The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1309–1348
DOI: 10.1214/17-AOAS1027

BAYESIAN ESTIMATES OF ASTRONOMICAL TIME DELAYS BETWEEN GRAVITATIONALLY LENSED STOCHASTIC LIGHT CURVES[1]

BY HYUNGSUK TAK[*,2,3], KAISEY MANDEL[†,4], DAVID A. VAN DYK[‡,6], VINAY L. KASHYAP[‡,5], XIAO-LI MENG[†,2] AND ANETA SIEMIGINOWSKA[†,5]

Statistical and Applied Mathematical Sciences Institute[*], Harvard University[†] and Imperial College London[‡]

The gravitational field of a galaxy can act as a lens and deflect the light emitted by a more distant object such as a quasar. Strong gravitational lensing causes multiple images of the same quasar to appear in the sky. Since the light in each gravitationally lensed image traverses a different path length from the quasar to the Earth, fluctuations in the source brightness are observed in the several images at different times. The time delay between these fluctuations can be used to constrain cosmological parameters and can be inferred from the time series of brightness data or light curves of each image. To estimate the time delay, we construct a model based on a state-space representation for irregularly observed time series generated by a latent continuous-time Ornstein–Uhlenbeck process. We account for microlensing, an additional source of independent long-term extrinsic variability, via a polynomial regression. Our Bayesian strategy adopts a Metropolis–Hastings within Gibbs sampler. We improve the sampler by using an ancillarity-sufficiency interweaving strategy and adaptive Markov chain Monte Carlo. We introduce a profile likelihood of the time delay as an approximation of its marginal posterior distribution. The Bayesian and profile likelihood approaches complement each other, producing almost identical results; the Bayesian method is more principled but the profile likelihood is simpler to implement. We demonstrate our estimation strategy using simulated data of doubly- and quadruply-lensed quasars, and observed data from quasars *Q0957+561* and *J1029+2623*.

Received February 2016; revised January 2017.

[1] This work was conducted under the auspices of the CHASC International Astrostatistics Center. CHASC is supported by NSF Grants DMS-12-08791 and DMS-12-09232.

[2]Supported by the Harvard Statistics Department.

[3]Supported in part from the NSF under Grant DMS-11-27914 to the Statistical and Applied Mathematical Sciences Institute.

[4]Supported at Harvard by NSF Grants AST-1211196 and AST-156854.

[5]Supported by a NASA contract to the Chandra X-Ray Center NAS8-03060.

[6]Supported by a Wolfson Research Merit Award (WM110023) provided by the British Royal Society and a Marie-Curie Career Integration Grant (FP7-PEOPLE-2012-CIG-321865) provided by the European Commission.

Key words and phrases. Gravitational lensing, microlensing, Ornstein–Uhlenbeck process, Gibbs sampler, profile likelihood, ancillarity-sufficiency interweaving strategy, adaptive MCMC, Q0957+561, J1029+2623, LSST, quasar.

1. Introduction. Quasars are the most luminous active galaxies in the universe that host an accreting supermassive black hole at the center. The path that light takes from a quasar to Earth can be altered by the gravitational field of a massive intervening galaxy, acting as a lens and bending the trajectory of the emitted light; see the first panel of Figure 1. When the quasar, lensing galaxy, and Earth are geometrically aligned, multiple images of the quasar can appear in slightly different locations in the sky, from the perspective of an observer on Earth. This phenomenon is known as strong gravitational lensing [Schneider, Ehlers and Falco (1992); Schneider, Wambsganss and Kochanek (2006)]. In this case, there are typically two or more replicate images, referred to as doubly- or multiply-lensed quasars. Since quasars are highly luminous, they can be seen at great distances, which both enhances the possibility of lensing by an intervening galaxy and makes them useful for cosmology.

The light rays forming each of these gravitationally lensed quasar images take different routes from the quasar to Earth. Since both the lengths of the pathways and the gravitational potentials they traverse differ, the resulting multiple images are subject to differing lensing magnifications and their light rays arrive at the observer at different times. Because of this, any fluctuations in the source brightness are observed in each image at different times. From a statistical perspective, we can construct a time series of the brightness of each image, known as *a light curve*. Features in these light curves appear to be shifted in time and these shifts are called *time delays*.

Obtaining accurate time delay estimates is important in cosmology because they can be used to address fundamental questions regarding the origin and evolution of the universe. For instance, Refsdal (1964) suggested using time delay estimates to constrain the Hubble constant H_0, the current expansion rate of the universe; given a model for the mass distribution and gravitational potential of the lensing galaxy, the time delay between multiple images of the lensed quasar is inversely proportional to H_0 [Blandford and Narayan (1992), Suyu et al. (2013), Treu and

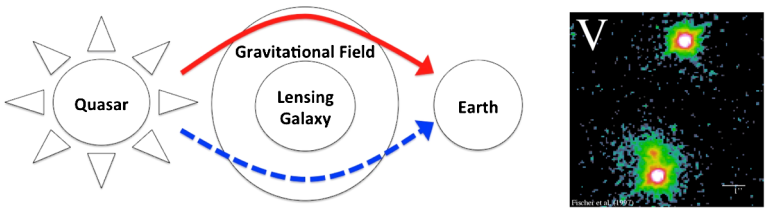

FIG. 1. *The gravitational field of an intervening galaxy acts as a lens deflecting two light rays of a quasar image toward the Earth as shown in the left panel. The arrival times can differ owing to the different lengths of pathways and different gravitational potentials they pass through. An optical V-band image of the doubly-lensed quasar Q*0957+561 *obtained with the Canada France Hawaii telescope* [*Fischer et al.* (1997); *Munoz et al.* (1998)] (*https://www.cfa.harvard.edu/castles*) *appears in the right panel. The two bright sources at the top and bottom are the lensed images of the quasar, and the small red point toward the top-left of the lower quasar image is the lensing galaxy.*

Marshall (2016)]. Also, Linder (2011) showed that an accurate time delay estimate could substantially constrain cosmological parameters and the equation of state of dark energy characterizing the accelerated expansion of the Universe.

The upcoming large-scale astronomical survey to be conducted with the Large Synoptic Survey Telescope [LSST, LSST Science Collaboration (2009)] will monitor thousands of gravitationally lensed quasars beginning in 2022. The LSST is the top-ranked ground-based telescope project in the 2010 Astrophysics Decadal Survey, and will produce extensive high-cadence time series observations of the full sky for ten years. The LSST will produce multi-band light curves (observed via multiple optical filters centered at different wavelengths) that form a vector time series for each image. In preparation for the era of the LSST, Dobler et al. (2015) organized a blind competition called the Time Delay Challenge (TDC) which ran from October 2013 to July 2014 with the aim of improving time delay estimation methods for application to realistic observational data sets. As a simplification for the first competition, the TDC organizers simulated thousands of single-band datasets, that is, scalar time series for each image, that mimic real quasar data. We are among 13 teams who took part in the TDC, each of which analyzed the simulated data using their own methods to estimate the blinded time delays.[7]

1.1. *Data and challenges*. We plot a pair of simulated light curves from a doubly-lensed quasar in Figure 2; the light curves are labeled as A and B. Each observation time is denoted by vertical dashed lines, at which the observer measures the brightness of each gravitationally lensed quasar image. In a real data analysis, these images would correspond to the two bright sources in the second panel of Figure 1. The brightness is reported on the magnitude scale, an astronomical logarithmic measure of brightness, in which smaller numbers correspond to brighter objects. The magnitudes in Figure 2 are presented up to an overall additive calibration constant as was the case in the TDC. Since the time delay is estimated via relative comparison between fluctuations in the two light curves, our analysis is insensitive to this overall additive constant.

For a doubly-lensed quasar, there are four variables recorded on an irregularly spaced sequence of observation times $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)^\top$; the observed magnitudes $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\top$ for light curve A and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ for light curve B as well as standard deviations, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^\top$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^\top$, representing their uncertainties due to heteroskedastic measurement error. In Figure 2, $\boldsymbol{x}$ and $\boldsymbol{y}$ are represented by red squares and blue circles,

[7]In the last stage of the TDC (called *rung4* in the TDC), an earlier version of our method achieved the smallest average coefficient of variation (*precision*), the TDC target for the average error level (*accuracy*) within one standard deviation, and acceptable average squared standardized residual (χ^2) after analyzing the second highest number of data sets (f). See Liao et al. (2015) for detailed results of the TDC.

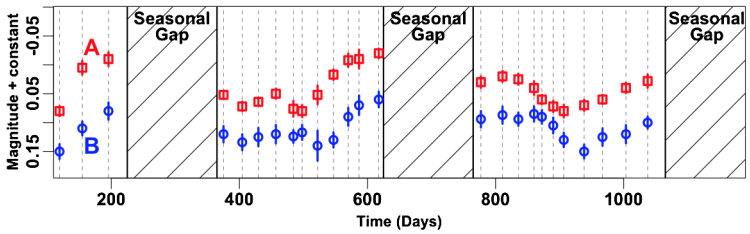

FIG. 2. *The red squares and blue circles indicate the observed magnitudes of the two simulated images at each observation time. The half lengths of vertical lines around the symbols represent the uncertainties* (*standard deviations*) *of the observed magnitudes. The convention in Astronomy is to plot the magnitude inversely so that smaller magnitudes* (*brighter object*) *appear on the top and larger ones* (*fainter object*) *on the bottom. The quasar magnitudes are vertically offset by an overall calibration constant, the value of which is unimportant for time delay estimation.*

and their standard deviations by the half-lengths of vertical lines around the symbols. Similarly, for a quadruply-lensed quasar, there are four light curves, each with their own measurement errors.

Since a quasar exhibits fluctuations in its brightness, it is possible to estimate time delays between different views of those fluctuations. In Figure 2, for example, the bottom of the V-shaped valley of light curve *A* at around 900 days precedes that of light curve B by around 50 days. Other features in the light curves exhibit a similar time shift of about 50 days.

However, a number of aspects of the light curves in Figure 2 make accurate time delay estimation statistically challenging. First, irregular observation times are inevitable because observations may be prevented in poor weather or during the day. Second, the motion of the Earth around the Sun causes seasonal gaps because the part of the sky containing the quasar is not visible at night from the location of a particular telescope during certain months. Third, since the light of each gravitationally lensed image traverses different paths through the gravitational potential, they are subject to differing degrees of lensing magnification. Thus, the light curves often exhibit different average magnitudes. Finally, observed magnitudes are measured with error, leading to relatively larger measurement errors for fainter images.

Moreover, some quasar images exhibit additional independent extrinsic variability, an effect called *microlensing*.[8] Significant microlensing occurs when a path of light passes unusually close to a star that is moving within the lensing galaxy. Lensing by this star introduces independent brightness magnification variations

[8]Microlensing is conceptually similar to strong lensing except that the lens is a star moving within the intervening galaxy. However, the lensed images produced by microlensing cannot be separately seen because their angular separation is too small for us to resolve with a telescope. Instead, astronomers observe only the combined magnification of both images, which changes with time due to the relative motions of the source and lensing star.

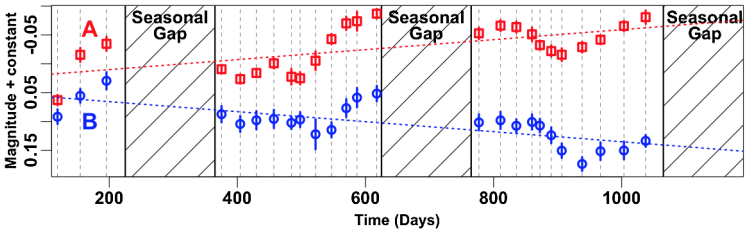

FIG. 3. *The light curves of two lensed images can have different long-term trends caused by microlensing due to stars moving within the lensing galaxy. This effect independently introduces a long-term magnification trend in each image. Here, we simulate the effect of two different long-term linear microlensing trends on the light curves in Figure 2. The dotted lines depict the linear microlensing trend for each image.*

into the corresponding image in addition to the overall magnifications caused by strong lensing of the galaxy [Chang and Refsdal (1979); Tewes, Courbin and Meylan (2013)]. The timescale of the microlensing variability is typically much larger than that of the intrinsic quasar variability if the lens is on a galaxy scale [Liao et al. (2015)]. Thus, the individual light curves may exhibit different long-term trends that are not related to the intrinsic variability of the source.[9] In Figure 3, as an illustration, we plot the same simulated light curves A and B plotted in Figure 2 but with different added linear trends to simulate the effect of microlensing.

1.2. *Other time delay estimation methods*. Conventional methods for time delay estimation have involved grid-based searches. One-dimensional grid methods estimate the time delay, Δ_{AB},[10] by minimizing the χ^2 distance or by maximizing the cross-correlation between two light curves, $\boldsymbol{x}$ and $\boldsymbol{y}_{\Delta_{\mathrm{AB}}}$, on a grid of values of Δ_{AB} [Fassnacht et al. (1999)], where $\boldsymbol{y}_{\Delta_{\mathrm{AB}}}$ denotes $\boldsymbol{y}$ shifted by Δ_{AB} days to the right. Both techniques require an interpolation scheme. The dispersion method [Pelt et al. (1994)] combines two light curves by shifting one of them in time and magnitude by Δ_{AB} and β_0, respectively. This is called the *curve-shifting* assumption. The method estimates Δ_{AB} and β_0 on a two-dimensional grid by minimizing the sum of squared differences between consecutive pairs of magnitudes on the combined curve. A bootstrapping method is used to produce standard errors of the time delay estimates. These methods account only for the intrinsic variability of a

[9] MacLeod et al. (2010) who analyzed about 9000 quasars obtained from the Sloan Digital Sky Survey [Berk et al. (2004)] show that the timescale of quasar intrinsic variability varies from days to years, and Mosquera and Kochanek (2011) indicate that the five shortest timescales of microlensing among 87 lensed quasars are between 8 and 12 years with respect to Einstein crossing timescales and are between 1 and 8 weeks with respect to source crossing timescale. Since the microlensing timescale is not always longer than the quasar intrinsic variability timescale, it is not always the case that we see the extrinsic long-term trends in the presence of microlensing.

[10] A positive value of Δ_{AB} indicates that features in light curve A appear before they appear in light curve B.

quasar. (When it is clear from the context, we suppress the subscript on Δ_{AB} and simply use Δ.)

Model-based methods have also been proposed in the past to avoid the computational burden of evaluating the fit on a fine grid. For example, Tewes, Courbin and Meylan (2013) model the intrinsic and extrinsic variabilities of light curves using high-order and low-order splines, respectively. They obtain the least square estimate of Δ by iterating a two-step fitting routine in which splines are first fit given Δ and then Δ is optimized given the model fit. They also use parametric bootstrapping for the standard error of the time delay estimate.

Harva and Raychaudhury (2006) (hereafter H&R) introduced the first fully Bayesian approach, though they do not account for microlensing. They assume each observed light curve is generated by an unobserved underlying process. One of the latent processes is assumed to be a shifted and scaled version of the other, with the time and magnitude shifts and the magnitude scale treated as unknown parameters. They use a collapsed Gibbs-type sampler for model fitting, with the latent process integrated out of the target posterior distribution. Unlike other existing methods, this approach unifies parameter estimation and uncertainty quantification into a single coherent analysis based on the posterior distribution of Δ.

1.3. *Our Bayesian and profile likelihood approaches*. The TDC motivated us to improve on H&R's fully Bayesian model by taking advantage of modeling and computational advances made since H&R's 2006 proposal. Specifically, we adopt an Ornstein–Uhlenbeck (O–U) process [Uhlenbeck and Ornstein (1930)] to model the latent light curve. The O–U process has been empirically shown to describe the stochastic variability of quasar data well [Kelly, Bechtold and Siemiginowska (2009); Kozłowski et al. (2010); MacLeod et al. (2010); Zu et al. (2013)]. We address the effect of microlensing by incorporating a polynomial regression on time into the model. We specify scientifically motivated prior distributions and conduct a set of systematic sensitivity analyses; see Appendix E for details of the sensitivity analyses. In contrast to H&R's strategy of sampling from a marginal distribution with the latent process integrated out, we use a Metropolis–Hastings (M–H) within the Gibbs sampler [Tierney (1994)] to sample the posterior in the full parameter space. We improve the convergence rate of our MCMC (Markov chain Monte Carlo) sampler by using an ancillarity-sufficiency interweaving strategy [Yu and Meng (2011)] and adaptive MCMC [Brooks et al. (2011)].

To complement the Bayesian method, we introduce a simple profile likelihood approach that allows us to remove nuisance parameters and focus on Δ [e.g., Davison (2003)]. We show that the profile likelihood function of Δ is approximately proportional to the marginal posterior distribution of Δ when a Jeffreys' prior is used for the nuisance parameters [Berger, Liseo and Wolpert (1999)]; see Appendix D. For the problems we investigate, the profile likelihood is nearly identical to the marginal posterior distribution in most cases, validating the approximation.

Our time delay estimation strategy combines these two complementary approaches. We first obtain the profile likelihood of Δ, which is simple to compute. A more principled fully Bayesian analysis focuses on the dominant mode identified by the profile likelihood and provides joint inference for the time delay and other model parameters via the joint posterior distribution.

The rest of this paper is organized as follows. We describe our Bayesian model in Section 2 and the MCMC sampler that we use to fit it in Section 3. In Section 4, we introduce the profile likelihood approach. We then specify our estimation strategy and illustrate it via a set of numerical examples in Section 5. An R package, `timedelay`, that implements the Bayesian and profile likelihood methods is publicly available at CRAN.[11]

2. A fully Bayesian model for time delay estimation.

2.1. *Latent time series.* We assume that each time-delayed light curve is generated from a latent curve representing the true source magnitude in continuous time. We denote these latent curves by $\boldsymbol{X} = \{X(t),\ t \in \mathbf{R}\}$ and $\boldsymbol{Y} = \{Y(t),\ t \in \mathbf{R}\}$, respectively, where $X(t)$ and $Y(t)$ are unobserved true magnitudes at time t. We use the vector notation $\boldsymbol{X}(\boldsymbol{t}) = (X(t_1), X(t_2), \ldots, X(t_n))^\top$ and $\boldsymbol{Y}(\boldsymbol{t}) = (Y(t_1), Y(t_2), \ldots, Y(t_n))^\top$ to denote the n magnitudes of each latent light curve at the irregularly-spaced observation times $\boldsymbol{t}$.

A curve-shifted model [Pelt et al. (1994); Kochanek et al. (2006)] assumes that one of the latent light curves is a shifted version of the other, that is,

$$Y(t) = X(t - \Delta) + \beta_0, \tag{2.1}$$

where Δ is a shift in time and β_0 is a magnitude offset, for example, in Figure 4, we displayed the solid red and dashed blue latent curves of images A and B, respectively, generated under the model in (2.1). Thus, the two curves exactly overlap if

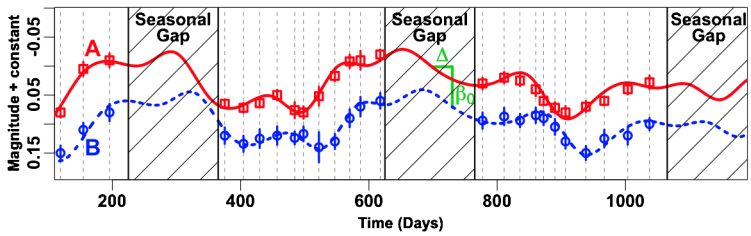

FIG. 4. *The solid red and dashed blue latent curves of images* A *and* B, *respectively, are generated under the model in* (2.1). *These two curves are superimposed on Figure* 2. *The curve-shifted model in* (2.1) *specifies that the dashed blue curve is a shifted version of the solid red curve by* Δ (=70) *days in time and by* β_0 (=0.07) *in magnitude. For illustration purposes,* $\boldsymbol{X}$ *is depicted as a solid red smooth curve*; *a more realistic model is described in Section* 2.3.

[11] https://cran.r-project.org/package=timedelay

the solid red curve is shifted by Δ days and by β_0 magnitude units. (For illustration purposes, $\boldsymbol{X}$ is depicted as a solid red smooth curve; a more realistic model is described in Section 2.3.) The key advantage of this model is that a single latent light curve, here $\boldsymbol{X}$, is sufficient to represent the true magnitude time series of the two (or more) lensed images. This model is a special case of H&R's scaled curve-shifted model, $Y(t) = sX(t-\Delta)+\beta_0$, where s is a magnitude scale change, mentioned at the end of Section 1.2. Setting $s = 1$ is reasonable because gravitational lensing only deflects the source light and magnifies it, that is, multiplies the source flux. Because magnitude is on the $\log_{10}$ scale of source flux, we expect an additive offset, that is, β_0, rather than a scale change. The curve-shifted model captures the essential physical effects of strong gravitational lensing (at least in the absence of microlensing), and thus is an appropriate model for estimating the time delay.

Microlensing causes additional long-term extrinsic variability unrelated to the intrinsic quasar variability driving the dynamics of $\boldsymbol{X}$. Thus, the curve-shifted model is not appropriate in the presence of microlensing. To account for microlensing, we assume that one of the latent light curves is a time-shifted version of the other, but with an additional polynomial regression of order m on $t-\Delta$, that is,

$$Y(t) = X(t-\Delta) + \boldsymbol{w}_m^\top(t-\Delta)\boldsymbol{\beta}, \tag{2.2}$$

where $\boldsymbol{w}_m(t-\Delta) \equiv (1, t-\Delta, (t-\Delta)^2, \ldots, (t-\Delta)^m)^\top$ is a covariate vector of length $m+1$, and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2, \ldots, \beta_m)^\top$ is a vector of regression coefficients.[12] The polynomial regression term in (2.2) accounts for the difference in the microlensing trends of the two light curves, that is, the difference between the long-term trends of $Y(t)$ and $X(t-\Delta)$. The microlensing model in (2.2) reduces to a curve-shifted model in (2.1) if $\beta_1 = \beta_2 = \cdots = \beta_m = 0$.

The best choice for the order of the polynomial regression depends on the extent of microlensing, and this varies from quasar to quasar. We set $m = 3$ as a default because the third-order polynomial regression has been successfully applied to model lensed quasars [Kochanek et al. (2006); Courbin et al. (2013); Morgan et al. (2012)]. If we find evidence via the profile likelihood that a third-order polynomial regression is not sufficient to reduce the effect of microlensing (see Section 5.1 for details), we can impose a reasonable upper bound of m by running preliminary regression on the observed light curves, and comparing the fits.

2.2. *Distribution of the observed data.* Observing the gravitationally-lensed images with a telescope, an astronomer measures the magnitude in each image, x_j and y_j, and reports standard deviations, δ_j and η_j, representing the uncertainties

[12] An orthonormal basis is more compatible with an independent prior on the regression coefficients, and thus may be preferred if a higher degree polynomial regression is used.

of the magnitudes due to measurement errors[13] at time t_j, $j = 1, 2, \ldots, n$. We assume that these measurements have independent Gaussian errors centered at the latent magnitudes $X(t_j)$ and $Y(t_j)$, that is,

$$x_j | X(t_j) \overset{\text{indep.}}{\sim} \mathrm{N}\big[X(t_j), \delta_j^2\big], \tag{2.3}$$

$$y_j | Y(t_j) \overset{\text{indep.}}{\sim} \mathrm{N}\big[Y(t_j), \eta_j^2\big], \tag{2.4}$$

where $\mathrm{N}[M, V]$ is a Gaussian distribution with mean M and variance V, and $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent given their true magnitudes. Using the model in (2.2), we can express (2.4) as

$$y_j | X(t_j - \Delta), \Delta, \boldsymbol{\beta} \overset{\text{indep.}}{\sim} \mathrm{N}\big[X(t_j - \Delta) + \boldsymbol{w}_m^\top (t_j - \Delta)\boldsymbol{\beta}, \eta_j^2\big]. \tag{2.5}$$

Given Δ, we define $\boldsymbol{t}^\Delta = (t_1^\Delta, t_2^\Delta, \ldots, t_{2n}^\Delta)^\top$ as the sorted vector of $2n$ times among the n observation times, $\boldsymbol{t}$, and the n time-delay-shifted observation times, $\boldsymbol{t} - \Delta$. Also, $\boldsymbol{X}(\boldsymbol{t}^\Delta) = (X(t_1^\Delta), X(t_2^\Delta), \ldots, X(t_{2n}^\Delta))^\top$ is the vector of $2n$ latent magnitudes at the times in $\boldsymbol{t}^\Delta$. The joint density function of the observed data given $\boldsymbol{X}(\boldsymbol{t}^\Delta)$, Δ and $\boldsymbol{\beta}$ is

$$p\big(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{X}(\boldsymbol{t}^\Delta), \Delta, \boldsymbol{\beta}\big) = \prod_{j=1}^{n} p\big(x_j | X(t_j)\big) \times p\big(y_j | X(t_j - \Delta), \Delta, \boldsymbol{\beta}\big), \tag{2.6}$$

where the two distributions in the product are given in (2.3) and (2.5).

2.3. *Prior distribution of the latent magnitudes.* We assume the latent continuous-time light curve, $\boldsymbol{X}$, is a realization of an O–U process [Uhlenbeck and Ornstein (1930)] as proposed in Kelly, Bechtold and Siemiginowska (2009). The stochastic differential equation,

$$dX(t) = -\frac{1}{\tau}\big(X(t) - \mu\big)\, dt + \sigma\, dB(t), \tag{2.7}$$

defines the O–U process, where μ and σ are on the magnitude scale and govern the overall mean and short-term variability of the underlying process, τ is a timescale (in days) for the process to revert to the long-term mean μ, $\{B(t),\ t \geq 0\}$ is a standard Brownian motion and $dB(t)$ is an interval of the Brownian motion, whose distribution is Gaussian with mean zero and variance dt. We denote the three O–U parameters by $\boldsymbol{\theta} = (\mu, \sigma^2, \tau)^\top$.

Kelly, Bechtold and Siemiginowska (2009) empirically demonstrated that the power spectrum of the O–U process is consistent with the mean power spectrum of 55 well-sampled quasar light curves at a specific frequency range with

[13]The magnitude estimate and standard deviation typically summarize a Gaussian approximation to the likelihood of the latent magnitude for the flux data of an image. The standard deviation does not necessarily represent a standard error of a repeated sampling measurement error distribution.

timescales shorter than τ. They also investigated the associations between model parameters and the physical properties of quasars, for example, τ has a positive correlation with black hole mass, which is consistent with previous astrophysical studies. Kozłowski et al. (2010) and MacLeod et al. (2010) were concerned about a possible selection bias in the sample of quasars used in Kelly, Bechtold and Siemiginowska (2009), and thus they analyzed thousands of light curves. Kozłowski et al. (2010) found further support for the O–U process in their analyses of about 2700 quasars obtained from the Optical Gravitational Lensing Experiment [OGLE, Kozłowski and Kochanek (2009)]. They showed that the distribution of the goodness-of-fit statistic obtained by fitting the O–U process to their light curves was consistent with the expected distribution of the statistic under the assumption that the light curve variation was stochastic. MacLeod et al. (2010) further verified the argument about the correlations between model parameters and physical properties in Kelly, Bechtold and Siemiginowska (2009) by analyzing about 9000 quasars obtained from the Sloan Digital Sky Survey [Berk et al. (2004)]. Zu et al. (2013) also supported the O–U process by comparing it to the Gaussian process with three different covariance functions in fitting about 200 OGLE light curves. Their numerical results based on the F-test and Bayesian information criterion supported the O–U process. These studies popularized the O–U process among astrophysicists to the extent that the TDC simulated its quasar light curves under an O–U process[14] [Dobler et al. (2015)]. The earlier approach of H&R (2006) preceded these more recent advances in astrophysical and statistical modeling of quasars.

The solution of the stochastic differential equation in (2.7) provides the prior distribution for the time-sorted latent magnitudes $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ via its Markovian property. Specifically,

$$
\begin{aligned}
X\bigl(t_1^{\Delta}\bigr)|\Delta,\boldsymbol{\theta} &\sim \mathrm{N}\biggl[\mu,\frac{\tau\sigma^2}{2}\biggr], \qquad \text{and for } j=2,3,\ldots,2n,\\
X\bigl(t_j^{\Delta}\bigr)|X\bigl(t_{j-1}^{\Delta}\bigr),\Delta,\boldsymbol{\theta} &\sim \mathrm{N}\biggl[\mu+a_j\bigl(X\bigl(t_{j-1}^{\Delta}\bigr)-\mu\bigr),\frac{\tau\sigma^2}{2}\bigl(1-a_j^2\bigr)\biggr],
\end{aligned}
\tag{2.8}
$$

where $a_j \equiv \exp(-(t_j^{\Delta}-t_{j-1}^{\Delta})/\tau)$ is a shrinkage factor that depends on the observational cadence and τ. If two adjacent latent magnitudes are close in time, that is,

[14]The TDC organizers generated 500 10-year-long light curves by using the O–U process ($\mu=0$, $\log(\tau)\in[1.5,3.0]$, and $\log(\sigma)\in[-1.1,-0.3]$), and re-used these to make about 5000 doubly- or quadruply-lensed light curves with different starting points, different seasonal gaps, etc. Microlensing is simulated via a catalog convergence of Oguri and Marshall (2010), shear, and surface density. The measurement errors were heteroskedastic Gaussian. The organizers intentionally contaminated the data to make the time delay estimation difficult; the reported standard deviations may be underestimated, measurement errors may be correlated due to time-dependent calibration error, and magnitudes may be temporarily offset due to time-dependent systematic effects in the telescope optics.

$t_j^{\Delta} - t_{j-1}^{\Delta}$ is small, a_j is close to unity and under this prior $X(t_j^{\Delta})$ borrows more information or shrinks more toward the previous latent magnitude, $X(t_{j-1}^{\Delta})$, and exhibits less uncertainty. On the other hand, if neighboring latent magnitudes are distant in time, for example, due to a seasonal gap, a_j is close to zero, and under this prior $X(t_j^{\Delta})$ borrows little information from the distant value $X(t_{j-1}^{\Delta})$ and instead approaches the overall mean μ with more uncertainty. This is known as *the mean reversion* property of the O–U process.

The joint prior density function of the $2n$ latent magnitudes is

$$p(\boldsymbol{X}(\boldsymbol{t}^{\Delta})|\Delta,\boldsymbol{\theta}) = p(X(t_1^{\Delta})|\Delta,\boldsymbol{\theta}) \times \prod_{j=2}^{2n} p(X(t_j^{\Delta})|X(t_{j-1}^{\Delta}),\Delta,\boldsymbol{\theta}), \tag{2.9}$$

where the distributions on the right-hand side are given in (2.8).

2.4. *Prior distributions for the time delay and the magnitude offset.* We adopt independent proper prior distributions for Δ and $\boldsymbol{\beta}$,

$$p(\Delta,\boldsymbol{\beta}) = p(\Delta)p(\boldsymbol{\beta}) \propto I_{\{u_1\le\Delta\le u_2\}} \times \mathrm{N}_{m+1}(\boldsymbol{\beta}|\mathbf{0}, 10^5 \times I_{m+1}), \tag{2.10}$$

where $I_{\{D\}}$ is the indicator function of D, $\mathrm{N}_{m+1}(\boldsymbol{\beta}|\mathbf{0}, 10^5 \times I_{m+1})$ is an $m+1$ dimensional Gaussian density evaluated at $\boldsymbol{\beta}$ whose mean is $\mathbf{0}$, a vector of zeros with length $m+1$, and variance-covariance matrix is $10^5 \times I_{m+1}$, with an $m+1$ dimensional identity matrix I_{m+1}. We put a diffuse Gaussian prior on $\boldsymbol{\beta}$ to minimize impact on the posterior inference and to ensure posterior propriety.

The range of the uniform prior distribution on Δ, $[u_1, u_2]$, reflects the range of interest. One choice is the *entire feasible range* (or feasible range) of Δ, $[t_1 - t_n, t_n - t_1]$; only values of Δ in this range can correspond to adjusted light curves that overlap by at least one data point. (H&R uses a diffuse Gaussian prior distribution on Δ that is defined even outside this range.)

In some cases, information about the likely range of Δ is available from previous analyses or possibly from astrophysical probes, for example, we can find the likely range of Δ using a physical model for the mass and gravitational potential of the lens, as well as the redshifts (an astronomical measure of distance) and relative spatial locations of a quasar and lens.

In reality, the time delay and lensing magnification may be correlated a priori. We assume a priori independence, however, because it is difficult to construct an informative joint prior distribution without more information about the lens system, that is, image positions, distances and a lens model.

2.5. *Prior distributions for the parameters in the O–U process.* Considering both scientific knowledge and the dynamics of the O–U process, we put a uniform distribution on the O–U mean μ, an independent inverse-Gamma (IG) distribution,

IG$(1, b_\sigma)$, on its short-term variance σ^2 and an independent IG$(1, b_\tau)$ distribution on its timescale τ, that is,

$$\begin{aligned} p(\mu, \sigma^2, \tau) &= p(\mu)p(\sigma^2)p(\tau) \\ &\propto \frac{\exp(-b_\sigma/\sigma^2)}{(\sigma^2)^2} \times \frac{\exp(-b_\tau/\tau)}{\tau^2} \\ &\quad \times I_{\{-30\le\mu\le 30\}} \times I_{\{\sigma^2>0\}} \times I_{\{\tau>0\}}. \end{aligned} \tag{2.11}$$

The units of b_σ are magnitude squared per day, hereafter mag^2/day, and the scale parameter of the IG distribution on τ is fixed at one day, that is, $b_\tau = 1$ day.

Here, the uniform distribution on μ encompasses a magnitude range from that of the Sun (magnitude $= -26.74$) to that of the faintest object visible with the Hubble Space Telescope (magnitude $= 30$). The IG distributions on τ and σ^2 set soft lower bounds[15] to focus on practical solutions in which Δ can be constrained. For example, in the limits when τ is much less than the observation cadence or when σ^2 is much smaller than the measurement variance divided by the cadence, the discrete observations of the continuous latent light curve appear as serially uncorrelated white noise sequence. In these limiting cases, it is impossible to estimate Δ by matching serially correlated fluctuation patterns. The soft lower bounds for τ and σ^2 discount these limiting cases, and allow us to focus on the relevant parameter space in which we expect time delay estimation to be feasible.

We set the shape parameter of the IG prior distribution on τ to unity and the scale parameter b_τ to one day to obtain a weakly informative prior. The resulting soft lower bound on τ is 0.5 day and is smaller than all of the estimates of τ in MacLeod et al. (2010), who analyzed 9275 quasars.

For the IG prior distribution of σ^2, we set the shape parameter to unity and the scale parameter to (Mean measurement standard deviation)^2/(Median cadence), that is,

$$b_\sigma = \frac{[\{\sum_{j=1}^n \delta_j + \sum_{j=1}^n \eta_j\}/2n]^2}{\text{Median}(t_2 - t_1, t_3 - t_2, \ldots, t_n - t_{n-1})}. \tag{2.12}$$

This scale parameter enables us to search for solutions for which we can constrain Δ by avoiding the above limiting case. Another viable choice for the scale parameter is $b_\sigma = 2 \times 10^{-7}$ because all estimates of σ^2 in MacLeod et al. (2010) are larger than this value. Sensitivity analyses for the choice of prior distributions of τ and σ^2 appear in Appendix E.

[15] Because the density function of IG(a, b) decreases exponentially from its mode, $b/(a+1)$, toward zero and geometrically decreases with a power of $a + 1$ toward infinity, it is relatively unlikely for the random variable to take on values much smaller than its mode.

3. Metropolis–Hastings within Gibbs sampler. Our overall hierarchical model is specified via the observation model in (2.3) and (2.5), the O–U process for the latent light curve in (2.8) and the prior distributions given in (2.10) and (2.11). Our first approach to model fitting uses a Gibbs-type sampler to explore the resulting full posterior distribution. It is possible to integrate out the latent magnitudes analytically and use a collapsed sampler based on the marginalized joint posterior distribution specified in Appendix A as H&R did. However, we treat $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ as latent variables, alternatively updating $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ and the other model parameters. [We could formulate our approach as data augmentation with $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ as the missing data; see van Dyk and Meng (2001).]

Specifically, we use a Metropolis–Hastings within Gibbs (MHwG) sampler [Tierney (1994)] that iteratively samples five complete conditional distributions of the full joint posterior density, $p(\boldsymbol{X}(\boldsymbol{t}^{\Delta}), \Delta, \boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$, proportional to the product of densities of observed data in (2.6) and prior densities in (2.9), (2.10) and (2.11). Iteration l of our sampler is composed of five steps:

$$\begin{aligned}\text{Step 1: Sample } &\bigl(\boldsymbol{X}^{(l)}\bigl(\boldsymbol{t}^{\Delta^{(l)}}\bigr), \Delta^{(l)}\bigr) \sim p\bigl(\boldsymbol{X}(\boldsymbol{t}^{\Delta}), \Delta|\boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}\bigr)\\ &= p\bigl(\boldsymbol{X}(\boldsymbol{t}^{\Delta})|\Delta, \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}\bigr) \times p\bigl(\Delta|\boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}\bigr)\\ &\qquad\text{by M–H,}\end{aligned} \tag{3.1}$$

$$\text{Step 2: Sample } \boldsymbol{\beta}^{(l)} \sim p\bigl(\boldsymbol{\beta}|\boldsymbol{\theta}^{(l-1)}, \boldsymbol{X}^{(l)}\bigl(\boldsymbol{t}^{\Delta^{(l)}}\bigr), \Delta^{(l)}\bigr), \tag{3.2}$$

$$\text{Step 3: Sample } \mu^{(l)} \sim p\bigl(\mu|(\sigma^2)^{(l-1)}, \tau^{(l-1)}, \boldsymbol{X}^{(l)}\bigl(\boldsymbol{t}^{\Delta^{(l)}}\bigr), \Delta^{(l)}, \boldsymbol{\beta}^{(l)}\bigr), \tag{3.3}$$

$$\text{Step 4: Sample } \bigl(\sigma^2\bigr)^{(l)} \sim p\bigl(\sigma^2|\tau^{(l-1)}, \boldsymbol{X}^{(l)}\bigl(\boldsymbol{t}^{\Delta^{(l)}}\bigr), \Delta^{(l)}, \boldsymbol{\beta}^{(l)}, \mu^{(l)}\bigr), \tag{3.4}$$

$$\begin{aligned}&\text{Step 5: Sample } \tau^{(l)} \sim p\bigl(\tau|\boldsymbol{X}^{(l)}\bigl(\boldsymbol{t}^{\Delta^{(l)}}\bigr), \Delta^{(l)}, \boldsymbol{\beta}^{(l)}, \mu^{(l)}, \bigl(\sigma^2\bigr)^{(l)}\bigr)\\ &\qquad\text{by M–H,}\end{aligned} \tag{3.5}$$

where we suppress conditioning on $\boldsymbol{x}$ and $\boldsymbol{y}$ in all five steps. The conditional distributions in (3.2), (3.3) and (3.4), are standard families that can be sampled directly, whereas those in (3.1) and (3.5) require M–H updates. We use the factorization in (3.1) to construct a joint proposal, $(\tilde{\boldsymbol{X}}(\boldsymbol{t}^{\tilde{\Delta}}), \tilde{\Delta})$, for $(\boldsymbol{X}(\boldsymbol{t}^{\Delta}), \Delta)$ and calculate its acceptance probability. First, $\tilde{\Delta}$ is proposed from $\mathrm{N}(\Delta^{(l-1)}, \psi^2)$, where ψ is a proposal scale and is set to produce a reasonable acceptance rate. Given $\tilde{\Delta}$, we propose $\tilde{\boldsymbol{X}}(\boldsymbol{t}^{\tilde{\Delta}}) \sim p(\boldsymbol{X}(\boldsymbol{t}^{\tilde{\Delta}})|\tilde{\Delta}, \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}, \boldsymbol{x}, \boldsymbol{y})$; this is a Gaussian distribution and is specified in Appendix B. Because the proposal for Δ and that for $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ given Δ are symmetric, $(\tilde{\boldsymbol{X}}(\boldsymbol{t}^{\tilde{\Delta}}), \tilde{\Delta})$ is accepted with a probability $\min(1, r)$, where

$$r = \frac{p(\tilde{\Delta}|\boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}, \boldsymbol{x}, \boldsymbol{y})}{p(\Delta^{(l-1)}|\boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}, \boldsymbol{x}, \boldsymbol{y})}. \tag{3.6}$$

Details of the marginalized density $p(\Delta|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})$ in (3.6) appear in Appendix A and details of Steps 2–5 appear in Appendix C.

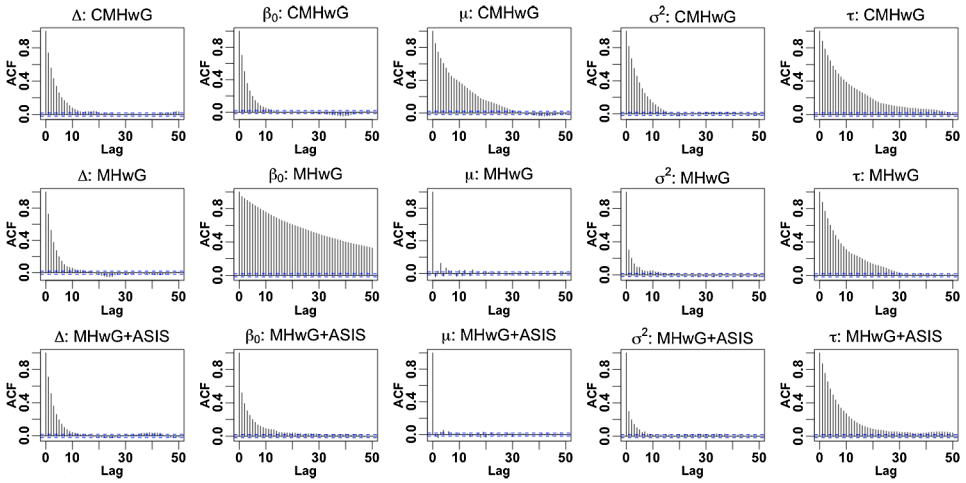

FIG. 5. *The autocorrelation functions for* $\Delta, \beta_0, \mu, \sigma^2$ *and* τ (*columns from left to right*) *based on* 10,000 *posterior samples after a burn-in of* 10,000. *Results are obtained using three different posterior samplers* (*CMHwG*, *MHwG and MHwG* + *ASIS*, *rows from top to bottom*). *We use the curve-shifted model for simplicity and the data from quasar Q*0957+671 *analyzed in Section* 5.3.

The direct updates for $\boldsymbol{\beta}$, μ and σ^2 are based on standard families that are not available using H&R's collapsed approach. Thus, the collapsed approach must update each of the model parameters via a Metropolis or M–H update, which can slow down convergence. (Collapsing Gibbs-type samplers, however, is known to improve their rate of convergence [Liu (2008)] if the complete conditionals can be sampled directly.) Also, the collapsed MHwG (CMHwG) sampler requires about three times more CPU time per iteration than the (noncollapsed) MHwG sampler that we propose. In Figure 5, we compare the autocorrelation functions (ACFs) of Δ, β_0, μ, σ^2 and τ obtained by the CMHwG sampler (first row) and those obtained by our MHwG sampler (second row). The sampler in the third row is discussed in Section 3.1. All algorithms are run using the curve-shifted model in (2.1) fit to data for quasar *Q0957+561* [Hainline et al. (2012)]. Except for that of β_0, the ACFs generated with CMHwG (first row), decay more slowly than those obtained with MHwG (second row). The effective sample sizes per second (ESS/sec) tend to improve with MHwG over CMHwG, for example, for Δ the ESS/sec is 5.23 with CMHwG and 21.09 with MHwG. The exception is β_0, for which ESS/sec is 6.33 with CMHwG, but only 1.74 with MHwG. The mixing for $\boldsymbol{\beta}$ in our microlensing model is slow in general. In the following section, we discuss a way to improve the convergence rate of β_0 (or $\boldsymbol{\beta}$ in general) for the MHwG sampler, while retaining its fast running time.

3.1. *Ancillarity-sufficiency interweaving strategy.* To improve the convergence rate of $\boldsymbol{\beta}$, we adopt the ancillarity-sufficiency interweaving strategy [ASIS,

Yu and Meng (2011)]. In a general hierarchical modeling setting, ASIS interweaves trajectories of the Markov chains obtained by two discordant parameterizations of the unknown quantities, which reduces dependence between the adjoining iterates. A different parameterization for the location parameters, for example, $\boldsymbol{\beta}$ in our case, can be derived by shifting, and that for scale parameters by rescaling. The two parameterizations are designed so that the original and transformed parameters can be viewed as ancillary and sufficient statistics for $\boldsymbol{\beta}$, respectively. ASIS is always faster to converge than the slower of the data augmentation samplers based on either of the two parameterizations and is geometrically convergent even when neither of the two data augmentation samplers is.

In the parameterization used up until now, $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ is an *ancillary augmentation* (AA) for $\boldsymbol{\beta}$ in that it is an ancillary statistic for $\boldsymbol{\beta}$, that is, the distribution of $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ in (2.8) does not depend on $\boldsymbol{\beta}$. On the other hand, a *sufficiency augmentation* (SA) for $\boldsymbol{\beta}$ is based on a transformation of $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ that has sufficient information to estimate $\boldsymbol{\beta}$, that is, a sufficient statistic for $\boldsymbol{\beta}$. To derive an SA for $\boldsymbol{\beta}$, we introduce the parameterization:

$$K\big(t_j^{\Delta}\big) \equiv X\big(t_j^{\Delta}\big) + \boldsymbol{w}_m^{\top}\big(t_j^{\Delta}\big)\boldsymbol{\beta} \times I_{\boldsymbol{t}-\Delta}\big(t_j^{\Delta}\big), \qquad \text{for } j = 1, 2, \ldots, 2n, \tag{3.7}$$

where

$$I_{\boldsymbol{t}-\Delta}\big(t_j^{\Delta}\big) = \begin{cases} 1, & \text{if } t_j^{\Delta} \in \boldsymbol{t} - \Delta, \\ 0, & \text{if } t_j^{\Delta} \in \boldsymbol{t}. \end{cases} \tag{3.8}$$

This indicator is one if t_j^{Δ} is an element of $\boldsymbol{t} - \Delta = \{t_1 - \Delta, t_2 - \Delta, \ldots, t_n - \Delta\}$ and zero otherwise. Thus, $\boldsymbol{K}(\boldsymbol{t}^{\Delta})$ represents the time-sorted latent magnitudes of $\boldsymbol{X}(\boldsymbol{t})$ and of microlensing-adjusted $\boldsymbol{Y}(\boldsymbol{t})$, that is, $\boldsymbol{X}(\boldsymbol{t} - \Delta) + \boldsymbol{w}_m^{\top}(\boldsymbol{t} - \Delta)\boldsymbol{\beta}$. Using (3.7), we express the observation model in (2.3) and (2.5) as

$$x_j | K(t_j) \overset{\text{indep.}}{\sim} \mathrm{N}\big[K(t_j), \delta_j^2\big]. \tag{3.9}$$

$$y_j | K(t_j - \Delta), \Delta \overset{\text{indep.}}{\sim} \mathrm{N}\big[K(t_j - \Delta), \eta_j^2\big]. \tag{3.10}$$

The distributions for the latent light curve in (2.8) are replaced by

$$\begin{aligned} K\big(t_1^{\Delta}\big) | \Delta, \boldsymbol{\beta}, \boldsymbol{\theta} &\sim \mathrm{N}\bigg[\mu + \boldsymbol{w}_m^{\top}\big(t_1^{\Delta}\big)\boldsymbol{\beta} \times I_{\{\boldsymbol{t}-\Delta\}}\big(t_1^{\Delta}\big), \frac{\tau\sigma^2}{2}\bigg], \\ K\big(t_j^{\Delta}\big) | K\big(t_{j-1}^{\Delta}\big), \Delta, \boldsymbol{\beta}, \boldsymbol{\theta} &\\ &\sim \mathrm{N}\Big[\mu + \boldsymbol{w}_m^{\top}\big(t_j^{\Delta}\big)\boldsymbol{\beta} \times I_{\{\boldsymbol{t}-\Delta\}}\big(t_j^{\Delta}\big) \\ &\quad + a_j\big(K\big(t_{j-1}^{\Delta}\big) - \mu - \boldsymbol{w}_m^{\top}\big(t_{j-1}^{\Delta}\big)\boldsymbol{\beta} \times I_{\{\boldsymbol{t}-\Delta\}}\big(t_{j-1}^{\Delta}\big)\big), \frac{\tau\sigma^2}{2}\big(1 - a_j^2\big)\Big]. \end{aligned} \tag{3.11}$$

Under this reparameterization of the model in terms of $\boldsymbol{K}(\boldsymbol{t}^{\Delta})$, $\boldsymbol{\beta}$ appears only in (3.11), which means that $\boldsymbol{K}(\boldsymbol{t}^{\Delta})$ contains sufficient information to estimate $\boldsymbol{\beta}$,

and thus $\boldsymbol{K}(\boldsymbol{t}^{\Delta})$ is an SA for $\boldsymbol{\beta}$. In contrast, $\boldsymbol{\beta}$ appears only in the distribution of observed magnitudes in (2.6), not in that of latent magnitudes in (2.8), and thus $\boldsymbol{X}(\boldsymbol{t}^{\boldsymbol{\Delta}})$ is an AA for $\boldsymbol{\beta}$. Because the parameterization does not affect the prior distributions of the model parameters in (2.10) and (2.11), the full joint posterior density in terms of $\boldsymbol{K}(\boldsymbol{t}^{\Delta})$, that is, $p(\boldsymbol{K}(\boldsymbol{t}^{\Delta}), \Delta, \boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$, is proportional to the product of densities of observed data given in (3.9) and (3.10) and prior densities in (3.11), (2.10) and (2.11). Consequently, the marginal posterior distribution of the model parameters, $\{\Delta, \boldsymbol{\beta}, \boldsymbol{\theta}\}$, is unchanged.

ASIS interweaves the trajectory of $\boldsymbol{\beta}$ from a sample constructed under AA and that constructed under SA. This can be accomplished by replacing Step 2 in (3.2) with the following four steps:

$$\text{Step } 2_{\mathrm{a}}\text{: Sample } \boldsymbol{\beta}_{\mathrm{AA}}^{(l)} \sim p\big(\boldsymbol{\beta}|\boldsymbol{\theta}^{(l-1)}, \boldsymbol{X}^{(l)}\big(\boldsymbol{t}^{\Delta^{(l)}}\big), \Delta^{(l)}\big). \tag{3.12}$$

$$\text{Step } 2_{\mathrm{b}}\text{: Set } K^{(l)}\big(t_j^{\Delta^{(l)}}\big) = X^{(l)}\big(t_j^{\Delta^{(l)}}\big) + \boldsymbol{w}_m^{\top}\big(t_j^{\Delta^{(l)}}\big)\boldsymbol{\beta}_{\mathrm{AA}}^{(l)} I_{\boldsymbol{t}-\Delta^{(l)}}\big(t_j^{\Delta^{(l)}}\big), \tag{3.13}$$

$$\text{Step } 2_{\mathrm{c}}\text{: Sample } \boldsymbol{\beta}_{\mathrm{SA}}^{(l)} \sim p\big(\boldsymbol{\beta}|\boldsymbol{\theta}^{(l-1)}, \boldsymbol{K}^{(l)}\big(\boldsymbol{t}^{\Delta^{(l)}}\big), \Delta^{(l)}\big), \tag{3.14}$$

$$\text{Step } 2_{\mathrm{d}}\text{: Set } X^{(l)}\big(t_j^{\Delta^{(l)}}\big) = K\big(t_j^{\Delta^{(l)}}\big) - \boldsymbol{w}_m^{\top}\big(t_j^{\Delta^{(l)}}\big)\boldsymbol{\beta}_{\mathrm{SA}}^{(l)} I_{\boldsymbol{t}-\Delta^{(l)}}\big(t_j^{\Delta^{(l)}}\big). \tag{3.15}$$

Again, we suppress the conditioning on $\boldsymbol{x}$ and $\boldsymbol{y}$. In Step 2_c, we set $\boldsymbol{\beta}^{(l)}$ to $\boldsymbol{\beta}_{\mathrm{SA}}^{(l)}$ sampled from its conditional posterior distribution specified in (C.2). In Step 2_d, ASIS updates $\boldsymbol{X}^{(l)}(\boldsymbol{t}^{\Delta^{(l)}})$ to adjust for the inconsistency between the updates sampled in (3.3)–(3.5) that are based on $\boldsymbol{X}^{(l)}(\boldsymbol{t}^{\Delta^{(l)}})$ and the update $\boldsymbol{\beta}^{(l)}$ that is based on $\boldsymbol{K}^{(l)}(\boldsymbol{t}^{\Delta^{(l)}})$. Updating $\boldsymbol{X}^{(l)}(\boldsymbol{t}^{\Delta^{(l)}})$ in (3.15) synchronizes this inconsistency and preserves the stationary distribution [Yu and Meng (2011)]. The additional computational cost of ASIS is negligible because the conditional updates in (3.12) and (3.14) include quick multivariate Gaussian sampling; see (C.1) and (C.2) for details.

ACFs of the model parameters obtained by MHwG equipped with ASIS, denoted by MHwG+ASIS, appear on the third row of Figure 5; the ACF of β_0 in the second column shows a noticeable improvement compared to that obtained by the MHwG sampler. The ESS/sec for β_0 is 20.95 with MHwG+ASIS and 1.74 with MHwG. In general, ASIS improves the mixing of all the regression coefficients in our microlensing model. Although it improves the ACF for the components of $\boldsymbol{\beta}$, ASIS has little effect on the ACF of Δ. The ESS/sec for Δ is 21.35 with MHwG+ASIS and 21.09 with MHwG. This small improvement implies that the dependence between Δ and $\boldsymbol{\beta}$ may be weak a posteriori; this is confirmed by our data analyses in Section 5.1. Nevertheless, ASIS improves overall convergence of the chain which we expect to improve the reliability of all inferences based on the chain.

3.2. *Adaptive MCMC.* Our MHwG sampler (either with or without ASIS) requires a proposal distribution in each of its two Metropolis steps, that is,

Algorithm 1 Steps of the adaptive MHwG+ASIS sampler

Set $\boldsymbol{X}^{(\mathbf{0})}(\boldsymbol{t}^{\boldsymbol{\Delta}^{(\mathbf{0})}})$, $\Delta^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\mu^{(0)}$, $(\sigma^2)^{(0)}$, $\tau^{(0)}$, $\psi^{(0)}$, $\phi^{(0)}$.
For $l = 1, 2, \ldots$
Step 1: Sample $\Delta^{(l)}$ using a Metropolis step with proposal rule $\mathrm{N}[\Delta^{(l-1)}, (\psi^{(l-1)})^2]$.
  If a new proposal for $\Delta^{(l)}$ is accepted, then sample $\boldsymbol{X}^{(l)}(\boldsymbol{t}^{\Delta^{(l)}})$,
  or otherwise set $\boldsymbol{X}^{(l)}(\boldsymbol{t}^{\Delta^{(l)}})$ to $\boldsymbol{X}^{(l-1)}(\boldsymbol{t}^{\Delta^{(l-1)}})$.
Step 2: (ASIS) Update $\boldsymbol{\beta}^{(l)}$ and $\boldsymbol{X}^{(l)}(\boldsymbol{t}^{\Delta^{(l)}})$ via (3.12)–(3.15).
Step 3: Sample $\mu^{(l)}$ via (3.3).
Step 4: Sample $(\sigma^2)^{(l)}$ via (3.4).
Step 5: Sample $\tau^{(l)}$ using an M–H step with proposal rule $\mathrm{N}[\log(\tau^{(l-1)}), (\phi^{(l-1)})^2]$.
Step 6: (Adaptation) If $l \bmod 100 = 0$
  if the acceptance rate of Δ in iterations $l-99, l-98, \ldots, l > 0.44$ **then**
    $\psi^{(l)} \leftarrow \psi^{(l-1)} \times \exp(\min(0.01, 1/\sqrt{(l/100)}))$
  else if the acceptance rate of Δ in iterations $l-99, l-98, \ldots, l < 0.23$ **then**
    $\psi^{(l)} \leftarrow \psi^{(l-1)} \times \exp(-\min(0.01, 1/\sqrt{(l/100)}))$
  end if
  if the acceptance rate of τ in iterations $l-99, l-98, \ldots, l > 0.44$ **then**
    $\phi^{(l)} \leftarrow \phi^{(l-1)} \times \exp(\min(0.01, 1/\sqrt{(l/100)}))$
  else if the acceptance rate of τ in iterations $l-99, l-98, \ldots, l < 0.23$ **then**
    $\phi^{(l)} \leftarrow \phi^{(l-1)} \times \exp(-\min(0.01, 1/\sqrt{(l/100)}))$
  end if
Otherwise, $\psi^{(l)} = \psi^{(l-1)}$ and $\phi^{(l)} = \phi^{(l-1)}$.

$\mathrm{N}[\Delta^{(l-1)}, \psi^2]$ used to update $\Delta^{(l)}$ in (3.1) and $\mathrm{N}[\log(\tau^{(l-1)}), \phi^2]$ used to update $\log(\tau^{(l)})$ in (3.5), where ψ and ϕ are the proposal scales. To avoid burdensome off-line tuning of the proposal scales, we implement an adaptive MCMC sampler [Brooks et al. (2011)] that allows automatic adjustment during the run. The steps of the adaptive MHwG+ASIS sampler are specified in Algorithm 1. Specifically, we implement an algorithm that updates the two proposal scales every 100 iterations, based on the most recent 100 proposals as outlined in Step 6 of Algorithm 1. The Markov chains equipped with the adaptive MCMC converge to the stationary distribution because the adjustment factors, $\exp(\pm\min(0.01, 1/\sqrt{i}))$, in Step 6 of of Algorithm 1 approach unity as i goes to infinity. This condition is called the diminishing adaptation condition [Roberts and Rosenthal (2007)]. We set the lower and upper bounds of the acceptance rate to 0.23 and 0.44, respectively [Gelman et al. (2014)].

All of the numerical results presented in Figure 5 were obtained using algorithms that similarly adapted their M–H updates, that is, the M–H updates of all the parameters in CMHwG and of Δ and τ in both MHwG and MHwG+ASIS.

4. Profile likelihood of the time delay. We use the profile likelihood of Δ [e.g., Davison (2003)] to obtain a simple approximation to its marginal posterior

distribution, $p(\Delta|\boldsymbol{x}, \boldsymbol{y})$. This profile likelihood is

$$L_{\text{prof}}(\Delta) \equiv \max_{\boldsymbol{\beta},\boldsymbol{\theta}} L(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta}) = L(\Delta, \hat{\boldsymbol{\beta}}_\Delta, \hat{\boldsymbol{\theta}}_\Delta), \tag{4.1}$$

where $L(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta})$ is the marginal likelihood function of the model parameters with the latent light curve integrated out, that is,

$$\begin{aligned} L(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta}) &= p(\boldsymbol{x}, \boldsymbol{y}|\Delta, \boldsymbol{\beta}, \boldsymbol{\theta}) \\ &= \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{X}(\boldsymbol{t}^\Delta), \Delta, \boldsymbol{\beta}) \times p(\boldsymbol{X}(\boldsymbol{t}^\Delta)|\Delta, \boldsymbol{\theta})\, d\boldsymbol{X}(\boldsymbol{t}^\Delta), \end{aligned} \tag{4.2}$$

and $(\hat{\boldsymbol{\beta}}_\Delta, \hat{\boldsymbol{\theta}}_\Delta)$ are the values of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ that maximize $L(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta})$ for each Δ.

The profile likelihood of a parameter, say φ, may asymptotically approximate its marginal posterior distribution with a uniform prior on φ. This happens, for example, if the log likelihood of the model parameters is approximately quadratic given φ under standard asymptotic arguments. The prior distribution on the parameters other than φ is chosen in such a way as to approximately cancel the determinant of Hessian matrix of the log likelihood, for example, as happens asymptotically with the Jeffreys' prior; see Appendix D for details.

Treating $L_{\text{prof}}(\Delta)$ as an approximation to $p(\Delta|\boldsymbol{x}, \boldsymbol{y})$, we evaluate $L_{\text{prof}}(\Delta)$ on a fine grid of values over the interesting range of Δ. We set w values from Δ_1 to Δ_w, that is, $\{\Delta_1, \Delta_2, \ldots, \Delta_w\}$, where, for example, $\Delta_j - \Delta_{j-1} = 0.1$ $(j = 2, 3, \ldots, w)$ for a high-resolution mapping. Unfortunately, this can be computationally burdensome due to the large number of values on the grid, for example, if the feasible range for Δ is $[-1500, 1500]$, the grid consists of 30,001 values. At one second per evaluation, this requires about 8 hours and 20 minutes. Though computationally expensive, the high-resolution mapping of $L_{\text{prof}}(\Delta)$ is useful because it clearly identifies the likely (modal) values of Δ. In practice, we use multiple cores in parallel to reduce the computation time and optimization is implemented using a general-purpose quasi-Newton method, `optim`, in R [R Core Team (2016)]. Initial values for numerical optimization at the first grid point are set just as with the Bayesian method described in Section 5 and the initial values for subsequent grid point are set to the values that maximize the profile likelihood at the previous grid point.

The profile likelihood evaluated on the grid can be used to approximate the posterior mean $E(\Delta|\boldsymbol{x}, \boldsymbol{y})$:

$$\hat{\Delta}_{\text{mean}} \equiv \frac{\sum_{j=1}^w \Delta_j \times L_{\text{prof}}(\Delta_j)}{\sum_{j=1}^w L_{\text{prof}}(\Delta_j)}, \tag{4.3}$$

and the posterior variance $\text{Var}(\Delta|\boldsymbol{x}, \boldsymbol{y})$,

$$\hat{V} \equiv \frac{\sum_{j=1}^w \Delta_j^2 \times L_{\text{prof}}(\Delta_j)}{\sum_{j=1}^w L_{\text{prof}}(\Delta_j)} - \left[\frac{\sum_{j=1}^w \Delta_j \times L_{\text{prof}}(\Delta_j)}{\sum_{j=1}^w L_{\text{prof}}(\Delta_j)}\right]^2. \tag{4.4}$$

Moreover, the posterior mode of Δ can be approximated by a value of Δ in the grid that maximizes the profile likelihood, which is a discrete approximation to the maximum likelihood estimator, $\hat{\Delta}_{\mathrm{MLE}} \equiv \arg\max_{\Delta} L_{\mathrm{prof}}(\Delta)$. If the profile likelihood exhibits multiple modes, however, the (approximate) posterior mean, mode and variance can be misleading. Instead, each mode requires separate investigation based on their (approximate) relative size.

5. Time delay estimation strategy and numerical illustrations. The first step of our analysis is to plot $L_{\mathrm{prof}}(\Delta)$ over the range of Δ to check for multimodality that may indicate multiple modes [e.g., Brooks et al. (1997)] in the marginal posterior distribution of Δ. For some quasars, the interesting range of Δ can be narrowed using the results of past analyses or information from other astrophysical probes as discussed in Section 2.4. If prior information for Δ is unavailable, we explore the feasible range.

In our numerical studies, we find that when $L_{\mathrm{prof}}(\Delta)$ is dominated by one mode, the moment estimates of Δ based on $L_{\mathrm{prof}}(\Delta)$, that is, $\hat{\Delta}_{\mathrm{mean}}$ in (4.3) and $\hat{V}$ in (4.4), are almost identical to the posterior mean and variance obtained via MCMC. On the other hand, modes near the margins of the range of Δ may indicate microlensing; see Section 5.1. In this case, the order of polynomial regression must be increased. If there are multiple modes that are not near the margins of the feasible range, each mode merits investigation; evaluating $L_{\mathrm{prof}}(\Delta)$ divided by the square root of the observed Fisher information at each mode provides an approximation of the relative size of each mode. If the modes are so close that the MCMC chain readily jumps between them, it is easy to estimate their relative size; see Section 5.3.

As a cross-check, in all of our numerical examples we run three MCMC chains near each of the major mode(s) identified by $L_{\mathrm{prof}}(\Delta)$; The three starting values for each mode are {mode, mode $\pm$20 days}. Each chain is run for 510,000 iterations and the first 10,000 iterations are discarded as burn-in; the Gelman–Rubin diagnostic statistics [Gelman and Rubin (1992)] of all of the model parameters computed from the post burn-in chains in all of our numerical examples are smaller than 1.001, which justifies our burn-in size. Because the smallest effective sample size of the parameters computed from the post burn-in chains across all of our examples is about 11,000, we thin each chain by a factor of fifty (from length 500,000 to 10,000). We combine the three thinned chains to obtain our Monte Carlo sample from the posterior distribution. For all chains, we set the starting value of $\boldsymbol{\beta}$ to the estimated regression coefficients obtained by regressing $\boldsymbol{y} - \sum_j x_j/n$ on a covariate matrix $\boldsymbol{W}_m(\boldsymbol{t} - \Delta^{(0)})$ whose jth row is $\boldsymbol{w}_m^{\top}(t_j - \Delta^{(0)})$, where $\Delta^{(0)}$ is the initial value of Δ. The initial value of $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ is the combined light curve, that is, $\{\boldsymbol{x}, \boldsymbol{y}_{-\boldsymbol{\Delta}^{(0)}} - \boldsymbol{W}_m^{\top}(\boldsymbol{t} - \Delta^{(0)})\boldsymbol{\beta}^{(0)}\}$ sorted in time. The starting value of μ is set to the mean of $\boldsymbol{x}$, that of σ^2 to 0.01^2, and that of τ to 200. We set the initial standard deviations of the proposal distributions to $\psi = 10$ days for Δ and $\phi = 3$ for $\log(\tau)$. (The unit of τ is days.)

We use simulated data of doubly- and quadruply-lensed quasars publicly available at the TDC website (http://timedelaychallenge.org) to illustrate our time delay estimation strategy when prior information for Δ is not available. We also analyze observed data of quasars *Q0957+561* and *J1029+2623* over the feasible range of Δ for illustrative purpose, though prior information is available to limit the range of Δ.

We report the CPU time in seconds using a server equipped with two 8-core Intel Xeon E5-2690 at 2.9 GHz and 64 GB of memory. We report the entire mapping time for $L_{\text{prof}}(\Delta)$.

5.1. *A doubly-lensed quasar simulation.* The simulated data for a doubly-lensed quasar are plotted in the first panel of Figure 6; the median cadence is 3 days, the cadence standard deviation is 1 day, observations are made for 4 months in each of 5 years for 200 observations in total and measurement errors are heteroskedastic Gaussian. The light curves suffer from microlensing which can be identified from their different long-term linear trends and similar short-term (intrinsic) variability.

To show the effect of microlensing on the time delay estimation, we fit both the curve-shifted model ($m = 0$) in (2.1) and the microlensing model with $m = 3$ in (2.2). We plot $\log(L_{\text{prof}}(\Delta))$ and $L_{\text{prof}}(\Delta)$ based on the curve-shifted model over the feasible range, $[t_1 - t_n, t_n - t_1] = [-1575.85, 1575.85]$, in the two panels of Figure 7. The profile likelihood exhibits large modes near the margins that overwhelm the profile likelihood near the true time delay (5.86 days denoted by the vertical dashed line).

In the presence of microlensing, the curve-shifted model cannot identify the time delay because the latent curves are not shifted versions of each other. The

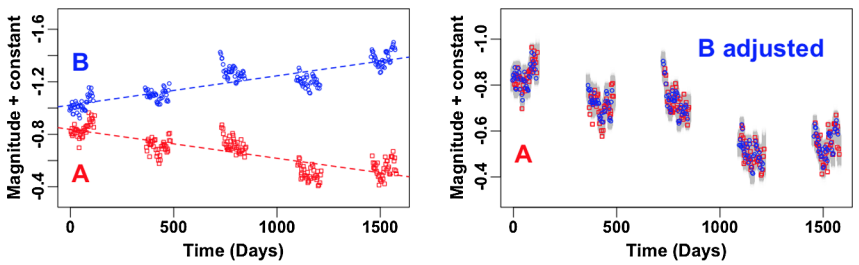

FIG. 6. *The first panel shows a TDC data set suffering from microlensing that results in light curves with different long-term trends. The dashed lines denote fitted linear regression lines. In the second panel, we combine the two light curves by shifting light curve B by* $E(\Delta|\boldsymbol{x}, \boldsymbol{y})$ *in the horizontal axis and by subtracting the estimated third-order polynomial regression based on* $E(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y})$ *from light curve B. The microlensing model finds matches between the intrinsic fluctuations of the light curves after removing the relative microlensing trend from light curve B. We plot the posterior sample of* $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ *in gray in the right panel to represent the point-wise prediction interval for the combined latent light curve. The gray areas encompass most of the combined observed light curve, indicating that the fitted model predicts the observed data well.*

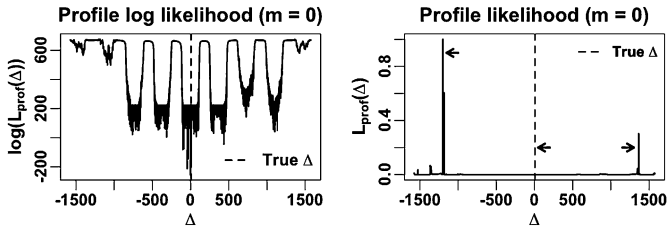

FIG. 7. *The profile log likelihood* (*left*) *and the profile likelihood* (*right*) *of* Δ *over its feasible range under the curve-shifted model* ($m = 0$). *We exponentiate and normalize* $\log(L_{\text{prof}}(\Delta_i))$ *as* $\exp[\log(L_{\text{prof}}(\Delta_i)) - \max_j(\log(L_{\text{prof}}(\Delta_j)))]$ *for all* i. *The vertical dashed line indicates the true time delay. The profile likelihood near the true time delay* (5.86 *days*) *is overwhelmed by the modes near margins.*

modes of $L_{\text{prof}}(\Delta)$ near the margins of the range of Δ occur because, in the small overlap between the tips of two light curves, spurious matches may be made by chance between similar fluctuation patterns. In Figure 8, for instance, we shift light curve B in the x-axis by the three values of Δ indicated by three arrows in the second panel of Figure 7. In the first panel of Figure 8, the two light curves shifted by the true time delay do not match for any shift in magnitude. However, given the time delays at around -1200 or 1360 days, the two light curves look well connected as shown in the second and third panels. Thus, the profile likelihood near the true time delay is overwhelmed by the values of the profile likelihood near -1200 and 1360 days.

To correct this effect, we fit the microlensing model with a third-order polynomial regression ($m = 3$). Both $\log(L_{\text{prof}}(\Delta))$ and $L_{\text{prof}}(\Delta)$ are plotted in Figure 9. One mode clearly dominates $L_{\text{prof}}(\Delta)$. Using a uniform prior for Δ over its feasible range and setting $\sigma^2 \sim \text{IG}(1, 2/10^7)$, we initialize three MCMC chains near

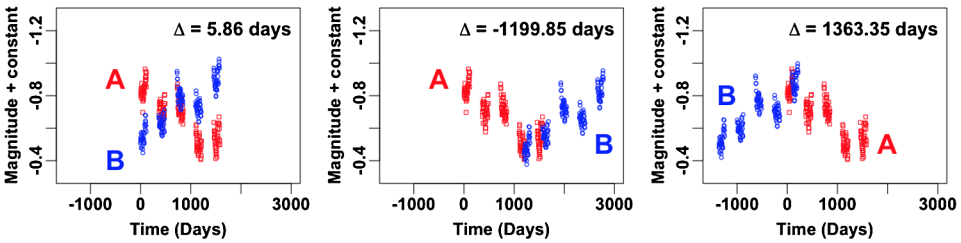

FIG. 8. *We shift light curve* B (*blue*) *by the true time delay* (5.86 *days*) *in the first panel, by* $-1{,}199.85$ *days in the second panel, and by* 1363.35 *days in the third panel. These three time delays correspond to three arrows in the second panel of Figure* 7. *The shift in magnitude used is the value of* β_0 *that maximizes the profile likelihood given each time delay. Without accounting for microlensing, the curve-shifted model fails because the light curves do not match even at the true time delay. The curve-shifted model may produce large modes near the margins because, in the small overlap between the tips of two light curves, spurious matches may be made by chance between similar fluctuation patterns as shown in the second and third panels.*

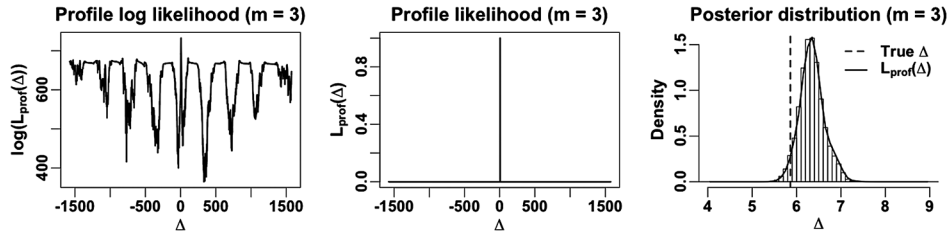

FIG. 9. *The profile log likelihood* (*first panel*) *and the profile likelihood* (*second panel*) *of* Δ *over its feasible range under the microlensing model* ($m = 3$). *The profile likelihood shows one mode near the true time delay* (5.86 *days*). *The third panel shows the marginal posterior distribution of* Δ *as a histogram of the MCMC samples with re-normalized* $L_{\text{prof}}(\Delta)$ *superimposed. The vertical dashed line indicates the true time delay.*

$\hat{\Delta}_{\text{mean}} = 6.36$ days. It took 14,457 seconds to map $L_{\text{prof}}(\Delta)$ and 5115 seconds on average for each MCMC chain. The profile likelihood and marginal posterior near the dominant mode are almost identical and are consistent with the true value of Δ as shown in the third panel of Figure 9.

In the second panel of Figure 6, we combine two light curves by shifting light curve B by the posterior mean of Δ in the horizontal axis and by subtracting the estimated polynomial regression based on the posterior means of $\boldsymbol{\beta}$ from light curve B. The microlensing model finds matches between the intrinsic fluctuations of the light curves after removing the relative microlensing trend from light curve B. We also plot the posterior sample of $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ in gray in the right panel of Figure 6. The gray regions represent the point-wise prediction intervals for the combined latent light curve. The gray areas encompass most of the combined observed light curve, indicating that the fitted model predicts the observed data well.

We summarize the Bayesian and profile likelihood estimates for Δ in Table 1. The true delay is within two posterior standard deviation of the posterior mean; similar accuracy is obtained with the profile likelihood approximation. This is anecdotal evidence that our model works well when microlensing is properly accounted for; there is no severe multi-modality near edges of the range of Δ in the second panel of Figure 9.

TABLE 1

Estimates of Δ; *the profile likelihood estimates,* $\hat{\Delta}_{\text{mean}}$ *and* $\hat{V}^{0.5}$ *are given in the* $E(\Delta|\boldsymbol{x}, \boldsymbol{y})$ *and* $SD \equiv SD(\Delta|\boldsymbol{x}, \boldsymbol{y})$ *columns, where* $\text{Error} \equiv |\Delta_{\text{true}} - E(\Delta|\boldsymbol{x}, \boldsymbol{y})|$ *with* Δ_{true} *indicating the true time delay* (5.86 *days*), *and* $\chi \equiv \text{Error}/SD(\Delta|\boldsymbol{x}, \boldsymbol{y})$

Method	$E(\Delta\|\boldsymbol{D}_{\text{obs}})$	$\hat{\Delta}_{\text{MLE}}$	**SD**	Δ_{true}	**Error**	χ
Bayesian	6.33		0.28	5.86	0.47	1.68
Profile likelihood	6.36	6.35	0.28	5.86	0.50	1.79

TABLE 2

Coverage estimates calculated from 1000 *simulated data sets*; *we generate these simulations using* (2.3), (2.4) *and* (2.8) *given the posterior median values of* $\{\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau\}$ *as generative values. After fitting our Bayesian model on each simulation, we check the proportion of interval estimates containing the generative values*

	Δ	β_0	β_1	β_2	β_3	μ	σ^2	τ
Coverage estimate	1.000	0.996	0.997	0.994	0.994	0.959	0.336	0.922

We also conduct a simulation study, generating 1000 datasets from our final model with an adjustment for microlensing ($m = 3$), and report the frequency coverage of the 95% posterior intervals [Tak, Kelly and Morris (2017)]. The result is over-coverage for Δ, conservatively meeting the spirit of the frequentist confidence level, reasonable coverage for $\boldsymbol{\beta}$, and under-coverage for both σ^2 and τ; see Table 2. The severe under-coverage for σ^2 does not seem to affect the coverage rate of Δ; the scatter plot of Δ and $\log(\sigma)$ in Figure 10 indicates that the two parameters are almost independent a posteriori. In a numerical sensitivity analysis in Appendix E, we show that the posterior mode of Δ is close to the true time delay even when the posterior mode of $\log(\sigma)$ is substantially different from its true value. (See Figure 17 in Appendix E.)

Figure 10 displays scatterplots of the posterior sample of Δ against each of the other model parameters. The time delay Δ exhibits weak correlations with the regression coefficients and nonlinear relationships with μ and $\log(\tau)$, though β_0 and $\log(\sigma)$ appear nearly independent of Δ.

5.2. *A quadruply-lensed quasar simulation*. The simulated data for a quadruply-lensed quasar are plotted in Figure 11 and are composed of four light curves, A, B, C and D; the median cadence is 6 days, the cadence standard deviation is 1 day, observations are made for 4 months in each of 10 years with 200

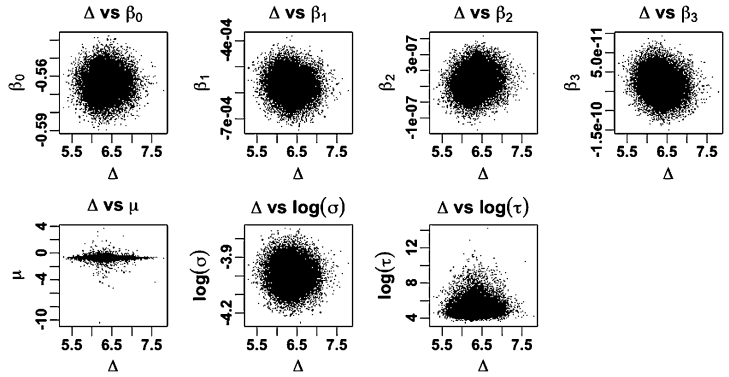

FIG. 10. *Scatter plots of the posterior sample of* Δ *and each of the other model parameters.*

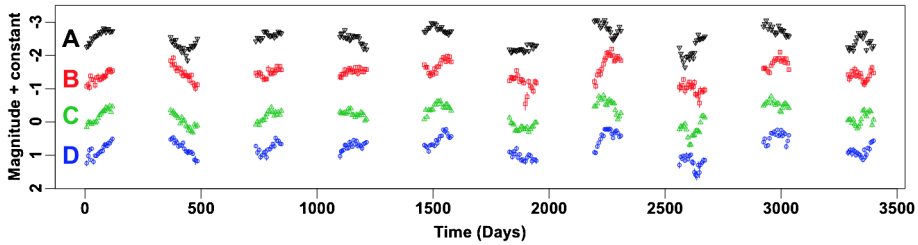

FIG. 11. *Simulated quadruply-lensed quasar data used in the TDC.*

observations in total, and measurement errors are heteroskedastic Gaussian. The feasible range for each Δ is $[-3391.62, 3391.62]$.

With quadruply lensed data, there are three time delay parameters because the four light curves are generated by one underlying process. Our model, however, is designed to analyze pairs of light curves independently, and thus we focus on Δ_{AB}, Δ_{AC} and Δ_{AD}, where the subscripts index the two light curves being compared, among the six possible pairs. This pair-wise approach proceeds by applying the method developed for doubly-lensed data in Section 5.1 to the pair of light curves corresponding to each of Δ_{AB}, Δ_{AC} and Δ_{AD} in turn [Fassnacht et al. (1999)].

By focusing on pairwise comparisons of the four time series, we do not account for the correlations between the time delays. A coherent model would consider all four light curves in a single model simultaneously [Hojjati, Kim and Linder (2013); Tewes, Courbin and Meylan (2013)]; the four light curves are generated from one latent process and the three distinct time delays may have a posteriori correlations. It is conceptually straightforward, but properly modeling quadruply lensed data would involve three time delays, 12 regression coefficients ($m = 3$) and three O–U parameters. Extending our model to simultaneously consider all of the data is a topic for future research.

We analyze the quadruply lensed simulated light curves, using the microlensing model ($m = 3$). After confirming a single dominating mode in the profile likelihood for each time delay parameter, we initiate three MCMC chains near this

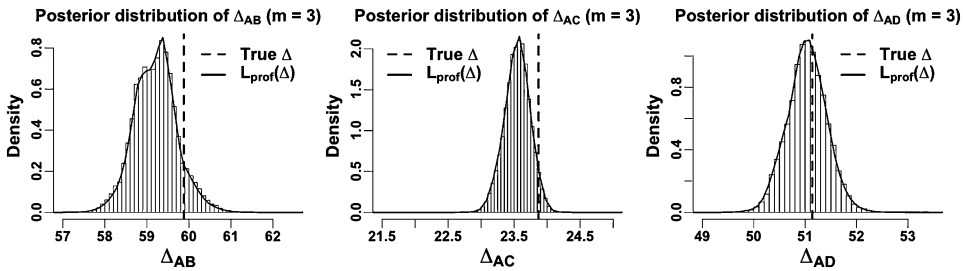

FIG. 12. *The marginal posterior distributions of* Δ_{AB} (*first panel*), Δ_{AC} (*second panel*), *and* Δ_{AD} (*third panel*) *with re-normalized* $L_{prof}(\Delta)$ *superimposed. Vertical dashed lines indicate blinded true time delays.*

TABLE 3

Estimates of Δ_{AB}, Δ_{AC}, *and* Δ_{AD}; *the profile likelihood estimates*, $\hat{\Delta}_{\text{mean}}$ *and* $\hat{V}^{0.5}$ *are given in the* $E(\Delta|\boldsymbol{x}, \boldsymbol{y})$ *and SD* $\equiv SD(\Delta|\boldsymbol{x}, \boldsymbol{y})$ *columns*, *where* $\text{Error} \equiv |\Delta_{\text{true}} - E(\Delta|\boldsymbol{x}, \boldsymbol{y})|$ *with* Δ_{true} *indicating the true time delay*, *that is*, $\Delta_{AB} = 59.88$, $\Delta_{AC} = 23.87$ *and* $\Delta_{AD} = 51.14$, *and* $\chi \equiv \text{Error}/SD(\Delta|\boldsymbol{x}, \boldsymbol{y})$

	Method	$E(\Delta\|\boldsymbol{x}, \boldsymbol{y})$	$\hat{\Delta}_{\text{MLE}}$	SD	Δ_{true}	Error	χ
Δ_{AB}	Bayesian	59.21		0.51	59.88	0.67	1.33
	Profile likelihood	59.21	59.38	0.51	59.88	0.67	1.33
Δ_{AC}	Bayesian	23.55		0.19	23.87	0.32	1.68
	Profile likelihood	23.54	23.58	0.19	23.87	0.33	1.74
Δ_{AD}	Bayesian	51.04		0.38	51.14	0.10	0.26
	Profile likelihood	51.03	51.08	0.38	51.14	0.11	0.29

mode. The posterior distributions of Δ_{AB}, Δ_{AC} and Δ_{AD} appear in Figure 12 with $L_{\text{prof}}(\Delta)$ superimposed. The profile likelihood is almost identical to the posterior distribution of each parameter and both estimate the true time delays well. The average CPU time taken to map $L_{\text{prof}}(\Delta)$ is about 73,000 seconds (averaging over the three time delays). The average CPU time taken for each MCMC chain is about 5500 seconds (averaging over nine chains; three chains for each time delay). Our estimation results are summarized in Table 3. The Bayesian estimates and profile likelihood approximations are quite similar and both produce estimates within two standard deviations of the truth.

5.3. *Quasar* Q0957+561. The first known gravitationally (doubly) lensed quasar *Q0957+561* was discovered by Walsh, Carswell and Weymann (1979) who suggested that a strong gravitational lensing may have formed the two images. Here, we analyze the most recent observations of this quasar. These observations were made by the United States Naval Observatory in 2008–2011 [Hainline et al. (2012). The data were observed on 57 nights and are plotted in the first panel of Figure 13. The feasible range for Δ is $[-1178.939, 1178.939]$.

Inspection of $L_{\text{prof}}(\Delta)$ reveals four modes close to each other near 425 days. Using a uniform prior distribution of Δ over its range and $\sigma^2 \sim \text{IG}(1, 2/10^7)$, we ran three MCMC chains near the highest mode. The second panel of Figure 13 shows the marginal posterior distribution of Δ with $L_{\text{prof}}(\Delta)$ superimposed. Here, the profile likelihood approximation to the marginal posterior distribution of Δ is less accurate. This may be because the approximation depends on an asymptotic argument while the data size is small. Nonetheless, the profile likelihood identifies each of the dominant modes of the posterior distribution. Mapping $L_{\text{prof}}(\Delta)$ took 1243 seconds and each MCMC chain took on average 1955 seconds.

In the third panel of Figure 13, we shift light curve B by the posterior mean of Δ in the horizontal axis and subtract the estimated third-order polynomial regression

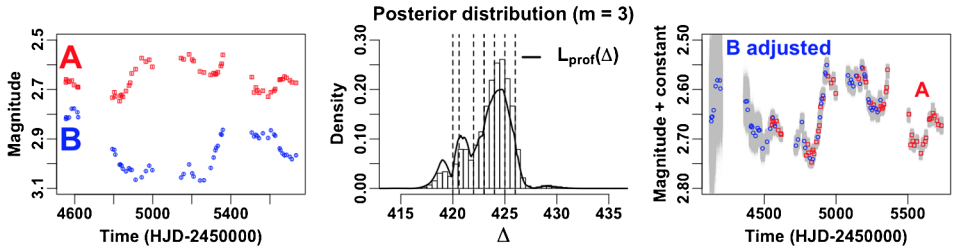

FIG. 13. *Observations of Quasar Q*0957+561 *from Hainline et al.* (2012) *are plotted in the first panel. The second panel exhibits the marginal posterior distribution of* Δ *with re-normalized* $L_{\text{prof}}(\Delta)$ *superimposed. The vertical dashed lines represent the historical estimates given in Table 4 which are concentrated near the highest and the second highest modes. In the third panel, we combine the two light curves by shifting light curve* B *by* $E(\Delta|\boldsymbol{x}, \boldsymbol{y})$ *in the horizontal axis and by subtracting the estimated third-order polynomial regression based on* $E(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y})$. *The gray regions represent the point-wise prediction intervals for the combined latent light curve, that is, the posterior sample of* $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$. *The gray areas encompass most of the observed light curve, which shows how well the fitted model predicts the observed data. HJD indicates the Heliocentric Julian date.*

based on the posterior means of $\boldsymbol{\beta}$. The fitted microlensing model matches the intrinsic fluctuations of the two light curves well. We also plot the posterior sample of $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ in gray which represents the point-wise prediction intervals for the latent light curves. The gray areas encompass most of the observed light curve, which shows how well the fitted model predicts the observed data.

Estimates based on different observations of Q0957+561 appear in Table 4. Though the posterior mean and standard deviation may be difficult to interpret with a multimodal posterior distribution, we include them for comparison. Our estimates are broadly consistent with the others. We emphasize here that our methods reveal several possible time delays in that there are four modes in the marginal posterior distribution of Δ, whereas previous analyses report only a single point estimate for Δ and its standard error. By investigating the entire posterior distribution, we learn that previous estimates, denoted by vertical dashed lines in the second panel of Figure 13, are located near the highest and the second highest modes of our marginal posterior distribution of Δ. Thus, our approach is more informative in that it provides a summary of several possible values of Δ and their relative likelihoods, corresponding to locations and sizes of the several modes of the posterior distribution of Δ.

5.4. *Quasar J*1029+2623. Inada et al. (2006) discovered the gravitationally lensed quasar *J1029+2623* whose estimated time delay is the second largest yet observed. Though *J1029+2623* has three images (*A*, *B* and *C*), Fohlmeister et al. (2013) merged *B* and *C* because they overlap significantly and thus their light curves are difficult to disentangle. They published its data (A, $B + C$) with 279 epochs monitored at the Fred Lawrence Whipple Observatory from January 2007

TABLE 4

Historical time delay estimates ($\hat{\Delta}$) and standard errors (SE) for Q0957+561 (r-band). We compute the posterior mean and standard deviation of Δ (423.71 ± 2.03); the profile likelihood approximate the posterior mean and standard deviation as 423.21 ± 2.81. Pelt et al. (1996), Oscoz et al. (1997), Oscoz et al. (2001), Serra-Ricart et al. (1999) and Shalyapin et al. (2014) adopted various methods to estimate Δ using different data sets spanning different periods. We report the average measurement standard deviation (SD) of their data; two average measurement SDs are reported if their data come from two sources. In all cases except our method, a bootstrapping method was used to calculate the SE

Researchers	Number of observations	Observation period	Measurement SD (mag.)	$\hat{\Delta}$	SE
Pelt et al. (1996)	831	1979–1994	0.0159	423	6
Oscoz et al. (1997)	86	1994–1996	0.01, 0.02	424	3
Serra-Ricart et al. (1999)	197	1996–1998	0.023, 0.025	425	4
Oscoz et al. (2001)	100	1994–1996	0.009, 0.01	426	5
				423	2
				420	8
				422	3
Shalyapin et al. (2014)	371	2005–2010	0.012	420.6	1.9
This work	57	2008–2011	0.004	423.71	2.03
				423.21	2.81

to June 2012. The time delay estimate obtained by analyzing the combined image can be different from that obtained by analyzing images B or C separately. Nonetheless, we follow Fohlmeister et al. (2013) in order to provide a fair comparison. The first panel in Figure 14 shows these data. The feasible range of Δ is $[-2729.759, 2729.759]$.

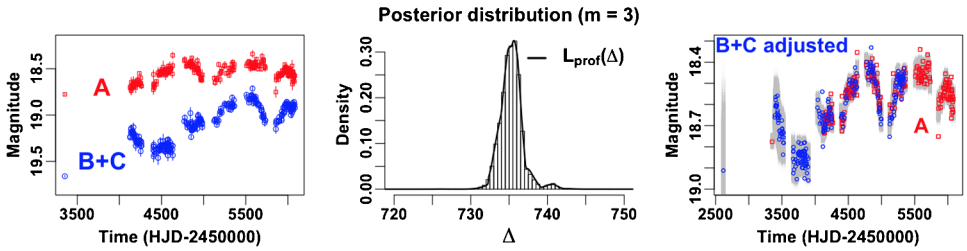

FIG. 14. *We plot the observations of Quasar J1029+2623 from Fohlmeister et al. (2013) in the first panel. The second panel exhibits the marginal posterior distribution of Δ with re-normalized $L_{\text{prof}}(\Delta)$ superimposed. In the last panel, we combine the two light curves by shifting light curve B + C by $E(\Delta|\boldsymbol{x}, \boldsymbol{y})$ in the horizontal axis and by subtracting the estimated third-order polynomial regression based on $E(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y})$. The gray areas represent the range of the point-wise prediction intervals for the latent light curve. Comparing the gray regions with the observed data shows how well the model fits the observed data. HJD indicates the Heliocentric Julian date.*

We confirm a dominant mode near 735 days and invisibly small modes near -2000 and 1800 days via $L_{\text{prof}}(\Delta)$. Since the mode near 735 days overwhelms the other modes, we focus on the dominant mode. We initiated three MCMC chains near 735 days using a uniform prior distribution of Δ over its range and $\sigma^2 \sim \text{IG}(1, 2 \times 10^{-7})$. We display the marginal posterior distribution of Δ in the second panel of Figure 14 with $L_{\text{prof}}(\Delta)$ superimposed. Mapping $L_{\text{prof}}(\Delta)$ took 33,683 seconds and each MCMC chain took an average of 8555 seconds. The posterior distribution and the profile likelihood are almost identical. In the third panel, we shift light curve B by the posterior mean of Δ in the horizontal axis and subtract the estimated third-order polynomial regression based on the posterior mean of $\boldsymbol{\beta}$. Again, the fitted microlensing model is a good match of the two light curves and our graphical model checking shows that the range of our predicted values for the combined latent light curve, denoted by the gray areas, encompasses the observed light curve well.

In Table 5, we compare our estimates with historical estimates that are based on the same data. The Bayesian method uses 5% and 95% quantiles of the posterior samples of Δ as the 90% posterior interval. To obtain the 90% interval estimate for Δ via the profile likelihood, we draw a sample of size 50,000 of Δ using the empirical CDF of the normalized profile likelihood and report the 5% and 95% quantiles.

The shape of the posterior distribution of Δ is almost identical to that of the profile likelihood in the second panel of Figure 14. However, the posterior mean of Δ is larger than the profile approximation, $\hat{\Delta}_{\text{mean}}$, by about two days. This is because of invisibly small modes near -2000 and 1800 days.

Overall, our point estimates are smaller than the historical estimates by about ten days and our 90% posterior intervals are much shorter than the historical 90% confidence intervals in Table 5. We suspect that the discrepancy between the estimates might arise from the overly simple microlensing models used in the historical analyses. Fohlmeister et al. (2013), for example, combine the output obtained from fitting two different models, one with a linear model for the microlensing

TABLE 5

Historical time delay estimates and 90% *confidence intervals for J*1029+2623. *Our work provides the posterior mean and* 90% *posterior interval of* Δ *and profile likelihood approximations to them. Fohlmeister et al.* (2013) *did not specify how they produced the sampling distribution of* Δ. *Kumar, Stalin and Prabhu* (2014) *used a parametric bootstrapping method*

Researchers	**Method**	**Estimate**	**90% Interval**
Fohlmeister et al. (2013)	χ^2-minimization (AIC, BIC)	744	(734, 754)
Kumar, Stalin and Prabhu (2014)	Difference-smoothing	743.5	(734.6, 752.4)
This work	Bayesian	735.30	(733.09, 737.62)
	Profile likelihood	733.11	(732.94, 738.44)

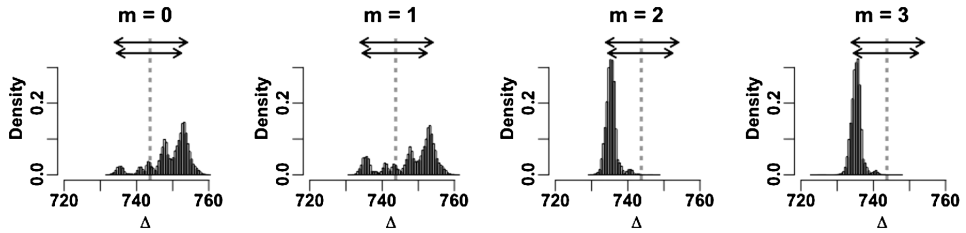

FIG. 15. *The marginal posterior distributions of* Δ *with* $m = 0, 1, 2$ *and* 3. *The* 90% *intervals of Fohlmeister et al.* (2013) *and Kumar, Stalin and Prabhu* (2014) *are denoted by the lengths of the arrows at the top of each panel and cover all the modes of the posterior distributions of* Δ *shown in the first* ($m = 0$) *and second* ($m = 1$) *panels. This implies that their microlensing models might have produced results similar to our microlensing model with either* $m = 0$ *or* $m = 1$. *Their point estimates are denoted by vertical dashed lines. As we increase the order of the polynomial regression for microlensing, the severe multi-modality dissipates and one mode becomes dominant. This implies why our point and interval estimates are quite different from theirs.*

polynomial (which was optimal with respect to AIC) and the other with no adjustment for microlensing (which was optimal with respect to BIC). Unfortunately, they do not describe *how* they combine the fits. Since they also apply high-order splines for each season in addition to the microlensing polynomial, neither of their models are directly comparable to ours. Kumar, Stalin and Prabhu (2014), on the other hand, account for the microlensing by using a Gaussian kernel smoothing technique.

Figure 15 shows our marginal posterior distribution of Δ with $m = 0$, 1, 2 and 3, respectively. The 90% intervals of Fohlmeister et al. (2013) and Kumar, Stalin and Prabhu (2014) denoted by the lengths of the arrows at the top of each panel cover all the modes of the posterior distributions of Δ shown in the first ($m = 0$) and second ($m = 1$) panels. This implies that their microlensing models might have produced results similar to our microlensing model with either $m = 0$ or $m = 1$. As we increase the order of the polynomial regression for microlensing, the severe multi-modality dissipates and one mode becomes dominant. Thus, the discrepancy between their estimates and ours with $m = 3$ might be due to their use of an overly simple microlensing model.

6. Concluding remarks. Accurately estimating time delays among gravitationally lensed quasar images is a key to making fundamental measurements of the current expansion rate of the universe and dark energy [Refsdal (1964); Linder (2011)]. The Large Synoptic Survey Telescope [LSST Science Collaboration (2009)] will produce extensive time series data on thousands of multiply lensed quasars starting in 2022. Anticipating this era of the LSST, we have improved the fully Bayesian model of Harva and Raychaudhury (2006) by leveraging recent advances in astrophysical and statistical modeling. We have added an Ornstein–Uhlenbeck process to model the fluctuations in quasar light curves, a

polynomial regression to account for microlensing, and a profile-likelihood-guided Bayesian strategy.

We proposed our original model in the context of the Time Delay Challenge [TDC, Dobler et al. (2015); Liao et al. (2015)]. This original model worked well for data without severe microlensing, leading to the best precision and targeted average bias level among the methods submitted to the TDC. Our original model, however, did not properly account for microlensing, and thus produced poor time delay estimates in some cases resulting in a mediocre performance in one of the evaluation criteria. This motivated us to develop our current microlensing model.

The upcoming second TDC, called the TDC2, aims to further improve estimation methods under a more realistic setting. Since the LSST will produce multiband optical data observed in filters centered at six different wavelengths for each lens system, proper analysis will require jointly modeling a vector of light curves to estimate the common time delay in each system. Modeling microlensing will be more challenging in TDC2, because its effect depends on the wavelength of the quasar light.

There are several opportunities to build upon our work in preparation for the TDC2 and eventually for the LSST. It is desirable to implement more sophisticated methods of model selection such as information criteria to choose the complexity of the microlensing trend. Though astrophysicists have used a cubic polynomial trend for microlensing models for some quasars so far, it would be better to have a fast and principled mechanism to determine the order given any data of gravitationally lensed quasars. Another avenue for further improvement is to constrain the range of the time delay by incorporating additional astrophysical information such as spatial positions of the images relative to the lensing galaxy, and an astrophysical model for the mass distribution of the lens. For quadruply-lensed quasar systems, constructing a Bayesian model to simultaneously analyze the four light curves, would allow us to coherently estimate the relative time delays without loss of information. Further improvements to the computational efficiency of our profile likelihood and MCMC strategies for analyzing extensive vector time series will enhance their effectiveness in the era of the LSST.

APPENDIX A: THE LIKELIHOOD FUNCTION

We define a combined light curve $\boldsymbol{z} = (z_1, z_2, \ldots, z_{2n})^\top$ as follows. The observed magnitude at time t_i^Δ $(i = 1, 2, \ldots, 2n)$ is denoted by z_i, which is either x_j or $y_j - \boldsymbol{w}_m^\top(t_j - \Delta)\boldsymbol{\beta}$ for some j $(j = 1, 2, \ldots, n)$ depending on whether t_i^Δ is one of the elements of $\boldsymbol{t}$ or of $\boldsymbol{t} - \Delta$. The measurement standard deviation is denoted by ξ_i, which is either δ_j for x_j or η_j for $y_j - \boldsymbol{w}_m^\top(t_j - \Delta)\boldsymbol{\beta}$ for some j. We also define z_i' as the centered observed magnitude at time t_i^Δ, which is either $x_j - \mu$ or $y_j - \boldsymbol{w}_m^\top(t_j - \Delta)\boldsymbol{\beta} - \mu$ for some j. Let $D_i = \{z_1', z_2', \ldots, z_i'\}$ and derive the likelihood

$$L(\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau) \propto p(z_1') \times \prod_{i=2}^{2n} p(z_i'|D_{i-1}) \tag{A.1}$$

with $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ integrated out. The sampling distribution of D_{2n} given $\Delta, \boldsymbol{\beta}, \mu, \sigma^2$ and τ factors as

$$z_1' \sim \mathrm{N}\big[0, \xi_1^2 + \tau\sigma^2/2\big], \tag{A.2}$$

$$z_i' | D_{i-1} \sim \mathrm{N}\big[a_i\mu_{i-1}, \xi_i^2 + a_i^2\Omega_{i-1} + \tau\sigma^2(1-a_i^2)/2\big], \tag{A.3}$$

where $\mu_1 = (1-B_1)z_1'$, $\mu_i = (1-B_i)z_i' + B_i a_i \mu_{i-1}$, $\Omega_i = (1-B_i)\xi_i^2$, $B_1 = \xi_1^2/[\xi_1^2 + \tau\sigma^2/2]$, $B_i = \xi_i^2/[\xi_i^2 + a_i^2\Omega_{i-1} + \tau\sigma^2(1-a_i^2)/2]$. Thus, the likelihood function of $(\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau)$ in (A.1) is the product of the Gaussian densities.

By multiplying (A.1) by the prior density functions for Δ and $\boldsymbol{\beta}$ in (2.10) and those for μ, σ^2, and τ in (2.11), we can obtain their joint posterior density function with the latent magnitudes $\boldsymbol{X}(\boldsymbol{t}^{\Delta})$ marginalized out. Given the values of $(\boldsymbol{\beta}, \mu, \sigma^2, \tau)$ and a Uniform$[u_1, u_2]$ prior distribution for Δ, $L(\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau)$ is proportional to the marginalized conditional posterior density $p(\Delta | \boldsymbol{\beta}, \mu, \sigma^2, \tau, \boldsymbol{x}, \boldsymbol{y})$ used in (3.1) and (3.6).

APPENDIX B: CONDITIONAL POSTERIOR DISTRIBUTIONS OF THE LATENT MAGNITUDES

We use the same notation for the observed data as is defined in Appendix A, that is, z_i' and ξ_i. We introduce the centered latent magnitudes $\boldsymbol{X}'(\boldsymbol{t}^{\Delta}) = \boldsymbol{X}(\boldsymbol{t}^{\Delta}) - \mu$ for notational simplicity. Also, let "$< t_i^{\Delta}$" denote the set $\{t_j^{\Delta} : j = 1, 2, \ldots, i-1\}$ and "$> t_i^{\Delta}$" denote $\{t_j^{\Delta} : j = i+1, i+2, \ldots, 2n\}$. To sample $p(\boldsymbol{X}'(\boldsymbol{t}^{\Delta}) | \Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau, \boldsymbol{x}, \boldsymbol{y})$ used in (3.1), we sample the following conditional posterior distributions of each latent magnitude. (We suppress conditioning on $\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau, \boldsymbol{x}, \boldsymbol{y}$.)

$$X'(t_1^{\Delta}) | \boldsymbol{X}'(> t_1^{\Delta}) \sim \mathrm{N}\big[(1-B_1)z_1' + B_1 a_2 X'(t_2^{\Delta}), (1-B_1)\xi_1^2\big], \tag{B.1}$$

where $B_1 = \xi_1^2/[\xi_1^2 + \tau\sigma^2(1-a_2^2)/2]$. For $i = 2, 3, \ldots, 2n-1$,

$$\begin{aligned} &X'(t_i^{\Delta}) | \boldsymbol{X}'(< t_i^{\Delta}), \boldsymbol{X}'(> t_i^{\Delta}) \\ &\quad \sim \mathrm{N}\left[(1-B_i)z_i' + B_i\left((1-B_i^*)\frac{X'(t_{i+1}^{\Delta})}{a_{i+1}} + B_i^* a_i X'(t_{i-1}^{\Delta})\right), (1-B_i)\xi_i^2\right], \end{aligned} \tag{B.2}$$

where $B_i = \xi_i^2/[\xi_i^2 + \frac{\tau\sigma^2}{2}\frac{(1-a_i^2)(1-a_{i+1}^2)}{1-a_i^2 a_{i+1}^2}]$ and $B_i^* = \frac{1-a_{i+1}^2}{1-a_i^2 a_{i+1}^2}$. Last,

$$X'(t_{2n}^{\Delta}) | \boldsymbol{X}'(< t_{2n}^{\Delta}) \sim \mathrm{N}\big[(1-B_{2n})z_{2n}' + B_{2n} a_{2n} X'(t_{2n-1}^{\Delta}), (1-B_{2n})\xi_{2n}^2\big], \tag{B.3}$$

where $B_{2n} = \xi_{2n}^2/[\xi_{2n}^2 + \tau\sigma^2(1-a_{2n}^2)/2]$ and $a_i = \exp(-(t_i^{\Delta} - t_{i-1}^{\Delta})/\tau)$. Having sampled $\boldsymbol{X}'(\boldsymbol{t}^{\Delta})$, we set $\boldsymbol{X}(\boldsymbol{t}^{\Delta}) = \boldsymbol{X}'(\boldsymbol{t}^{\Delta}) + \mu$.

APPENDIX C: CONDITIONAL POSTERIOR DISTRIBUTIONS OF $\boldsymbol{\beta}$, μ, σ^2, AND τ USED IN ALGORITHM 1

We specify the conditional posterior distributions of $\boldsymbol{\beta}$, μ, σ^2, and τ used in Steps 2–5 of the MHwG+ASIS sampler in Algorithm 1. We suppress explicitly conditioning on $\boldsymbol{x}$ and $\boldsymbol{y}$ in the condition.

Step 2 of Algorithm 1 first samples $\boldsymbol{\beta}$ from the following Gaussian conditional posterior distribution [Step 2_a in (3.12)]; with an n by n diagonal matrix V whose diagonal elements are $\boldsymbol{\eta}^2$,

$$\boldsymbol{\beta}|\mu,\sigma,\tau,\boldsymbol{X}\big(\boldsymbol{t}^{\Delta}\big),\Delta\sim \mathrm{N}_{m+1}\big[J^{-1}\boldsymbol{W}_m(\boldsymbol{t}-\Delta)^{\top}V^{-1}\boldsymbol{u},\,J^{-1}\big], \tag{C.1}$$

where $J\equiv \boldsymbol{W}_m^{\top}(\boldsymbol{t}-\Delta)V^{-1}\boldsymbol{W}_m(\boldsymbol{t}-\Delta)+10^{-5}I_{m+1}$ and $\boldsymbol{u}\equiv \boldsymbol{y}-\boldsymbol{X}(\boldsymbol{t}-\Delta)$. To implement ASIS, we need the conditional posterior distribution for $\boldsymbol{\beta}$ given $\boldsymbol{K}(\boldsymbol{t}^{\Delta})$ used in (3.14). Let $\boldsymbol{K}'(\boldsymbol{t}^{\Delta})\equiv \boldsymbol{K}(\boldsymbol{t}^{\Delta})-\mu$, B be a $2n$ by $(m+1)$ matrix whose jth row is $(\boldsymbol{w}_m(t_j^{\Delta})-a_j\times \boldsymbol{w}_m(t_{j-1}^{\Delta}))^{\top}$ with $a_1=0$, L be a $2n$ by $2n$ diagonal matrix whose jth diagonal element is $\tau\sigma^2(1-a_j^2)/2$, $\boldsymbol{b}$ be a $2n$ by 1 vector whose jth element is $K'(t_j^{\Delta})-a_jK'(t_{j-1}^{\Delta})$, and finally $A\equiv B^{\top}L^{-1}B+10^{-5}I_{m+1}$. Then

$$\boldsymbol{\beta}|\mu,\sigma^2,\tau,\boldsymbol{K}\big(\boldsymbol{t}^{\Delta}\big),\Delta,\boldsymbol{x},\boldsymbol{y}\sim \mathrm{N}_{m+1}\big[A^{-1}B^{\top}L^{-1}\boldsymbol{b},\,A^{-1}\big]. \tag{C.2}$$

In Step 3 of Algorithm 1, we sample μ from a truncated Gaussian conditional posterior distribution whose support is $[-30,30]$;

$$\begin{aligned}&\mu|\sigma^2,\tau,\boldsymbol{X}\big(\boldsymbol{t}^{\Delta}\big),\Delta,\boldsymbol{\beta}\\&\quad\sim \mathrm{N}\Bigg[\frac{X(t_1^{\Delta})+\sum_{i=2}^{2n}\frac{X(t_i^{\Delta})-a_iX(t_{i-1}^{\Delta})}{1+a_i}}{1+\sum_{i=2}^{2n}\frac{1-a_i}{1+a_i}},\,\frac{\tau\sigma^2/2}{1+\sum_{i=2}^{n}\frac{1-a_i}{1+a_i}}\Bigg].\end{aligned} \tag{C.3}$$

In Step 4 of Algorithm 1, the parameter σ^2 has an inverse-Gamma conditional posterior distribution, that is,

$$\begin{aligned}&\sigma^2|\mu,\tau,\boldsymbol{X}\big(\boldsymbol{t}^{\Delta}\big),\Delta,\boldsymbol{\beta}\\&\quad\sim \mathrm{IG}\Bigg(n+1,\,b_{\sigma}+\frac{(X(t_1^{\Delta})-\mu)^2}{\tau}\\&\qquad+\sum_{i=2}^{2n}\frac{[(X(t_i^{\Delta})-\mu)-a_i(X(t_{i-1}^{\Delta})-\mu)]^2}{\tau(1-a_i^2)}\Bigg).\end{aligned} \tag{C.4}$$

The conditional posterior density function of τ used in Step 5 of Algorithm 1 is known up to a normalizing constant, that is,

$$\begin{aligned}&p\big(\tau|\mu,\sigma^2,\boldsymbol{X}\big(\boldsymbol{t}^{\Delta}\big),\Delta,\boldsymbol{\beta}\big)\\&\quad\propto \frac{\exp\big(-\frac{1}{\tau}-\frac{(X(t_1^{\Delta})-\mu)^2}{\tau\sigma^2}-\sum_{i=2}^{2n}\frac{[(X(t_i^{\Delta})-\mu)-a_i(X(t_{i-1}^{\Delta})-\mu)]^2}{\tau\sigma^2(1-a_i^2)}\big)}{\tau^{n+2}\prod_{i=2}^{2n}(1-a_i^2)^{1/2}}\\&\qquad\times I_{\{\tau>0\}}.\end{aligned} \tag{C.5}$$

To sample τ from (C.5), we use an M–H step with a Gaussian proposal density $\mathrm{N}[\log(\tau), \phi^2]$ on a logarithmic scale where ϕ is a proposal scale tuned to produce a reasonable acceptance rate.

APPENDIX D: PROFILE LIKELIHOOD APPROXIMATION TO THE MARGINAL POSTERIOR DISTRIBUTION OF Δ

We show that $L_{\mathrm{prof}}(\Delta)$ with a uniform prior distribution on Δ is approximately proportional to $p(\Delta|\boldsymbol{x}, \boldsymbol{y})$. Let $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$. Then

$$p(\Delta|\boldsymbol{x}, \boldsymbol{y}) = \int L(\Delta, \boldsymbol{\nu}) p(\Delta, \boldsymbol{\nu})\, d\boldsymbol{\nu} = k \int L(\Delta, \boldsymbol{\nu}) p(\boldsymbol{\nu}|\Delta)\, d\boldsymbol{\nu}, \tag{D.1}$$

where k is a normalizing constant of the uniform prior distribution for Δ and the likelihood function is the marginal likelihood function defined in (4.2) or (A.1). We specify a Jeffreys' prior on $\boldsymbol{\nu}$ given Δ, that is, $p(\boldsymbol{\nu}|\Delta) \propto |I_\Delta(\boldsymbol{\nu})|^{0.5}\, d\boldsymbol{\nu}$, where $I_\Delta(\boldsymbol{\nu})$ is the Fisher information defined as $-E[\partial^2 \log(L(\Delta, \boldsymbol{\nu}))/\partial \boldsymbol{\nu} \boldsymbol{\nu}^\top]$. The resulting $p(\Delta|\boldsymbol{x}, \boldsymbol{y})$ is a Jeffreys-integrated marginal likelihood under a uniform prior [Berger, Liseo and Wolpert (1999)]. (With a uniform prior on $\boldsymbol{\nu}$ given Δ, that is, $p(\boldsymbol{\nu}|\Delta) \propto 1$, the likelihood is a uniform-integrated marginal likelihood which can be approximated by the Laplace method.[16]) If we can approximate $l(\Delta, \boldsymbol{\nu}) \equiv \log(L(\Delta, \boldsymbol{\nu}))$ with respect to $\boldsymbol{\nu}$ by a second-order Taylor's series, for example, under standard asymptotic arguments, then

$$l(\Delta, \boldsymbol{\nu}) \approx l(\Delta, \hat{\boldsymbol{\nu}}_\Delta) - (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)^\top \big[-l''(\Delta, \hat{\boldsymbol{\nu}}_\Delta)\big](\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)/2, \tag{D.2}$$

where $\hat{\boldsymbol{\nu}}_\Delta = \arg\max_{\boldsymbol{\nu}} l(\Delta, \boldsymbol{\nu})$, and $l''(\Delta, \hat{\boldsymbol{\nu}}_\Delta) \equiv \partial^2 l(\Delta, \boldsymbol{\nu})/\partial \boldsymbol{\nu} \boldsymbol{\nu}^\top|_{\boldsymbol{\nu} = \hat{\boldsymbol{\nu}}_\Delta}$, which results in

$$L(\Delta, \boldsymbol{\nu}) \approx \exp\big(l(\Delta, \hat{\boldsymbol{\nu}}_\Delta) - (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)^\top \big[-l''(\Delta, \hat{\boldsymbol{\nu}}_\Delta)\big](\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)/2\big). \tag{D.3}$$

Using this, we approximate the marginal posterior density function of Δ by

$$\begin{aligned} p(\Delta|\boldsymbol{x}, \boldsymbol{y}) &\approx k \times L(\Delta, \hat{\boldsymbol{\nu}}_\Delta) \\ &\quad \times \int \exp\big(-(\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)^\top \big[-l''(\Delta, \hat{\boldsymbol{\nu}}_\Delta)\big](\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)/2\big) \big| I_\Delta(\boldsymbol{\nu}) \big|^{0.5}\, d\boldsymbol{\nu}. \end{aligned} \tag{D.4}$$

If we replace the Fisher information in (D.4), that is, $I_\Delta(\boldsymbol{\nu})$, with the observed information, $-l''_\Delta(\hat{\boldsymbol{\nu}}_\Delta)$, under standard asymptotic arguments, the integral in (D.4)

[16] The Laplace approximation based on the uniform prior requires the Hessian but most optimizers numerically evaluate the Hessian (or its approximation) automatically. However, the closed-form of the Hessian matrix is not available and a numerical approximation to the Hessian matrix is unstable in the small data example in Section 5.3. The profile likelihood approximation based on the Jeffreys' prior does not require calculating the Hessian, which is a computational advantage especially since we must evaluate the profile likelihood at every point on a grid of values of Δ. We note, however, that the Jeffreys' prior can be inappropriate for a reference prior because it sometimes becomes too informative in high dimensions [e.g., Berger, Bernardo and Sun (2015)].

converges to $(2\pi)^2$ because the integrand converges to a multivariate Gaussian density up to $(2\pi)^{-2}$. Finally,

$$\begin{aligned} p(\Delta|\boldsymbol{x}, \boldsymbol{y}) &\approx k \times (2\pi)^2 \times L(\Delta, \hat{\boldsymbol{v}}_\Delta) \\ &= k \times (2\pi)^2 \times L_{\text{prof}}(\Delta) \propto L_{\text{prof}}(\Delta). \end{aligned} \tag{D.5}$$

APPENDIX E: SENSITIVITY ANALYSES

To assess the influence of prior distributions of τ and σ^2 on the posterior distribution of Δ, we conduct sensitivity analyses, varying the scale and shape parameters of their IG prior distributions.

As an example, we generate 80 observations with $(\Delta, \beta_0, \mu, \sigma^2, \tau) = (50, 2, 0, 0.03^2, 100)$. The median observation cadence is 3 days and the measurement standard deviations are set to 0.005 magnitude. When fitting the Bayesian model, we assume for simplicity that $\Delta \sim \text{Uniform}[0, 100]$ a priori. We run three Markov chains, each for 10,000 iterations after 10,000 burn-in iterations.

E.1. Sensitivity analysis of the prior distribution of τ. We investigate the sensitivity of the posterior distribution of Δ to the shape parameter of the IG prior distribution of τ. We denote the shape parameter by a_τ and fix the scale parameter at one day. A reasonably small value of the scale parameter does not make any differences in the resultant posterior distributions of τ or Δ because $a_j = \exp(-(t_j^\Delta - t_{j-1}^\Delta)/\tau)$ dominates the scale parameter in the conditional posterior density of τ in (C.5). We fix the $\text{IG}(1, b_\sigma)$ prior distribution for σ^2, where $b_\sigma = 2 \times 10^{-7}$ mag^2/day as described in Section 2.5.

Figure 16 shows the result of sensitivity analysis with three values of the shape parameter, $a_\tau = 0.1$, 10 and 80 (columns). Each column shows the posterior distribution of Δ (first row), that of $\log(\tau)$ (second row), and a scatter plot of posterior samples of $\log(\sigma)$ and $\log(\tau)$ (third row) obtained under each shape parameter. The dashed lines indicate the generative true values.

The modes of the first two posterior distributions of Δ are near the generative value of Δ. However, with the informative choice of $a_\tau = 80$, the posterior distribution of Δ is flat. A large value of a_τ concentrates the prior density on the O–U processes on mean-reversion timescales τ much shorter than the observational cadence. Moreover, a large value results in a prior mode, $1/(1 + a_\tau)$, that is close to zero and a large value of the degrees of freedom $(2 \times a_\tau)$ for the prior distribution strongly influences the posterior of τ. Hence, the latent light curves governed by these O–U processes with small τ will effectively appear as white noise time series. The result is a model that is ineffective at constraining the time delay because it is unable to match serially correlated fluctuation patterns in the light curves. The second row in Figure 16 shows that as a_τ increases, the mode of the posterior distribution of $\log(\tau)$ becomes smaller with a shorter right tail, and thus moves away

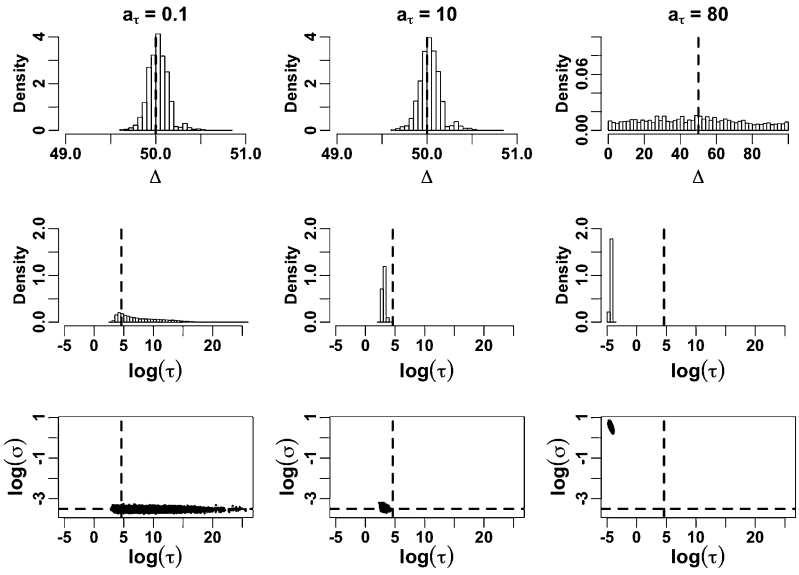

FIG. 16. *Each column shows posterior distribution of* Δ (*first row*), *that of* $\log(\tau)$ (*second row*), *and a scatter plot of* $\log(\sigma)$ *over* $\log(\tau)$ (*third row*) *obtained under three values of* a_τ (*columns*, a_τ =0.1, 10 *and* 80). *The generative values of* $(\Delta, \log(\sigma), \log(\tau))$ *are* $(50, -3.5, 4.6)$ *and represented by the dashed lines on each plot. The posterior distribution of the time delay is robust to the shape parameter* (a_τ) *as long as it is reasonably small. The ESS of* Δ *is* 3235, 3068, 3080, 478 *and* 3625 *from the left.*

from the generative value of $\log(\tau) = 4.6$. When the mode of $\log(\tau)$ reaches -5 ($\tau = \exp(-5) = 0.007 \ll$ 3-day observation cadence), the posterior distribution of Δ becomes flat.

E.2. Sensitivity analysis of the prior distribution of σ^2. We check the sensitivity of the posterior distribution of Δ to the scale parameter b_σ of the IG$(1, b_\sigma)$ prior distribution for σ^2. The effect of the unit shape parameter is negligible because the resultant shape parameter of the IG conditional posterior distribution of σ^2 in (C.4) is $n + 1$ so that n plays a dominant role in controlling the right tail behavior. We fix the IG$(1, b_\tau)$ prior distribution for τ, where b_τ is fixed at one day, as described in Section 2.5.

We display the result of the sensitivity analysis in Figure 17, where the values of b_σ are increasing from 0.001 to 10 from the first column. As the soft lower bound ($= b_\sigma/2$) increases from the left, the posterior distribution of the time delay becomes flatter. This is because the generative value of $\sigma^2 (= 0.03^2)$ is less than the soft lower bound for large values of b_σ, for example, when $b_\sigma = 10$ in the right most column, the IG$(1, 10)$ prior distribution of σ^2 exponentially cuts off the probability density in the region to left of the mode, 5 mag^2/day, which excludes the generative value of σ^2 $(= 0.03^2)$. Because the generative σ^2 is much smaller

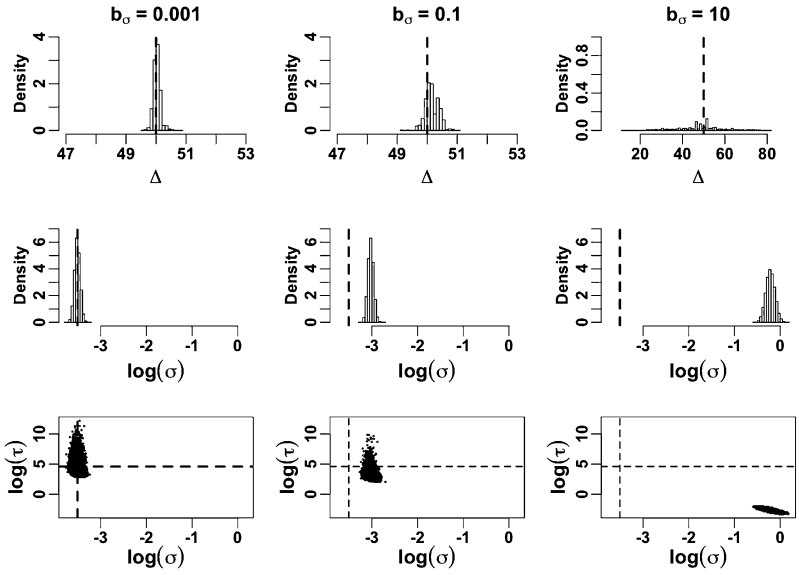

FIG. 17. *Each column shows posterior distribution of* Δ (*first row*), *that of* $\log(\sigma)$ (*second row*) *and a scatter plot of* $\log(\tau)$ *over* $\log(\sigma)$ (*third row*) *obtained under three values of* b_σ (*columns*, $b_\sigma = 0.001, 0.1$ *and* 10). *The generative values of* $(\Delta, \log(\sigma), \log(\tau))$ *are* $(50, -3.5, 4.6)$ *and represented by the dashed lines on each plot. The modes of the posterior distributions of parameters are near the generative values as the scale parameter* (*soft lower bound*) *decreases. The ESS of* Δ *is* 2957, 3082, 4148, 2459 *and* 23 *from the left.*

than the soft lower bound, the posterior distribution of σ^2 has negligible mass near the generative value of σ^2. Also, because the posterior samples of σ and τ are negatively correlated a posteriori as shown in the scatter plots, posterior distributions that favor large values of σ^2 also favor small values of τ. As discussed in Appendix E.1, when the posterior distribution of τ is concentrated on values smaller than the observational cadence, the posterior latent light curve $\boldsymbol{X}(\boldsymbol{t}^\Delta)$ effectively becomes a white noise sequence. In this case, it is difficult to constrain Δ.

The second row of Figure 17 shows that as the soft lower bound decreases from the right, the posterior distribution of $\log(\sigma)$ moves toward the generative value of $\log(\sigma) = -3.5$. Though not shown here, posterior distributions obtained under a value of b_σ smaller than 0.001 do not noticeably differ from that obtained with $b_\sigma = 0.001$. With the small soft lower bound, the modes of the posterior distributions of the other parameters tend to be near their generative values. We also found that the choice of b_σ is less important for large data sets, for example, with $n > 400$.

Acknowledgments. This work was conducted under the auspices of the CHASC International Astrostatistics Center. We thank CHASC members for many

helpful discussions and the editor, associate editor and reviewer for their careful reading and insightful suggestions.

SUPPLEMENTARY MATERIAL

R codes and data (DOI: 10.1214/17-AOAS1027SUPP; .zip). This zip file [Tak et al. (2017)] contains all the computer code (Rcode.R) and data (Data.zip) used in this article. An R package, `timedelay`, that implements the Bayesian and profile likelihood methods is publicly available at CRAN (https://cran.r-project.org/package=timedelay).

REFERENCES

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2015). Overall objective priors. *Bayesian Anal.* **10** 189–221. MR3420902

BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28. MR1702200

BERK, D. E. V., WILHITE, B. C., KRON, R. G., ANDERSON, S. F., BRUNNER, R. J., HALL, P. B., IVEZIĆ, Ž., RICHARDS, G. T., SCHNEIDER, D. P., YORK, D. G., BRINKMANN, J. V., LAMB, D. Q., NICHOL, R. C. and SCHLEGEL, D. J. (2004). The ensemble photometric variability of ∼25,000 quasars in the sloan digital sky survey. *Astrophys. J.* **601** 692.

BLANDFORD, R. and NARAYAN, R. (1992). Cosmological applications of gravitational lensing. *Annu. Rev. Astron. Astrophys.* **30** 311–358.

BROOKS, S. P., MORGAN, B. J., RIDOUT, M. S. and PACK, S. (1997). Finite mixture models for proportions. *Biometrics* **53** 1097–1115.

BROOKS, S., GELMAN, A., JONES, G. L. and MENG, X.-L., eds. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton, FL. MR2742422

CHANG, K. and REFSDAL, S. (1979). Flux variations of QSO 0957+561 A, B and image splitting by stars near the light path. *Nature* **282** 561–564.

COURBIN, F., CHANTRY, V., REVAZ, Y., SLUSE, D., FAURE, C., TEWES, M., EULAERS, E., KOLEVA, M., ASFANDIYAROV, I., DYE, S., MAGAIN, P., VAN WINCKEL, H., COLES, J., SAHA, P., IBRAHIMOV, M. and MEYLAN, G. (2013). COSMOGRAIL: The COSmological MOnitoring of GRAvItational lenses IX. time delays, lens dynamics and baryonic fraction in He 0435-1223. *Astron. Astrophys.* **536** A53.

DAVISON, A. C. (2003). *Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913

DOBLER, G., FASSNACHT, C., TREU, T., MARSHALL, P. J., LIAO, K., HOJJATI, A., LINDER, E. and RUMBAUGH, N. (2015). Strong lens time delay challenge. I. Experimental design. *Astrophys. J.* **799** 168.

FASSNACHT, C., PEARSON, T., READHEAD, A., BROWNE, I., KOOPMANS, L., MYERS, S. and WILKINSON, P. (1999). A determination of h_o with the CLASS gravitational lens B1608+656. I. time delay measurements with the VLA. *Astrophys. J.* **527** 498.

FISCHER, P., BERNSTEIN, G., RHEE, G. and TYSON, J. A. (1997). The mass distribution of the cluster Q0957+561 from gravitational lensing. *Astron. J.* **113** 521.

FOHLMEISTER, J., KOCHANEK, C. S., FALCO, E. E., WAMBSGANSS, J., OGURI, M. and DAI, X. (2013). A two-year time delay for the lensed quasar SDSS J1029+ 2623. *Astrophys. J.* **764** 186.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

HAINLINE, L. J., MORGAN, C. W., BEACH, J. N., KOCHANEK, C., HARRIS, H. C., TILLEMAN, T., FADELY, R., FALCO, E. E. and LE, T. X. (2012). A new microlensing event in the doubly imaged quasar Q 0957+561. *Astrophys. J.* **744** 104.

HARVA, M. and RAYCHAUDHURY, S. (2006). *Bayesian Estimation of Time Delays Between Unevenly Sampled Signals*. IEEE, New York.

HOJJATI, A., KIM, A. G. and LINDER, E. V. (2013). Robust strong lensing time delay estimation. *Phys. Rev. D* **87** 123512.

INADA, N., OGURI, M., MOROKUMA, T., DOI, M., YASUDA, N., BECKER, R. H., RICHARDS, G. T., KOCHANEK, C. S., KAYO, I., KONISHI, K. et al. (2006). SDSS J1029+2623: A gravitationally lensed quasar with an image separation of 225. *Astrophys. J. Lett.* **653** L97.

KELLY, B. C., BECHTOLD, J. and SIEMIGINOWSKA, A. (2009). Are the variations in quasar optical flux driven by thermal fluctuations? *Astrophys. J.* **698** 895.

KOCHANEK, C., MORGAN, N., FALCO, E., MCLEOD, B., WINN, J., DEMBICKY, J. and KETZEBACK, B. (2006). The time delays of gravitational lens He 0435–1223: An early-type galaxy with a rising rotation curve. *Astrophys. J.* **640** 47.

KOZŁOWSKI, S. and KOCHANEK, C. S. (2009). Discovery of 5000 active galactic nuclei behind the magellanic clouds. *Astrophys. J.* **701** 508.

KOZŁOWSKI, S., KOCHANEK, C. S., UDALSKI, A., SOSZYŃSKI, I., SZYMAŃSKI, M., KUBIAK, M., PIETRZYŃSKI, G., SZEWCZYK, O., ULACZYK, K. and POLESKI, R. (2010). Quantifying quasar variability as part of a general approach to classifying continuously varying sources. *Astrophys. J.* **708** 927.

KUMAR, S. R., STALIN, C. and PRABHU, T. (2014). H_0 from 11 well measured time-delay lenses. *Astron. Astrophys.* **580** A38.

LIAO, K., TREU, T., MARSHALL, P., FASSNACHT, C. D., RUMBAUGH, N., DOBLER, G., AGHAMOUSA, A., BONVIN, V., COURBIN, F., HOJJATI, A., JACKSON, N., KASHYAP, V., RATHNA KUMAR, S., LINDER, E., MANDEL, K., MENG, X. L., MEYLAN, G., MOUSTAKAS, L. A., PRABHU, T. P., ROMERO-WOLF, A., SHAFIELOO, A., SIEMIGINOWSKA, A., STALIN, C. S., TAK, H., TEWES, M. and VAN DYK, D. (2015). Strong lens time delay challenge: II. Results of TDC1. *Astrophys. J.* **800** 11.

LINDER, E. V. (2011). Lensing time delays and cosmological complementarity. *Phys. Rev. D* **84** 123529.

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR2401592

LSST SCIENCE COLLABORATION (2009). LSST Science Book, Version 2.0. Available at arXiv:0912.0201.

MACLEOD, C., IVEZIĆ, Ž., KOCHANEK, C., KOZŁOWSKI, S., KELLY, B., BULLOCK, E., KIMBALL, A., SESAR, B., WESTMAN, D., BROOKS, K., GIBSON, R., BECKER, A. C. and DE VRIES, W. H. (2010). Modeling the time variability of SDSS stripe 82 quasars as a damped random walk. *Astrophys. J.* **721** 1014.

MORGAN, C. W., HAINLINE, L. J., CHEN, B., TEWES, M., KOCHANEK, C. S., DAI, X., KOZLOWSKI, S., BLACKBURNE, J. A., MOSQUERA, A. M., CHARTAS, G., COURBIN, F. and MEYLAN, G. (2012). Further evidence that quasar X-ray emitting regions are compact: X-ray and optical microlensing in the lensed quasar q J0158-4325. *Astrophys. J.* **756** 52.

MOSQUERA, A. M. and KOCHANEK, C. S. (2011). The microlensing properties of a sample of 87 lensed quasars. *Astrophys. J.* **738** 96.

MUNOZ, J., FALCO, E., KOCHANEK, C., LEHÁR, J., MCLEOD, B., IMPEY, C., RIX, H.-W. and PENG, C. (1998). The CASTLES project. *Astrophys. Space Sci.* **263** 51–54.

OGURI, M. and MARSHALL, P. J. (2010). Gravitationally lensed quasars and supernovae in future wide-field optical imaging surveys. *Mon. Not. R. Astron. Soc.* **405** 2579–2593.

OSCOZ, A., MEDIAVILLA, E., GOICOECHEA, L. J., SERRA-RICART, M. and BUITRAGO, J. (1997). Time delay of QSO 0957+561 and cosmological implications. *Astrophys. J. Letters* **479** L89.

OSCOZ, A., ALCALDE, D., SERRA-RICART, M., MEDIAVILLA, E., ABAJAS, C., BARRENA, R., LICANDRO, J., MOTTA, V. and MUNOZ, J. (2001). Time delay in QSO 0957+561 from 1984-1999 optical data. *Astrophys. J.* **552** 81.

PELT, J., HOFF, W., KAYSER, R., REFSDAL, S. and SCHRAMM, T. (1994). Time delay controversy on QSO 0957+ 561 not yet decided. *Astron. Astrophys.* **286** 775–785.

PELT, J., KAYSER, R., REFSDAL, S. and SCHRAMM, T. (1996). The light curve and the time delay of QSO 0957+561. *Astron. Astrophys.* **305** 97–106.

R CORE TEAM (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

REFSDAL, S. (1964). The gravitational lens effect. *Mon. Not. R. Astron. Soc.* **128** 295–306. MR0175607

ROBERTS, G. O. and ROSENTHAL, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.* **44** 458–475. MR2340211

SCHNEIDER, P., EHLERS, J. and FALCO, E. (1992). *Gravitational Lenses*. Springer, Berlin.

SCHNEIDER, P., WAMBSGANSS, J. and KOCHANEK, C. S. (2006). *Gravitational Lensing: Strong, Weak and Micro*. Springer, Berlin.

SERRA-RICART, M., OSCOZ, A., SANCHÍS, T., MEDIAVILLA, E., GOICOECHEA, L. J., LICANDRO, J., ALCALDE, D. and GIL-MERINO, R. (1999). BVRI photometry of QSO 0957+561A, B: Observations, new reduction method, and time delay. *Astrophys. J.* **526** 40.

SHALYAPIN, V., GOICOECHEA, L. and GIL-MERINO, R. (2014). A 5.5-year robotic optical monitoring of Q0957+561: Substructure in a non-local cD galaxy. *Astron. Astrophys.* **540** A132.

SUYU, S., AUGER, M., HILBERT, S., MARSHALL, P., TEWES, M., TREU, T., FASSNACHT, C., KOOPMANS, L., SLUSE, D., BLANDFORD, R., COURBIN, F. and MEYLAN, G. (2013). Two accurate time-delay distances from strong lensing: Implications for cosmology. *Astrophys. J.* **766** 70.

TAK, H., KELLY, J. and MORRIS, C. N. (2017). Rgbp: An R package for Gaussian, Poisson, and Binomial Random Effects Models with Frequency Coverage Evaluations. *J. Stat. Softw.* **78** 1–33.

TAK, H., MANDEL, K., VAN DYK, D. A., KASHYAP, V. L., MENG, X.-L and SIEMIGINOWSKA, A. (2017). Supplement to "Bayesian estimates of astronomical time delays between gravitationally lensed stochastic light curves." DOI:10.1214/17-AOAS1027SUPP.

TEWES, M., COURBIN, F. and MEYLAN, G. (2013). COSMOGRAIL: The COSmological MOnitoring of GRAvItational lenses XI. techniques for time delay measurement in presence of microlensing. *Astrophys. J.* **605** 58.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. MR1329166

TREU, T. and MARSHALL, P. J. (2016). Time delay cosmography. *Astron. Astrophys. Rev.* **24** 11.

UHLENBECK, G. E. and ORNSTEIN, L. S. (1930). On the theory of the Brownian motion. *Phys. Rev.* **36** 823–841.

VAN DYK, D. A. and MENG, X.-L. (2001). The art of data augmentation. *J. Comput. Graph. Statist.* **10** 1–111. MR1936358

WALSH, D., CARSWELL, R. and WEYMANN, R. (1979). 0957+561 A, B- twin quasistellar objects or gravitational lens. *Nature* **279** 381–384.

YU, Y. and MENG, X.-L. (2011). To center or not to center: That is not the question—an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *J. Comput. Graph. Statist.* **20** 531–570. MR2878987

ZU, Y., KOCHANEK, C., KOZŁOWSKI, S. and UDALSKI, A. (2013). Is quasar optical variability a damped random walk? *Astrophys. J.* **765** 106.

H. TAK
STATISTICAL AND APPLIED MATHEMATICAL SCIENCES INSTITUTE
19 T.W. ALEXANDER DRIVE
DURHAM, NORTH CAROLINA 27703
USA
E-MAIL: hyungsuk.tak@gmail.com

K. MANDEL
V. L. KASHYAP
A. SIEMIGINOWSKA
HARVARD-SMITHSONIAN CENTER FOR ASTROPHYSICS
HARVARD UNIVERSITY
60 GARDEN STREET
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: kmandel@cfa.harvard.edu
vkashyap@cfa.harvard.edu
asiemiginowska@cfa.harvard.edu

D. A. VAN DYK
STATISTICS SECTION
DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE LONDON
LONDON SW7 2AZ
UNITED KINGDOM
E-MAIL: dvandyk@imperial.ac.uk

X.-L. MENG
DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
1 OXFORD STREET
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: meng@stat.harvard.edu