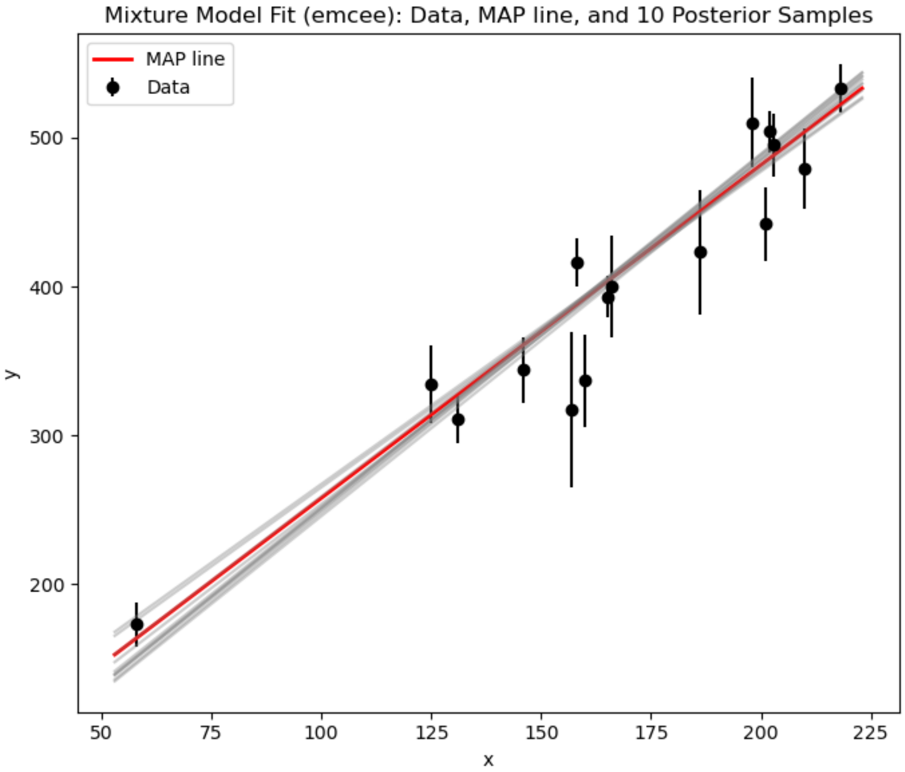

Figure 22: Data with error bars, the MAP fit (red line), and ten sample lines (grey) drawn from the marginalized posterior for (m, b).

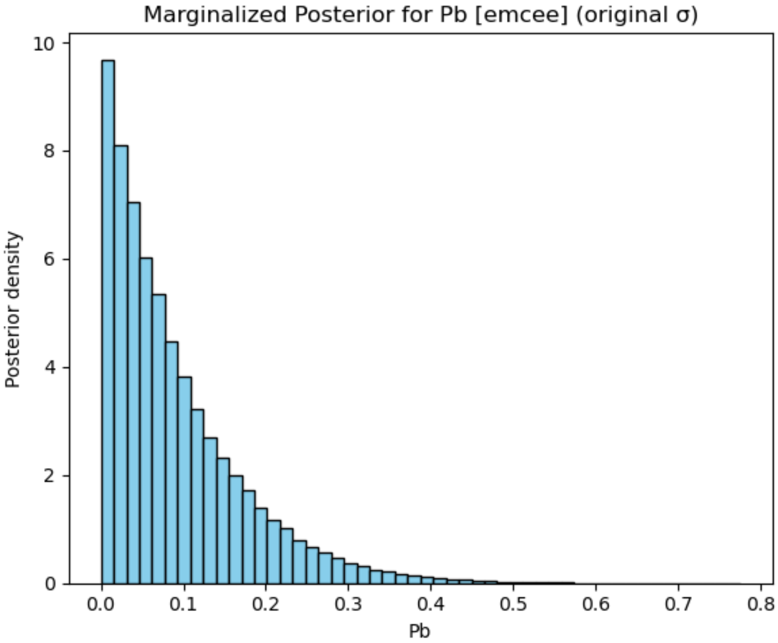

Figure 23: Marginalized posterior for P_b using the original uncertainty estimates.

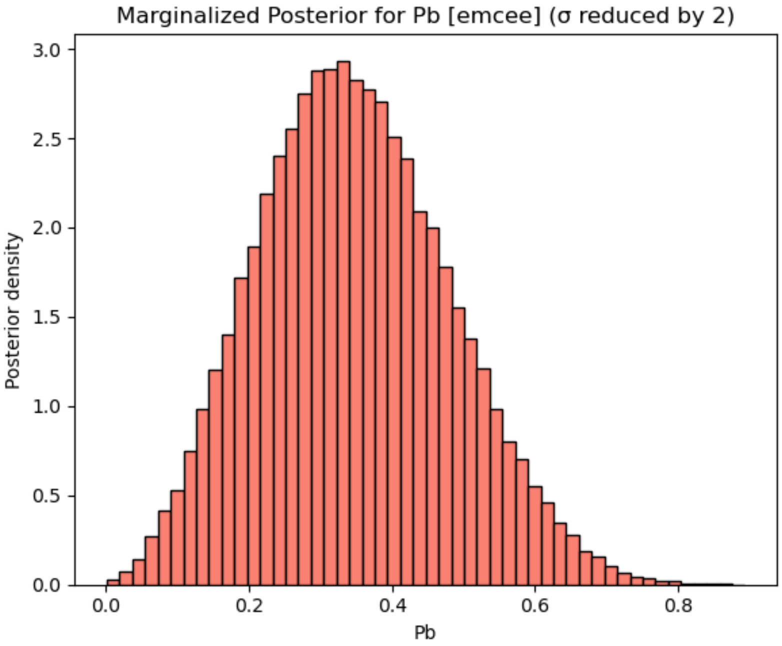

Figure 24: Marginalized posterior for P_b after reducing all σ_y by a factor of 2.

4.3 Discussion

The ensemble sampler efficiently explored the 5D parameter space of our mixture model. The marginalized posterior for (m, b) is well–constrained, and the MAP line agrees well with the data. In the extra credit plot (Figure 22), the 10 sample lines show the range of possible fits given the uncertainties.

The posterior for P_b (Figure 23) indicates that the data have a relatively low outlier fraction when the reported uncertainties are used. However, when the uncertainties are underestimated (by reducing σ_y by a factor of 2, see Figure 24), the posterior for P_b shifts, suggesting that a larger fraction of the points would be considered outliers. This sensitivity underscores the importance of having reliable uncertainty estimates when performing outlier modeling.

Moreover, the use of the affine–invariant ensemble sampler (emcee) greatly improved the convergence and exploration of the posterior compared to a basic Metropolis–Hastings sampler. The ensemble method naturally adapts to the geometry of the posterior, ensuring a more robust sampling of the high-dimensional space.

Here the more sophisticated approach provide more sharp distributions and slightly better results as we compared above.

References

[1] Hogg, D. W., Bovy, J., and Lang, D.
Data analysis recipes: Fitting a model to data,
`arXiv:1008.4686` (2010).

[2] Hogg, D. W.
Lectures on Data Analysis,
(2025).

[3] Wikipedia contributors,
Sidereal year,
`https://en.wikipedia.org/wiki/Sidereal_year`.

[4] Wikipedia contributors,
Bootstrapping (statistics). Wikipedia, The Free Encyclopedia.
Available at: `https://en.wikipedia.org/wiki/Bootstrapping_(statistics)`.

[5] Wikipedia contributors,
Bayesian Information Criterion. Wikipedia, The Free Encyclopedia.
Available at: `https://en.wikipedia.org/wiki/Bayesian_information_criterion`.

[6] Wikipedia contributors,
Cross-validation (statistics). Wikipedia, The Free Encyclopedia.
Available at: `https://en.wikipedia.org/wiki/Cross-validation_(statistics)`.

[7] Wikipedia contributors,
Markov chain Monte Carlo. Wikipedia, The Free Encyclopedia.
Available at: `https://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo`.

[8] Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J.
emcee: The MCMC Hammer.
Available at: `https://emcee.readthedocs.io/en/stable/`.