Part III Astrostatistics: Example Sheet 3 Solution
Example Class: Friday, 13 Mar 2020, 12:00pm, MR5

1 Doubly Lensed Quasar Time Delay Estimation

Quasar light curves (brightness time series) $f(t)$ (in magnitudes) are often modelled using the Orstein-Uhlenbeck (O-U) process, which is described mathematically by a stochastic differential equation of the form:

$$df(t) = \tau^{-1}[c - f(t)]\, dt + \sigma dW_t \tag{1}$$

where the second term is a Brownian motion (continuous-time limit of a random walk), with the variability scaled by σ, and the first term is a drag that tends to return the brightness back to the mean level c. The O-U process is a Gaussian process

$$f(t) \sim \mathcal{GP}(m(t), k(t, t')\,) \tag{2}$$

with mean function $m(t) = c$ and covariance function or kernel:

$$\mathrm{Cov}[f(t), f(t')] = k(t, t') = A^2 \exp(-|t - t'|/\tau) \tag{3}$$

with characteristic amplitude $A^2 = \tau\sigma^2/2$. The characteristic timescale for the quasar brightness to revert to the mean c is τ. Hence, astronomers often called this a "damped random walk". In a doubly-lensed quasar system, two images of the same quasar are observed. However, their brightness time series will have a time delay and magnification relative to each other due to the gravitational lensing effects. Find in the accompanying dataset, measurements of the brightness time series of two images of a lensed quasar, $y_1(t)$ and $y_2(t)$. Assume the measurement errors are Gaussian with the given standard deviations. Where possible, write down and derive the relevant equations before you implement them in code.

1. Plot the data. For each image (y_1 or y_2) time series separately, fit an O-U process by optimising the marginal likelihood to estimate c, A, and τ for each image. Are these estimates consistent between the two time series? Estimate the overall relative magnification factor (difference in magnitudes) between the two images. (The relative multiplicative magnification μ due to the gravitational lens is related to the magnitude shift by $\Delta m = -2.5 \log_{10} \mu$).

   Solution: See Figures 1, 2, and code. Optimising the GP marginal likelihood (derived below) for y_1, we find

   $$c_1 = -0.0063; A_1 = 0.0911; \tau_1 = 41.5204.$$

   Optimising the GP marginal likelihood (derived below) for y_2, we find

   $$c_2 = 0.1053; A_2 = 0.1026; \tau_2 = 40.5140.$$

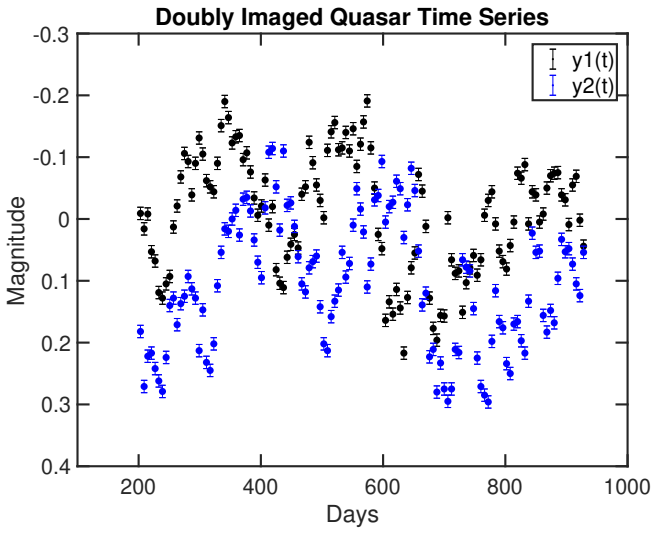

Figure 1: **Plot of both raw time series datasets.**

We find that A and τ are generally consistent between the two time series. The mean levels are different by $\Delta c = c_2 - c_1 \approx 0.10$, suggesting that the magnitude shift is $\Delta m \approx 0.10$.

Let $\boldsymbol{y}_1$ be the vector of time series measurements of quasar image 1 at observation times $\boldsymbol{t}_o$. That is, y_1^i is the measured magnitude of image 1 at time t_o^i. Let $\boldsymbol{f}_o$ be the vector of latent values of the underlying function at the observation times $\boldsymbol{t}_o$. Since $f(t)$ is a realisation of a Gaussian process, $\boldsymbol{f}_o$ has a multivariate Gaussian prior distribution,

$$\boldsymbol{f}_o \sim N(\mathbf{1}c, \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o))$$

where 1 is a vector of ones of the same dimension as $\boldsymbol{f}_o$, and $\boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)$ is a matrix, whose (i,j)th entry is an evaluation of the kernel $k(t_o^i, t_o^j)$ between the pair of observation times t_o^i, t_o^j, and depends on A, τ. The measurement covariance matrix $\boldsymbol{W}$ can be constructed by placing the stated measurement variances σ_i^2 on the diagonal W_{ii}, and assuming the measurement covariances are zero (as they are not reported). Then the measurement process can be described by:

$$\boldsymbol{y}_1 | \, \boldsymbol{f}_o \sim N(\boldsymbol{f}, \boldsymbol{W})$$

Using the properties of multivariate Gaussians, and/or the fact that $\boldsymbol{y}_1$ results from the sum of two Gaussian random vectors, $\boldsymbol{f}_o$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{W})$, we have the sampling distribution

$$\begin{aligned} P(\boldsymbol{y}_1 | \, c, A, \tau^2) &= \int P(\boldsymbol{y}_1, \boldsymbol{f}_o | \, c, A, \tau^2) \, d\boldsymbol{f}_o = \int N(\boldsymbol{y}_1 | \boldsymbol{f}_o, \boldsymbol{W}) N(\boldsymbol{f}_o | \mathbf{1}c, \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o)) \, d\boldsymbol{f}_o \\ &= N(\boldsymbol{y}_1 | \, \mathbf{1}c, \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W}) = N(\boldsymbol{y}_1 | \, \mathbf{1}c, \boldsymbol{\Sigma}(A, \tau)) \end{aligned}$$

where we define $\boldsymbol{\Sigma}(A, \tau) \equiv \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W}$. When viewed as a function of the parameters c, A, τ, with the data fixed, this is the marginal likelihood function

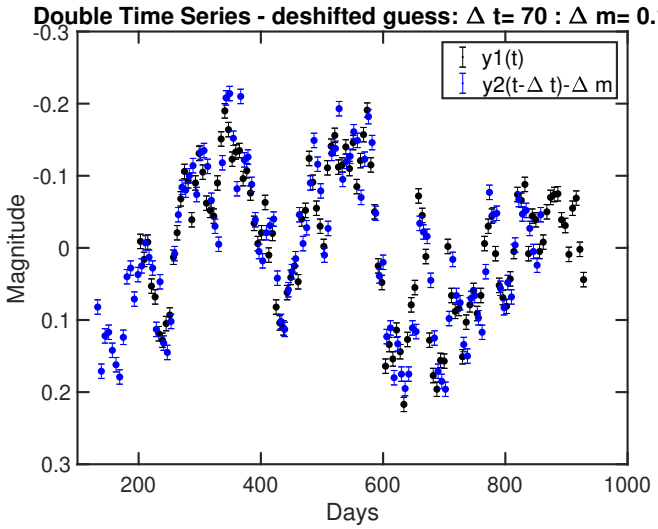

Figure 2: **Plot of time series $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ shifted by guesses for Δt and Δm.**

(because we have effectively marginalised out $\boldsymbol{f}_o$). The marginal likelihood for $\boldsymbol{y}_2$ can be derived the same way.

To optimise the marginal likelihood, it is best to numerically minimise the negative log likelihood function with respect to c, A, τ:

$$-\log P(\boldsymbol{y}_1|\, c, A, \tau^2) = \frac{1}{2}\log\det \boldsymbol{\Sigma}(A,\tau) + \frac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{1}c)^T\, \boldsymbol{\Sigma}(A,\tau)^{-1}\, (\boldsymbol{y}_1 - \boldsymbol{1}c)$$

It is inadvisable in the first term to directly compute $\det \boldsymbol{\Sigma}(A,\tau)$ and then take the log, because the determinant may likely be near zero. Instead, since $\boldsymbol{\Sigma}$ must be a proper covariance matrix, we can perform a Cholesky decomposition, $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$, where $\boldsymbol{L}$ is the lower triangular Cholesky factor. Then we can use the identity,

$$\log\det \boldsymbol{\Sigma} = 2\sum_{i=1}^{N} \log L_{ii}$$

along the diagonal of the Cholesky factor to directly compute the log determinant. Now that we have computed the Cholesky factor, we can also use it to compute the quadratic form in the second term. Letting $\boldsymbol{r} = \boldsymbol{y}_1 - \boldsymbol{1}c$,

$$\boldsymbol{r}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{r} = \boldsymbol{r}^T[\boldsymbol{L}\boldsymbol{L}^T]^{-1}\boldsymbol{r} = \boldsymbol{r}^T[(\boldsymbol{L}^{-1})^T\boldsymbol{L}^{-1}]\boldsymbol{r} = \boldsymbol{z}^T\boldsymbol{z}$$

where $\boldsymbol{z}$ is the solution to the linear system of equations $\boldsymbol{L}\boldsymbol{z} = \boldsymbol{r}$. While this is mathematically equivalent to $\boldsymbol{z} = \boldsymbol{L}^{-1}\boldsymbol{r}$, it can be solved using forward substitution without inverting $\boldsymbol{L}$. Thus, the quadratic form can be computed without directly inverting $\boldsymbol{\Sigma}$, which is usually an expensive and often numerically unstable operation.

2. Fixing the values of c, A, and τ you found for each image separately, overplot random light curves drawn from the GP prior on each separate time series dataset. Use the Gaussian Process machinery to estimate the underlying light curve of each image separately. Plot the expectation and standard deviation of the posterior prediction as a function of time.

   Solution: See Figures 3, 5, 4, 6, and the code.

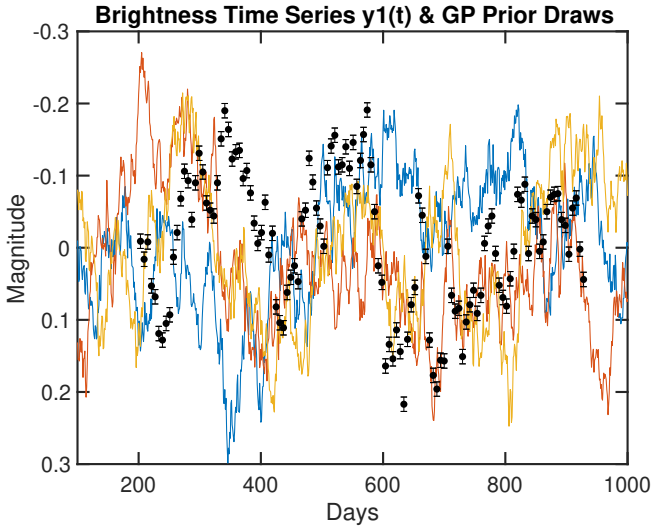

Figure 3: **Plot of y_1 data and draws from the GP prior with hyperparameters optimised for y_1.**

Let t_g be a grid of times on which we want to infer the latent function $f(t)$. Let f_g be the values of the latent function at those times, i.e. $f_g^i = f(t_g^i)$. The joint distribution of y_1 and f_g is

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{f}_g \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{1}c \\ \mathbf{1}c \end{pmatrix}, \begin{pmatrix} \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_o) + \boldsymbol{W} & \boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_g) \\ \boldsymbol{K}(\boldsymbol{t}_g, \boldsymbol{t}_o) & \boldsymbol{K}(\boldsymbol{t}_g, \boldsymbol{t}_g) \end{pmatrix} \right).$$

The covariance matrix components can be derived by considering $\boldsymbol{y}_1 = \boldsymbol{f}_o + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{W})$, and computing $\mathbf{Cov}[\boldsymbol{y}_1, \boldsymbol{y}_1]$, $\mathbf{Cov}[\boldsymbol{y}_1, \boldsymbol{f}_g]$, $\mathbf{Cov}[\boldsymbol{f}_o, \boldsymbol{f}_g]$, $\mathbf{Cov}[\boldsymbol{f}_g, \boldsymbol{f}_g]$, etc. Then using conditional property of multivariate Gaussians random vectors, we can compute the posterior of $\boldsymbol{f}_g$ given $\boldsymbol{y}_1$.

$$\boldsymbol{f}_g | \, \boldsymbol{y}_1 \sim N\left(\mathbb{E}[\boldsymbol{f}_g | \, \boldsymbol{y}_1], \mathbf{Var}[\boldsymbol{f}_g | \, \boldsymbol{y}_1]\right)$$

$$\mathbb{E}[\boldsymbol{f}_g | \, \boldsymbol{y}_1] = \mathbf{1}c + \boldsymbol{K}(\boldsymbol{t}_g, \boldsymbol{t}_o)\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}_g - \mathbf{1}c)$$

$$\mathbf{Var}[\boldsymbol{f}_g | \, \boldsymbol{y}_1] = \boldsymbol{K}(\boldsymbol{t}_g, \boldsymbol{t}_g) - \boldsymbol{K}(\boldsymbol{t}_g, \boldsymbol{t}_o)\boldsymbol{\Sigma}^{-1}\boldsymbol{K}(\boldsymbol{t}_o, \boldsymbol{t}_g)$$

and similarly for $\boldsymbol{f}_g | \, \boldsymbol{y}_2$.

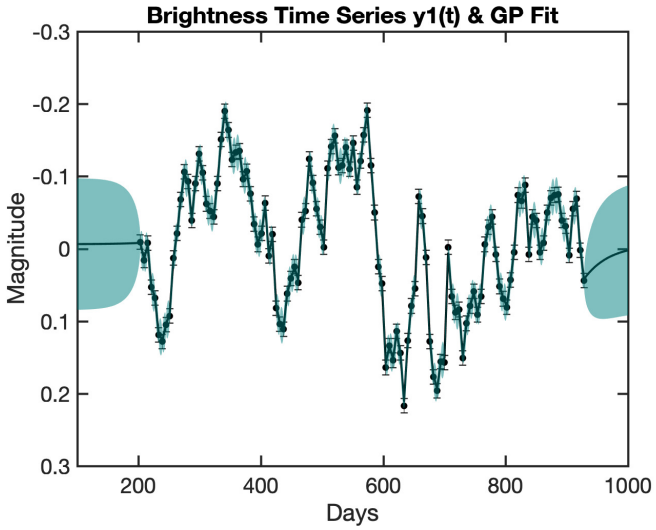

Figure 4: **Plot of y_1 data and the GP posterior fit to y_1.**

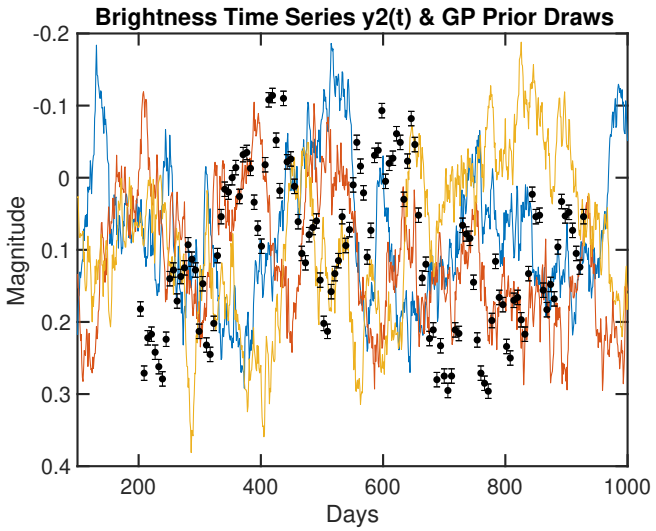

Figure 5: **Plot of y_2 data and draws from the GP prior with hyperparameters optimised for y_2.**

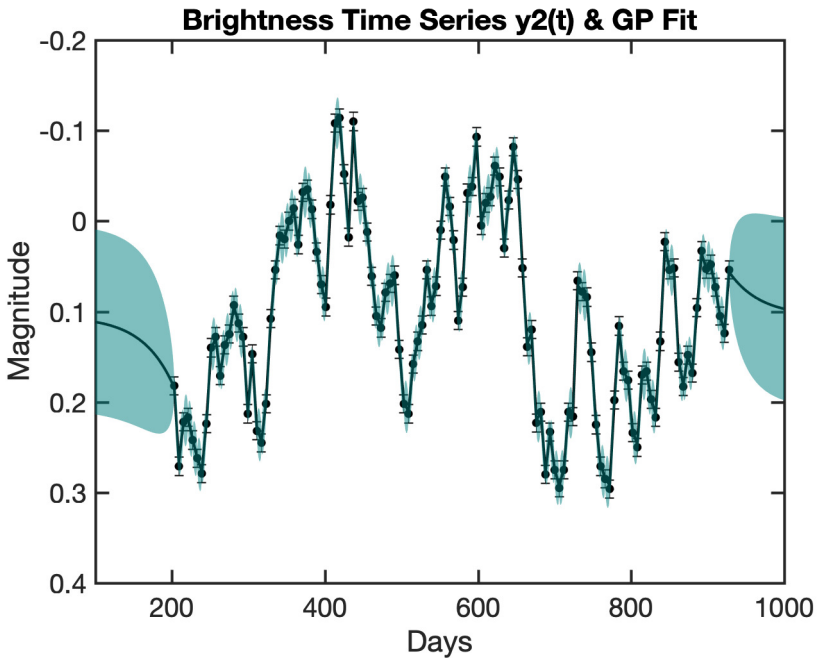

Figure 6: **Plot of y_2 data and the GP posterior fit to y_2.**

3. Derive a likelihood function for the two time series considered jointly, as two copies of the same realisation of the GP but shifted in time by the time delay Δt, and the magnification factor Δm (both relative to y_1), and measured with noise at the observed times. Thus $y_1(t)$ is a noisy measurement of $f(t)$ and $y_2(t)$ is a noisy measurement of $f(t-\Delta t)+\Delta m$. Using suitable non-informative priors, write down a posterior density $P(\Delta t, \Delta m, c, A, \tau | \, \boldsymbol{y}_1, \boldsymbol{y}_2)$.

Solution: Suppose we knew the true time delay Δt and magnitude shift Δm. Then we could construct the combined vectors:

$$\tilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 - \Delta m \end{pmatrix}$$

$$\tilde{\boldsymbol{t}} = \begin{pmatrix} \boldsymbol{t}_o \\ \boldsymbol{t}_o - \Delta t \end{pmatrix}$$

such that $\tilde{\boldsymbol{y}}$ contains noisy measurements of the same latent function $\tilde{\boldsymbol{f}}$ at times $\tilde{\boldsymbol{t}}$ (i.e. $\tilde{f}^i = f(\tilde{t}^i)$).

$$\tilde{\boldsymbol{f}} \sim N(\boldsymbol{1}c, \boldsymbol{K}(\tilde{\boldsymbol{t}}, \tilde{\boldsymbol{t}}))$$
$$\tilde{\boldsymbol{y}}_1 | \, \tilde{\boldsymbol{f}} \sim N(\boldsymbol{f}, \tilde{\boldsymbol{W}})$$

where

$$\tilde{\boldsymbol{W}} = \begin{pmatrix} \boldsymbol{W} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W} \end{pmatrix}.$$

Therefore the marginal likelihood function is

$$P(\tilde{\boldsymbol{y}} | \, , \Delta t, \Delta m, c, A, \tau) = N\left(\tilde{\boldsymbol{y}} | \, \boldsymbol{1}c, \boldsymbol{K}(\tilde{\boldsymbol{t}}, \tilde{\boldsymbol{t}}) + \tilde{\boldsymbol{W}}\right)$$

which again should be computed in log-space according the advice above.

4. Estimate Δt and Δm. Beware that the log likelihood is highly multimodal, so it is important to find the major mode. You may fix the O-U process parameters to reasonable values found previously, or estimate them jointly with $(\Delta t, \Delta m)$.

Solution: First, I optimise the previous marginal likelihood for both time series, using as initial guesses $c = c_1$, $\Delta m = c_2 - c_1$, $A = \frac{1}{2}(A_1 + A_2)$, and $\tau = \frac{1}{2}(\tau_1 + \tau_2)$. After optimising the marginal likelihood and using the observed Fisher information to derive parameter uncertainties, I find

$$\widehat{\Delta t} = 74.96 \pm 0.64$$

$$\widehat{\Delta m} = 0.10 \pm 0.29$$

$$\hat{c} = 0.01 \pm 0.04$$

$$\hat{A} = 0.10 \pm 0.01$$

$$\hat{\tau} = 43.87 \pm 0.19.$$

A contour plot of the marginal likelihood near the optimum is shown in Fig. 7. To explore the multi-modality of the marginal likelihood more closely, in Fig. 8, I have plotted the log of the profile likelihood:

$$L_{\mathbf{prof}}(\Delta t) = \underset{\Delta m, c, A, \tau}{\arg\max} \, P(\tilde{\boldsymbol{y}} | \, , \Delta t, \Delta m, c, A, \tau).$$

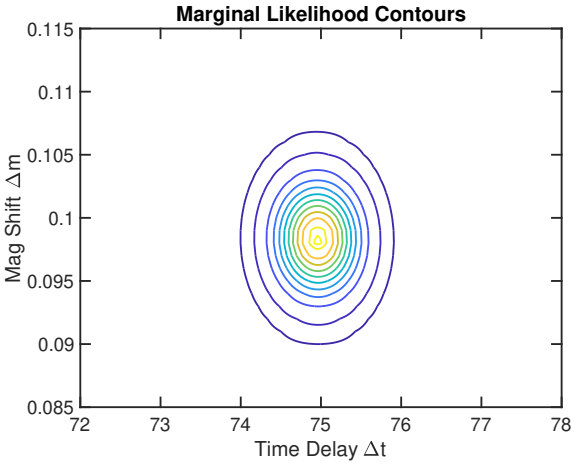

Figure 7: **Plot of the marginal likelihood of $(\Delta t, \Delta m)$ with the other hyperparameters fixed to their MLE values.**

In the code, I implemented a Metropolis-within-Gibbs algorithm to sample the posterior of the shifts with the hyperparameters fixed to their MLE values. I use flat improper priors on the time delay and magnitude shift.

$$P(\Delta t, \Delta m | \, \tilde{\boldsymbol{y}}, \hat{c}, \hat{A}, \hat{\tau}) \propto P(\tilde{\boldsymbol{y}} | \, \Delta t, \Delta m, \hat{c}, \hat{A}, \hat{\tau})$$

I ran 4 chains in parallel for 10^4 iterations, each initialised with the MLEs plus some random jitter. The two proposal jump scales in each parameter was tuned to give $30 - 50\%$ acceptance rates. The Gelman-Rubin ratios for both parameters was < 1.0004. In Fig. 9, I show the sample autocorrelation function of the slowest mixing parameter (Δm). The estimated autocorrelation time was 18 iterations, which I take as a thinning factor. I remove 20% of each chain as burn-in and thin the remainder by a factor of 18 (keeping only every 18th sample). Then I concatenated the chains to compute posterior inferences on the parameters. Figure 10, shows a scatter plot of the joint distribution of $(\Delta t, \Delta m)$ samples, showing that the parameters are well-constrained and uncorrelated in the posterior. Figure 11 show the 1D marginal posteriors distributions as 1D histograms of each parameter estimates. The posterior means and standard deviations are shown.

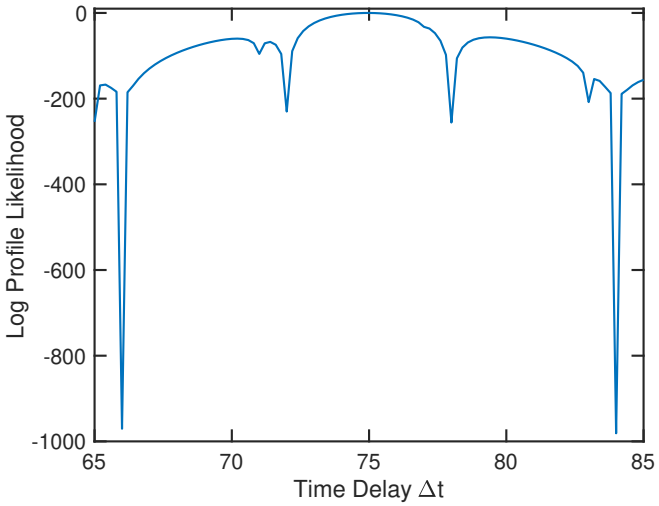

Figure 8: **Plot of the profile likelihood of Δt.**

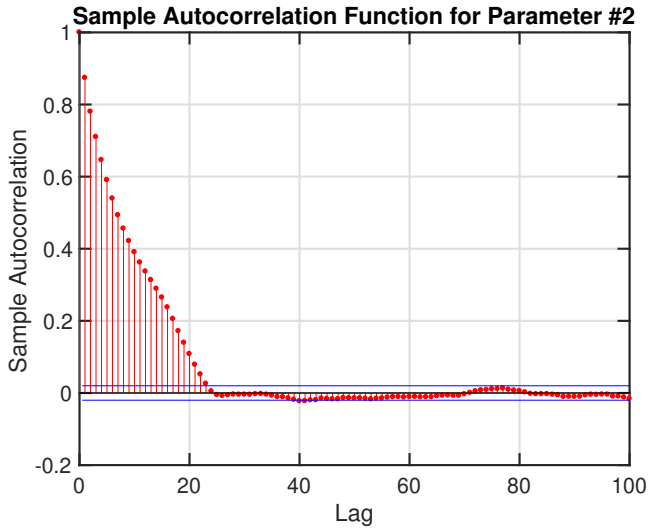

Figure 9: **Plot of the MCMC autocorrelation function of Δm.**

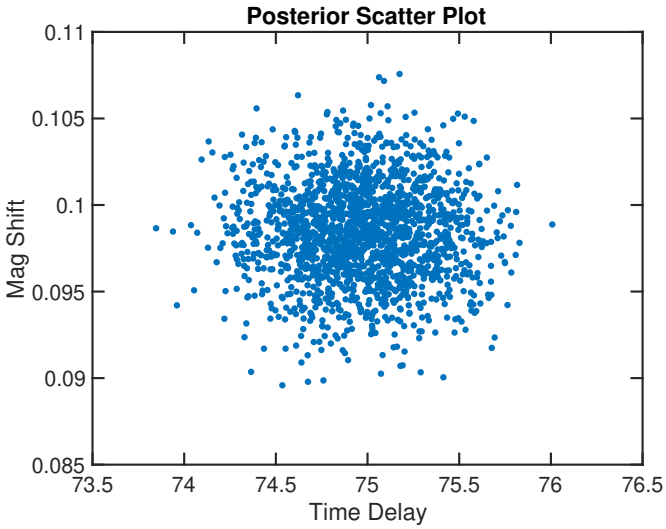

Figure 10: **Plot of the posterior distribution of** $(\Delta t, \Delta m)$ **MCMC samples.**

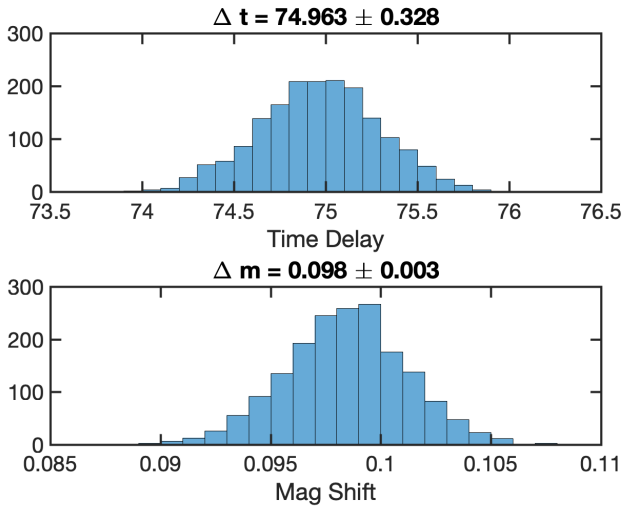

Figure 11: **Plot of the marginal posterior historigrams of** $(\Delta t, \Delta m)$ **MCMC samples.**

5. Overplot the two time series datasets, with y_2 shifted to the y_1 frame by subtracting the estimated Δt and Δm. Now using the O-U parameters you found, plot the posterior estimate of the underlying light curve using the two combined datasets.

   Solution: Figure 12 shows the posterior GP fit two both time series using the estimated time delay and magnitude shift.

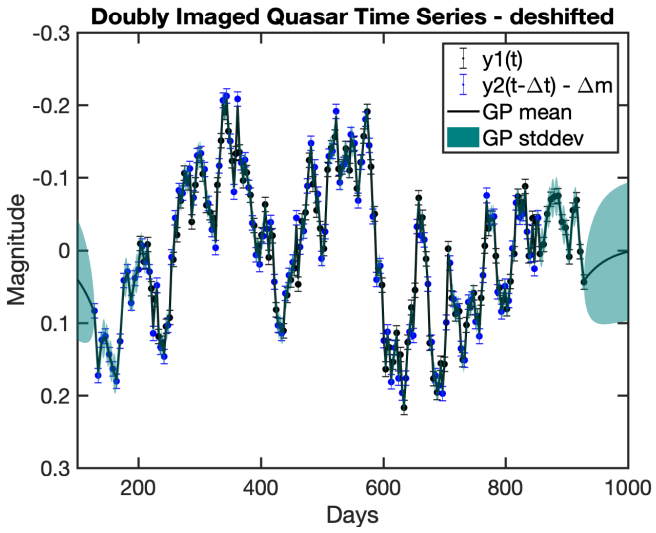

Figure 12: **Plot of the marginal posterior historigrams of $(\Delta t, \Delta m)$ MCMC samples.**

With $c, A, \tau, \Delta t, \Delta m$ set to their fitted estimates, we can compute the posterior of the latent function $\boldsymbol{f}_g$ evaluated on grid points $\boldsymbol{t}_g$:

$$\begin{pmatrix} \tilde{\boldsymbol{y}} \\ \boldsymbol{f}_g \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{1}c \\ \mathbf{1}c \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\tilde{\boldsymbol{t}}, \tilde{\boldsymbol{t}}) + \boldsymbol{W} & \boldsymbol{K}(\tilde{\boldsymbol{t}}, \boldsymbol{t}_g) \\ \boldsymbol{K}(\boldsymbol{t}_g, \tilde{\boldsymbol{t}}) & \boldsymbol{K}(\boldsymbol{t}_g, \boldsymbol{t}_g) \end{bmatrix} \right)$$

$$\boldsymbol{f}_g | \, \tilde{\boldsymbol{y}} \sim N\left(\mathbb{E}[\boldsymbol{f}_g | \, \tilde{\boldsymbol{y}}], \mathbf{Var}[\boldsymbol{f}_g | \, \tilde{\boldsymbol{y}}] \right)$$

$$\mathbb{E}[\boldsymbol{f}_g | \, \tilde{\boldsymbol{y}}] = \mathbf{1}c + \boldsymbol{K}(\boldsymbol{t}_g, \tilde{\boldsymbol{t}})[\boldsymbol{K}(\tilde{\boldsymbol{t}}, \tilde{\boldsymbol{t}}) + \boldsymbol{W}]^{-1}(\boldsymbol{f}_g - \mathbf{1}c)$$

$$\mathbf{Var}[\boldsymbol{f}_g | \, \tilde{\boldsymbol{y}}] = \boldsymbol{K}(\boldsymbol{t}_g, \boldsymbol{t}_g) - \boldsymbol{K}(\boldsymbol{t}_g, \tilde{\boldsymbol{t}})[\boldsymbol{K}(\tilde{\boldsymbol{t}}, \tilde{\boldsymbol{t}}) + \boldsymbol{W}]^{-1}\boldsymbol{K}(\tilde{\boldsymbol{t}}, \boldsymbol{t}_g).$$

Numerical Clue: A proper covariance matrix admits a Cholesky decomposition: $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is the lower triangular Cholesky factor. The log of the determinant $|\boldsymbol{\Sigma}| = \det \boldsymbol{\Sigma}$ can stably be computed from Equation A.18 of Rasmussen & Williams. If you have computed the Cholesky factor $\boldsymbol{L}$, solutions $\boldsymbol{x}$ to linear equations of the form $\boldsymbol{\Sigma}\boldsymbol{x} = \boldsymbol{b}$, i.e. $\boldsymbol{x} = \boldsymbol{\Sigma}^{-1}\boldsymbol{b}$, can be stably computed using forward/backward substitution, rather than by directly inverting $\boldsymbol{\Sigma}$, as described in Rasmussen & Williams, §A.4.