

Time delays and stationarity in quasar light curves

Namu Kroupa^{1,2,3,*}, David Yallup^{2,4} and Will Handley^{2,4}

¹*Astrophysics Group, Cavendish Laboratory, J.J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom*

²*Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, United Kingdom*

³*Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*

⁴*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom*



(Received 29 October 2025; accepted 23 February 2026; published 19 March 2026)

We present a fully Bayesian framework for time delay inference and stationarity tests in quasar light curves using marginalized Gaussian processes. The model separates a deterministic, nonstationary drift (piecewise linear mean) from stationary stochastic variability (Matérn and Spectral Mixture kernels), and jointly models multiple images with per-image microlensing. Bayesian evidence and parameter posteriors are obtained via nested sampling and marginalized over model choices. Applied to the quasars WFI J2033–4723, B 1608 + 656, and HE 0435–1223, we find strong evidence for nonstationarity in B 1608 + 656 and HE 0435–1223, while WFI J2033–4723 is consistent with stationarity. The stochastic component favors an Markovian exponential kernel for B 1608 + 656 and a non-Markovian Matérn- $\frac{3}{2}$ kernel for WFI J2033–4723 and HE 0435–1223. Multilength-scale Spectral Mixture kernels are disfavored. Time delays are shown to be robust to model assumptions and consistent with prior work within the error. We further identify and mitigate a likelihood pathology which biases toward large delays, providing a practical nested sampling convergence protocol.

DOI: [10.1103/87nn-klh9](https://doi.org/10.1103/87nn-klh9)

I. INTRODUCTION

The study of quasars, a particularly bright class of active galactic nuclei (AGN), serves as an important probe for understanding galaxy evolution and cosmology [1,2]. A defining characteristic is their pronounced variability across the electromagnetic spectrum, observed over a vast range of timescales, from minutes to decades [3–5]. This variability is intrinsically linked to the physical processes occurring in the vicinity of the supermassive black hole, including instabilities in the accretion disk, thermal fluctuations, and the reprocessing of radiation, providing information about the accretion disc, the supermassive black hole properties, and the circumnuclear environment [6–8].

The analysis of quasar light curves, therefore, provides a method for investigating these unresolved regions and testing accretion theories, but presents several difficulties. Ground-based surveys often produce irregularly sampled

data with seasonal gaps, complicating standard time series analysis [9,10]. Photometric measurements also include noise, which can be confused with low-amplitude intrinsic variability. Quasar variability is largely stochastic, often characterized as red noise, with greater power on longer timescales [6,11]. This stochasticity makes it challenging to identify deterministic signals like periodicities. Moreover, quasar light curves exhibit a wide spectrum of complex behaviours, including long-term trends and distinct patterns that are not fully captured by simple stationary stochastic models [12,13]. These observations suggest the influence of nonstationarity in the underlying physical processes or the combined effect of multiple interacting mechanisms.

The challenge of accurately estimating time delays from complex, noisy, and irregularly sampled light curves is central not only to understanding intrinsic AGN physics but also to their cosmological applications. Time delay cosmography, which utilizes the time delays between multiple images of a gravitationally lensed variable quasar to measure absolute distances and infer the Hubble parameter H_0 [14,15], critically depends on robust delay measurements. Within the context of the Hubble tension [16,17], time delay cosmography provides precise and accurate H_0 values [18–21]. However, observational challenges, including distortions from microlensing, persist [22–24]. A statistically analogous problem to

*Contact author: nk544@cam.ac.uk

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

gravitational lens time delay estimation is faced by reverberation mapping, used to probe the broad-line region and accretion disc sizes [25–29].

Traditional Fourier-based methods for time series analysis, such as those based on the power spectral density or the cross-spectrum, struggle with the irregular sampling typical of astronomical data [4,30–32]. While localized variability can be identified using methods like Bayesian Blocks [33] and machine learning offers new approaches [34,35], stochastic process models remain a cornerstone. The damped random walk model gained popularity for its simplicity in describing quasar optical variability [6,11,36]. The empirical linear *rms-flux* relation [37] and the associated lognormal flux distribution of quasars found theoretical backing in the damped random walk framework, where the flux is given by $x(t) = \exp[\ell(t)]$ with $\ell(t)$ being a linear Gaussian stationary stochastic process [38]. However, deviations from damped random walk predictions, particularly steeper power spectral density slopes at short timescales [39,40] and nonstationarity in highly variable AGN like IRAS 13224–3809 [41], evidenced by nonlognormal flux distributions and nonlinear rms-flux relations [42,43], necessitate more flexible models. Knowing whether a stochastic process is stationary is crucial for linking variability to physical models and understanding connections between emission components [41].

Gaussian processes (GPs) provide a nonparametric Bayesian framework for modeling such time series, naturally handling irregular sampling and offering principled uncertainty quantification [22,28,44,45]. The choice of the GP kernel and its hyperparameters is critical. Traditional maximum likelihood estimation can be problematic for complex likelihood surfaces or influential priors. Furthermore, preselecting a single kernel introduces strong biases, especially for extrapolation tasks like time delay estimation, where stationary kernels with simple mean functions revert to the prior mean far from data [46,47]. The GP approach also avoids the insensitivity of binned data methods to nonstationary variations on timescales shorter than the bin width.

A standard frequentist hypothesis test for stationarity, the augmented Dickey-Fuller test [43,48], has its own limitations. Firstly, it allows only pairwise comparison between a specific nonstationary and stationary model. Reformulated in terms of GPs, the null hypothesis may be rewritten as possessing a polynomial mean function with a nonstationary kernel, while the alternative hypothesis possesses a mean function which is a polynomial of one order lower and a stationary kernel. When extrapolating a time series, which is strictly necessary for time delay inference [47], polynomial mean functions may lead to unphysical diverging light curve magnitudes unless higher-order coefficients are heavily regularized. Secondly, the precise form of the GP kernel in each hypothesis is set by

the number of time lags included in the test, which can itself be determined by an Akaike information criterion or similar. However, as observed previously [46,49,50], Gaussian process posterior distribution may deviate strongly from the assumption of Gaussianity underlying such information criteria and hence bias the inference.

As a Bayesian method, GPs are amenable to extension in a fully Bayesian framework by marginalizing over kernel and mean function choices to improve robustness [46]. This paper focuses on flexible, data-driven mean functions parameterized as piecewise linear splines with flexible knots [51–53] combined with stationary kernels from the Matérn and Spectral Mixture Kernel (SM) families. Especially, we do not rely on time series descriptors, which are ultimately proxies for underlying stochastic processes [54]. Finite data, especially when the observation length is insufficient relative to intrinsic timescales, limits inference [46,55,56]. We therefore use Bayesian inference to make probabilistic statements about model comparison. This approach of marginalizing over the mean function differs from and aims to be more robust than GP applications with fixed, simple mean functions [22]. The potential to link GP kernel parameters to underlying quasar properties [9,10] further motivates the development of robust kernel selection and mean function modeling. Time delays between different wavebands, another key observable related to disc reprocessing, also stand to benefit from such improved modeling.

In this paper, we analyze the light curves of the quadruply lensed quasars WFI J2033–4723 [57], B 1608 + 656 [58], and HE 0435–1223 [59]. These objects represent well-studied benchmark systems for time delay cosmography and microlensing studies due to their unique physical properties and the availability of long term, high-quality monitoring data, which make them ideal for Gaussian process modeling.

WFI J2033–4723 [57,60] is distinguished by a close pair of lensed images and exhibits significant microlensing, providing a valuable dataset for constraining quasar accretion disc sizes [61]. Long-term monitoring, notably by the COSMOGRAIL program [62], has yielded precise time delays for this system. In contrast, the radio-selected lens B 1608 + 656 [63] features a complex environment of two interacting lens galaxies and is largely free of microlensing in radio observations. This has made it a crucial target for clean measurements of intrinsic quasar variability and a foundational system for determining H_0 through VLA monitoring [58,64–66]. Finally, HE 0435–1223 [67] is a bright quad with a well-resolved Einstein ring and prominent microlensing [61]. It has become a cornerstone for H_0 measurements by the H0LiCOW [68–72] and TDCOSMO [20,21] Collaborations, owing to extensive and precise time delay data from COSMOGRAIL [73].

The multidecade and densely sampled light curves for these three objects provide multiple time-shifted views of

the same intrinsic signal. This wealth of data allows for the robust separation of intrinsic quasar variability from extrinsic microlensing effects, setting them apart from typical quasar observations and enabling precise astrophysical measurements.

The structure of this paper is as follows. Section II introduces the flux variability model. Section III describes the adopted protocol for sampling the GP model as applied to the light curves of WFI J2033–4723 [57], B 1608 + 656 [58], and HE 0435–1223 [59]. The inference results are explained in Sec. IV. Finally, the conclusions are provided in Sec. V.

II. BACKGROUND

In this section, we first explain the modeling choices for the quasar light curves and consequently describe the concrete flux variability model implemented. Finally, we describe how separate models are combined with marginalized GPs.

A. Time delayed Gaussian processes

We model quasar light curves with time delayed GPs. In the standard GP modeling paradigm [44], these are defined with a zero-mean function and a stationary kernel and tend to revert to their prior mean when predicting far from the observed data. This characteristic hinders their ability to extrapolate, which is crucial when dealing with significant time delays where the function and its shifted version become effectively uncorrelated [47]. This limitation can lead to pathological structures in the posterior distribution of the time delay, creating practical issues for inference, as was recently established as a primary challenge for time delay inference.

To improve extrapolation, we are therefore required to choose nonstationary GPs. For this, we may either choose nonstationary kernels, which do not decay over a certain length scale, or a nonconstant mean function. Due to the success of the damped random walk model and related stochastic processes [6] and motivated by previous modeling of quasar light curves with splines [74] or Bézier curves [12], we choose to retain a stationary kernel and proceed with a nonconstant mean function. One possibility to proceed is to use the reformulation of a spline as the posterior mean of a GP [75–77]. However, this formulation restricts the knots of the spline to lie on the data points. A less restrictive method of spline regression are *flexknots* [51–53, 78–89], which promotes the knots to parameters which can be sampled.

A final modeling choice lies in the order of the splines. While in principle this is a choice which can be marginalized over, we choose linear splines as this confines the effect of each linear segment of the spline to its own interval and does not induce correlations between data points far apart, compared to cubic splines

for instance. This is desirable from an interpretability standpoint.

In summary, we choose stationary kernels and a mean function parametrized by a linear spline with flexible knots. In the next section, we concretely describe the adopted GP model for the flux variability.

B. Flux variability model

Following Hojjati *et al.* [22], we assume that all observed light curves are instances of a latent GP but with shifts in magnitude and time, additional noise and a distortion from microlensing, also modeled as a GP. Departing from the standard GP modeling paradigm, we use parametric models for long range variations and nonparametric models to model variations over short scales. This separation of scales is obtained in a purely data-driven manner.

In the following, we describe the GP model for the case of $n_{lc} = 3$ light curves, A, B and C, with magnitudes $\mathbf{y}_A$, $\mathbf{y}_B$ and $\mathbf{y}_C$, respectively, as the generalization to more light curves is straightforward. From here on, we use the notation $\mathbf{f}(\mathbf{t})$ to denote the values of the function f at the observation times $\mathbf{t} = [t_1, \dots, t_{n_{data}}]^\top$ stacked into a column vector. Likewise, we denote a matrix evaluated at all pairs of observation times by $\mathbf{K}(\mathbf{t}, \mathbf{t})_{ij} = k(t_i, t_j)$.

1. Mean function

We set the mean function of light curve A to $m_A(t) = \bar{\mathbf{y}}_A + m_{fk}(t)$ by calculating the average $\bar{\mathbf{y}}_A$ of $\mathbf{y}_A$ and adding the flexknot m_{fk} , which we define further below. At the observation times $\mathbf{t}$, the values of the mean function for light curve A are thus

$$\mathbf{m}_A = \bar{\mathbf{y}}_A \mathbf{1} + \mathbf{m}_{fk}(\mathbf{t}), \quad (1)$$

where $\mathbf{1} = [1, 1, \dots, 1]^\top$. The mean functions of the other light curves differ from $\mathbf{m}_A$ by a relative offset in magnitude and a shift in time,

$$\mathbf{m}_B = \mathbf{m}_A(\mathbf{t} + \Delta t_{AB} \mathbf{1}) + \Delta m_{AB} \mathbf{1}, \quad (2)$$

$$\mathbf{m}_C = \mathbf{m}_A(\mathbf{t} + \Delta t_{AC} \mathbf{1}) + \Delta m_{AC} \mathbf{1}, \quad (3)$$

where Δm_{AB} and Δm_{AC} are the constant offsets of light curve B and C with respect to A. The light curves are therefore shifted instances of the latent GP with mean function $\mathbf{m}_{fk}$. We are thus using light curve A as a reference light curve and our mean function is equivalent to subtracting the average magnitude of light curve A from all light curves and modeling a latent GP. Therefore, we set uniform priors on Δm_{AB} and Δm_{AC} with range given by the maximum extent of the light curve shifted by light curve A, $[\min_i(\mathbf{y}_j - \bar{\mathbf{y}}_A)_i, \max_i(\mathbf{y}_j - \bar{\mathbf{y}}_A)_i]$ for light curve $j \in \{B, C\}$.

2. Flexknots

We follow the implementation in Ormondroyd *et al.* [53]. A flexknot is a linear spline whose knots $(t_n^{(\text{fk})}, y_n^{(\text{fk})})_{n=1}^{n_{\text{fk}}}$ are parameters. The number of knots n_{fk} is treated as a hyperparameter (Sec. II C). The t -values of the first $(t_1^{(\text{fk})})$ and last $(t_{n_{\text{fk}}}^{(\text{fk})})$ knot are fixed to the t -values of the first (t_1) and last $(t_{n_{\text{data}}})$ data points, respectively. A sorted-uniform prior [90,91] is set on the remaining $n_{\text{fk}} - 1$ t -values $(t_n^{(\text{fk})})_{n=2}^{n_{\text{fk}}-1}$. The sorted-uniform prior removes redundant regions of the parameter space given by relabeling of the knots. A uniform prior is set on the n_{fk} y -values. We take the convention that the spline continues to extrapolate linearly to the left and right. So far, the flexknot has $2n_{\text{fk}} - 2$ parameters with $n_{\text{fk}} \geq 2$. The value $n_{\text{fk}} = 0$ is now defined as the zero function and $n_{\text{fk}} = 1$ is set to the constant function with a uniform prior set on the constant c_{fk} . In summary, the flexknot now has $2n_{\text{fk}} - 1$ parameters with $n_{\text{fk}} \geq 0$ and is defined by

$$m_{\text{fk}}(t) = \begin{cases} 0 & \text{if } n_{\text{fk}} = 0, \\ c_{\text{fk}} & \text{if } n_{\text{fk}} = 1, \\ \text{spline}_{(t_n^{(\text{fk})}, y_n^{(\text{fk})})_{n=1}^{n_{\text{fk}}}}(t) & \text{if } n_{\text{fk}} \geq 2. \end{cases} \quad (4)$$

Increasing n_{fk} therefore increases the model complexity.

3. Covariance matrix

The covariance matrix takes a block form. The diagonal blocks account for the correlations within each light curve,

$$\mathbf{K}_A = \mathbf{K}(\mathbf{t}, \mathbf{t}), \quad (5)$$

$$\mathbf{K}_B = \mathbf{K}(\mathbf{t} + \Delta t_{AB} \mathbf{1}, \mathbf{t} + \Delta t_{AB} \mathbf{1}), \quad (6)$$

$$\mathbf{K}_C = \mathbf{K}(\mathbf{t} + \Delta t_{AC} \mathbf{1}, \mathbf{t} + \Delta t_{AC} \mathbf{1}). \quad (7)$$

The off-diagonal blocks represent the correlation between different light curves,

$$\mathbf{K}_{AB} = \mathbf{K}(\mathbf{t}, \mathbf{t} + \Delta t_{AB} \mathbf{1}), \quad (8)$$

$$\mathbf{K}_{AC} = \mathbf{K}(\mathbf{t}, \mathbf{t} + \Delta t_{AC} \mathbf{1}), \quad (9)$$

$$\mathbf{K}_{BC} = \mathbf{K}(\mathbf{t} + \Delta t_{AB} \mathbf{1}, \mathbf{t} + \Delta t_{AC} \mathbf{1}). \quad (10)$$

We choose the kernel function from the family of Matérn kernels [Exponential (E), Squared Exponential (SE), Matérn-3/2 (M32), Matérn-5/2 (M52), Matérn-7/2 (M72)], the Rational Quadratic (RQ) kernel and Spectral Mixture (SM) kernels. These choices are defined and discussed in the Appendix. The choice itself is treated as a hyperparameter (Sec. II C). Distortion from microlensing on each light curve is modeled with a GP with a SE kernel.

Hence, the microlensing term $\mathbf{K}_\mu$ is added to each block diagonal of the covariance matrix.

Finally, we include the error bars on the magnitude measurements, σ_A , σ_B and σ_C , as well as one adjustable white noise parameter σ so that the noise terms are

$$\Sigma_A = \text{diag}(\sigma_A)^2 + \sigma^2 \mathbf{I}, \quad (11)$$

$$\Sigma_B = \text{diag}(\sigma_B)^2 + \sigma^2 \mathbf{I}, \quad (12)$$

$$\Sigma_C = \text{diag}(\sigma_C)^2 + \sigma^2 \mathbf{I}, \quad (13)$$

where $\text{diag}(\sigma_A)$ places the entries of σ_A on the diagonal of a matrix and $\mathbf{I}$ is the identity matrix.

Uniform priors are set on all covariance matrix parameters. Priors on amplitude parameters (kernel amplitude A , SM kernel weights w_q , microlensing kernel amplitude A_μ) range from zero to the square root of half of the light curve magnitude range [46]. Length scale priors (kernel length scale ℓ , SM kernel scale ℓ_q , microlensing kernel length scale ℓ_μ) are set from zero to the difference between the largest and smallest observation times of the dataset, $\Delta t_{\text{range}} = t_{n_{\text{data}}} - t_1$. For the RQ parameter α , it was generically observed [46] that most of the posterior mass, covered by more than five standard deviations around the mean, is roughly within the interval $[0, 10]$ so that we use this range for the prior bounds here. The prior on each SM kernel frequency f_q is set to range from zero to the Nyquist frequency $f_{\text{Nyquist}} = \frac{n_{\text{data}}}{2\Delta t_{\text{range}}}$. The white noise term σ is set to range from zero to the maximum error bar among all light curves, $\max_{i,j}(\sigma_i)_j$.

In practice, we observe that the bulk of the posterior mass is well-contained within these prior ranges, except for the length scale of the E kernel, which is attributed to the data and not an intrinsic feature of the likelihood [47].

4. Full Gaussian process model

The magnitudes $\mathbf{y}_A$, $\mathbf{y}_B$ and $\mathbf{y}_C$ are stacked into a single column vector. Combining the individual light curves in such a way is known as a multioutput GP,

$$\begin{bmatrix} \mathbf{y}_A \\ \mathbf{y}_B \\ \mathbf{y}_C \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{m}_A \\ \mathbf{m}_B \\ \mathbf{m}_C \end{bmatrix}, \mathbf{K} \right) \quad (14)$$

with covariance matrix

TABLE I. Priors set on the parameters of the flux variability model. All priors are uniform in the ranges shown. “Sorted” indicates that a sorted-uniform prior [90,91] was used. The top set of parameters corresponds to the time delays, the middle set to the mean function and the bottom set to the covariance matrix. The prior on the time delay Δt was set differently for the WFI J2033–4723 and B 1608 + 656 (1) and HE 0435–1223 (2) datasets, respectively. The range of t -values of the light curves is denoted by $\Delta t_{\text{range}} = t_{n_{\text{data}}} - t_1$, mean and std denote the mean and standard deviation of time delays reported in Hojjati *et al.* [22], $\bar{y}_A$ is the average of the magnitudes y_A of light curve A and f_{Nyquist} is the Nyquist frequency computed from $t_1, \dots, t_{n_{\text{data}}}$.

Parameter	Prior range
(1) $\Delta t_{AB}, \Delta t_{AC}, \Delta t_{AD}$	$[-\Delta t_{\text{range}}, \Delta t_{\text{range}}]$
(2) $\Delta t_{AB}, \Delta t_{AC}, \Delta t_{AD}$	$[\text{mean} - 10 \cdot \text{std}, \text{mean} + 10 \cdot \text{std}]$
Δm_{Aj} for $j \in \{B, C, D\}$	$[\min_i (y_j - \bar{y}_A)_i, \max_i (y_j - \bar{y}_A)_i]$
n_{fk}	$\{0, 1, \dots, 6\}$
c_{fk}	$[\min_i (y_A - \bar{y}_A)_i, \max_i (y_A - \bar{y}_A)_i]$
$y_1^{(\text{fk})}, \dots, y_{n_{\text{fk}}}^{(\text{fk})}$	$[\min_i (y_A - \bar{y}_A)_i, \max_i (y_A - \bar{y}_A)_i]$
$t_2^{(\text{fk})}, \dots, t_{n_{\text{fk}}-1}^{(\text{fk})}$	Sorted($[t_1, t_{n_{\text{data}}}]$)
A, w_q, A_μ	$\left[0, \max_i \sqrt{\frac{\max_j (y_i)_j - \min_j (y_i)_j}{2}}\right]$
ℓ, ℓ_q, ℓ_μ	$[0, \Delta t_{\text{range}}]$
α	$[0, 10]$
f_q	$[0, f_{\text{Nyquist}}]$
σ	$[0, \max_{i,j} (\sigma_i)_j]$

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_A + \mathbf{K}_\mu + \Sigma_A & \mathbf{K}_{AB} & \mathbf{K}_{AC} \\ \mathbf{K}_{AB}^\top & \mathbf{K}_B + \mathbf{K}_\mu + \Sigma_B & \mathbf{K}_{BC} \\ \mathbf{K}_{AC}^\top & \mathbf{K}_{BC}^\top & \mathbf{K}_C + \mathbf{K}_\mu + \Sigma_C \end{bmatrix}. \quad (15)$$

In summary, the flexknot m_{fk} has $2n_{\text{fk}} - 1$ parameters, the offsets and time delays add $2(n_{\text{lc}} - 1)$ parameters, the white noise term adds one parameter and the microlensing kernels add two parameters since the SE kernel has two hyperparameters. The GP kernel contributes its own parameters, the number of which depends on the kernel family (Appendix).

The priors on these parameters are summarized in Table I. In general, the prior ranges were chosen to depend only on the bounding data, i.e., the minimum and maximum extent of the light curves along the t - and y -axes, or data averages, as this sets the rough scale of the data without explicitly using all data points. In practice, this achieves that the priors are both sufficiently wide and reasonably uninformative. A drawback of wide uninformative priors is that it becomes computationally more expensive to converge to the dominant mode of the posterior, as it occupies a smaller fraction of the parameter space. We will see in Sec. III B that incorporating more

a priori knowledge on the time delay parameter indeed accelerates the nested sampling (NS) run.

C. Marginalized Gaussian processes

The flexibility of our GP model is represented by the set of choices for the kernel and the mean function. A drawback of this is that we could potentially overfit the light curves. This motivates us to use Bayesian inference, wherein the Bayesian evidence selects the most probable model. Taking a step further, it is possible to marginalize over all such choices using marginalized GPs [46,92]. In comparison to traditional GPs, they infer the joint posterior of the discrete choice over different mean functions, kernels or noise models as well as GP kernel parameters. This is done by first inferring the posterior over GP parameters using Bayes theorem,

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}, \quad (16)$$

where $\mathcal{D}$ is a dataset, $\boldsymbol{\theta}$ are the parameters and $\mathcal{M}$ is the model choice, and using the computed evidences $p(\mathcal{D}|\mathcal{M})$ to infer the posterior over the choices,

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})p(\mathcal{M})}{p(\mathcal{D})}. \quad (17)$$

In all of our calculations, the prior $p(\mathcal{M})$ on the model choice is chosen to be uniform.

We note that this approach presents a significant generalization of the standard method of maximizing the likelihood $p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$ over the parameters (commonly called the hyperparameters in the context of GPs), known as type II maximum likelihood [44].

It is possible to perform the calculation of both Eqs. (16) and (17) in a single NS run by appropriately augmenting the NS parameter space with model choice parameters [46]. We take the alternative approach of separately computing the posterior and evidence for each choice [Eq. (16)] since proper sampling of the posterior requires manual choice-dependent tuning of NS convergence parameters (Sec. III) due to intrinsic pathologies of the time delay likelihood [47].

III. METHOD

In this section, we describe the method for sampling the posteriors. Finally, we explain how the inference of the time delays is stabilized with respect to the aforementioned pathologies in the likelihood.

A. Sampling

For each combination of flexknot mean function and kernel choice, a NS run is performed for the quasars WFI J2033–4723 [57] ($n_{\text{data}} = 218$, $n_{\text{lc}} = 3$), B 1608 + 656

[58] ($n_{\text{data}} = 228$, $n_{\text{lc}} = 4$) and HE 0435–1223 [59] ($n_{\text{data}} = 884$, $n_{\text{lc}} = 4$). In practice, the PolyChord [91,93] implementation of NS is used. Since the runtime of a NS run is proportional to the likelihood evaluation time [94,95], the NS runtime scales here as $\mathcal{O}((n_{\text{data}}n_{\text{lc}})^3)$ [44]. In practice, a single run typically took approximately 10^2 CPU hours for the smallest dataset, WFI J2033–4723, to 10^3 CPU days for the largest dataset, HE 0435–1223. Evidences and parameter posteriors are computed by postprocessing samples with ANESTHETIC [96]. The flux variability model was implemented with the TINYGP library [97].

We sample the posterior with NS [98], primarily for three reasons. Firstly, NS naturally incorporates uniform priors and priors with hard constraints, such as those from the flexknots, as NS can operate without gradients of the likelihood. Secondly, the GP likelihood with a joint uniform prior on all hyperparameters may include singular covariance matrices which lead to divergent negative log likelihoods, i.e., zero likelihood. NS, in particular PolyChord, adaptively ignores such regions without changing the evidence. Thirdly, kernel, flexknot and time delay posteriors are observed to be typically multimodal and exhibit extended degeneracies [47], which global methods such as NS handle automatically.

B. Protocol for nested sampling convergence

We observe that the time delay inference of all datasets suffers from the instability in Kroupa and Handley [47], i.e., the time delay posterior is peaked at values comparable to the width of the observation window. Notably, this occurs even for nonstationary mean functions, namely for flexknots with a large number of knots. This suggests that the structure of the likelihood described in Kroupa and Handley [47] persists even if the assumption of stationarity is relaxed.

To mitigate this problem, we adopt the following protocol for the WFI J2033–4723 and B 1608 + 656 light curves. First, the individual NS runs are performed with default PolyChord settings. At this stage most runs already produce time delays consistent with previously published values. Any runs for which the time delays are of order $\mathcal{O}(10^3 \text{ days})$ are rerun with the number of live points increased by a factor of 10. A second iteration of this (increasing the number of live points by a total factor of 100) is found to be sufficient to ensure that the time delays of all but one mean function and kernel combination are consistent within the error. We hypothesize that the remaining unconverged run (four knots with an SM_1 kernel for WFI J2033–4723) did not converge due to mode evaporation. That is, if multiple modes are populated with live points during a NS run, it is possible for one or more modes to spontaneously lose their live points to another mode (“evaporate”) due to stochastic fluctuations in the live point population per mode. This can be mitigated by further

increasing the number of live points, however this is costly in terms of the runtime [94,95] of NS. Changing the random number generator seed, however, resulted in convergence with default PolyChord settings, showing that NS is overall unstable with respect to this likelihood. For both WFI J2033–4723 and B 1608 + 656, the prior on the time delay is uniform on the range $\pm \Delta t_{\text{range}}$.

For HE 0435–1223, we observe that all NS runs initially yield large time delays. This is unchanged under the above protocol, within computational limitations. This is explained by noting that the dataset of HE 0435–1223 is significantly larger. Therefore, the posterior must be significantly more narrow around the true time delay when measured in terms of fractional parameter space volume. Due to the exceeding computational cost of increasing the number of live points further, we instead use a narrow uniform prior on the time delays. The prior is centered on the mean time delays inferred in Hojjati *et al.* [22] and widths set to 10 times the reported standard deviation. We note that approximate evidence values for the full-width prior can be obtained by appropriate rescaling, assuming that the posteriors are approximately zero outside the narrow prior. The rescaling factor is the same for all models here and the posterior probabilities of the models are therefore unaffected by such a rescaling. Therefore, we do not apply this correction. We find that inference with the narrow priors recovers well-constrained time delays with default PolyChord settings. Further, as the likelihood evaluation time is significantly larger, we fit a smaller number of models to the data, in particular limiting the maximum order of the Spectral Mixture kernels to SM_2 .

IV. RESULTS

The following results were all obtained by regressing the light curves in units of relative magnitude instead of flux, as supported by previous work [37,38]. As an independent test, we also refitted our GP model on the WFI J2033–4723 dataset in units of flux, from which we obtained evidence values less than -350 , which are strongly disfavored compared to the fits in units of magnitude. Thus, WFI J2033–4723 prefers a GP model which is log-normal in units of flux, and we proceed with fits in units of magnitude.

A. Kernel and mean function inference

1. Kernel inference

The posterior over mean function and kernel choices in terms of the model evidence $\log Z$ is shown in Fig. 1 and the corresponding marginalized fits are shown in Fig. 2. WFI J2033–4723, B 1608 + 656 and HE 0435–1223 exhibit a preference to the M_{32} , E and M_{32} kernel, respectively. The E kernel corresponds to a damped random walk and this preference is consistent with general observations on quasar light curves [40]. The preference to the

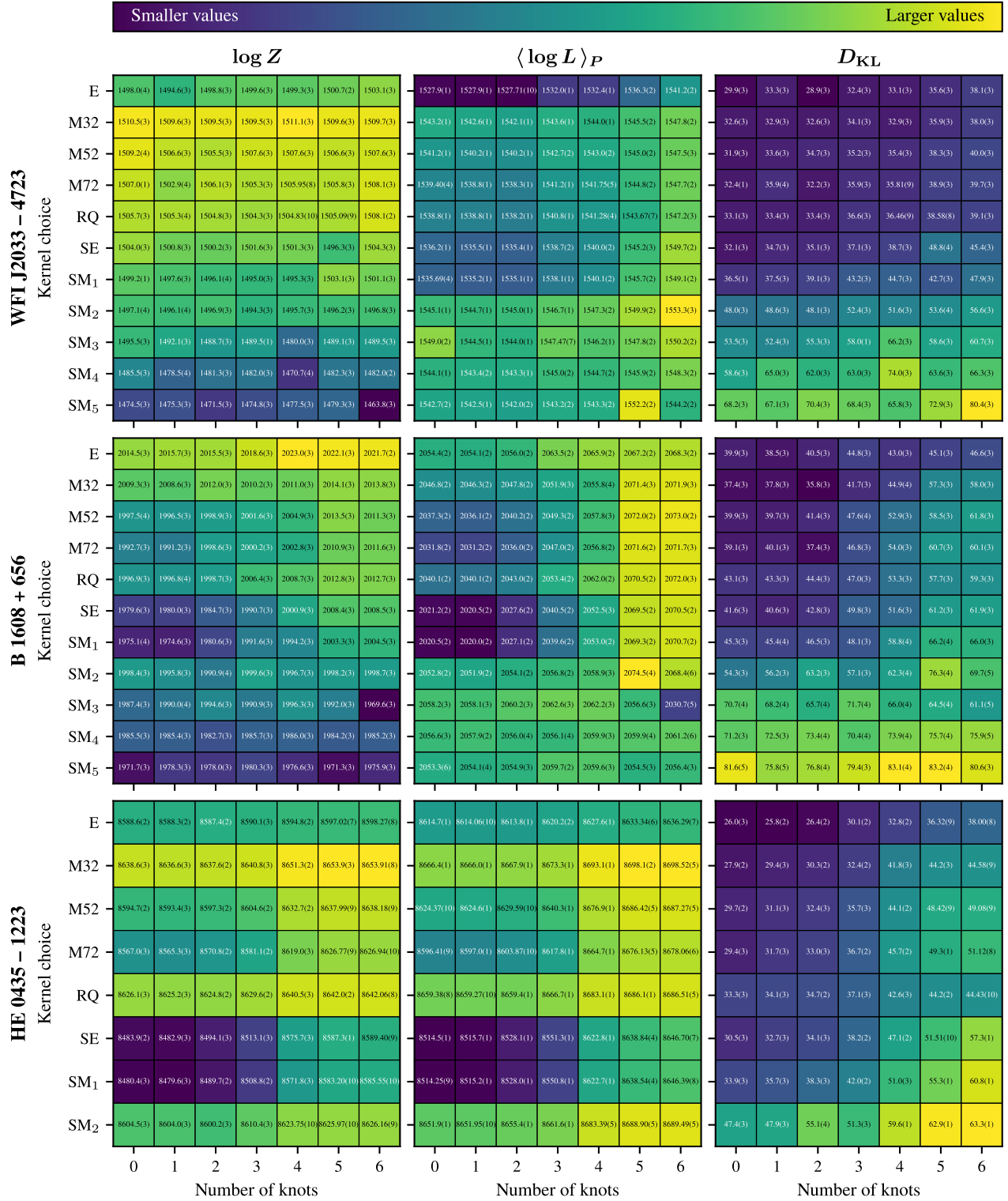


FIG. 1. Evidence $\log Z$, posterior-averaged log-likelihood $\langle \log L \rangle_P$ and KL-divergence D_{KL} for the three quasar light curve datasets. The three quantities are related by Occam's razor equation, $\log Z = \langle \log L \rangle_P - D_{KL}$. Each colored square in a subplot corresponds to a GP fit, performed with a given number of knots n_{fk} in the flexknot mean function and a given GP kernel. Zero and one knots correspond to a fixed and constant mean (with the constant a parameter), respectively. Note that the numbers in brackets denotes the standard deviation on the last digits. As shown by the KL-divergence, model complexity increases roughly from the top left to the bottom right corner. The goodness of fit, $\langle \log L \rangle_P$, roughly shows the opposite trend, with more complex models producing better fits. As the balance of these two, the evidence $\log Z$ shows an overall preference to the E and M32 kernels. Nonstationarity (more than one knot) is preferred for B 1608 + 656 and HE 0435-1223, whereas there is no clear indication of nonstationarity for WFI J2033-4723.

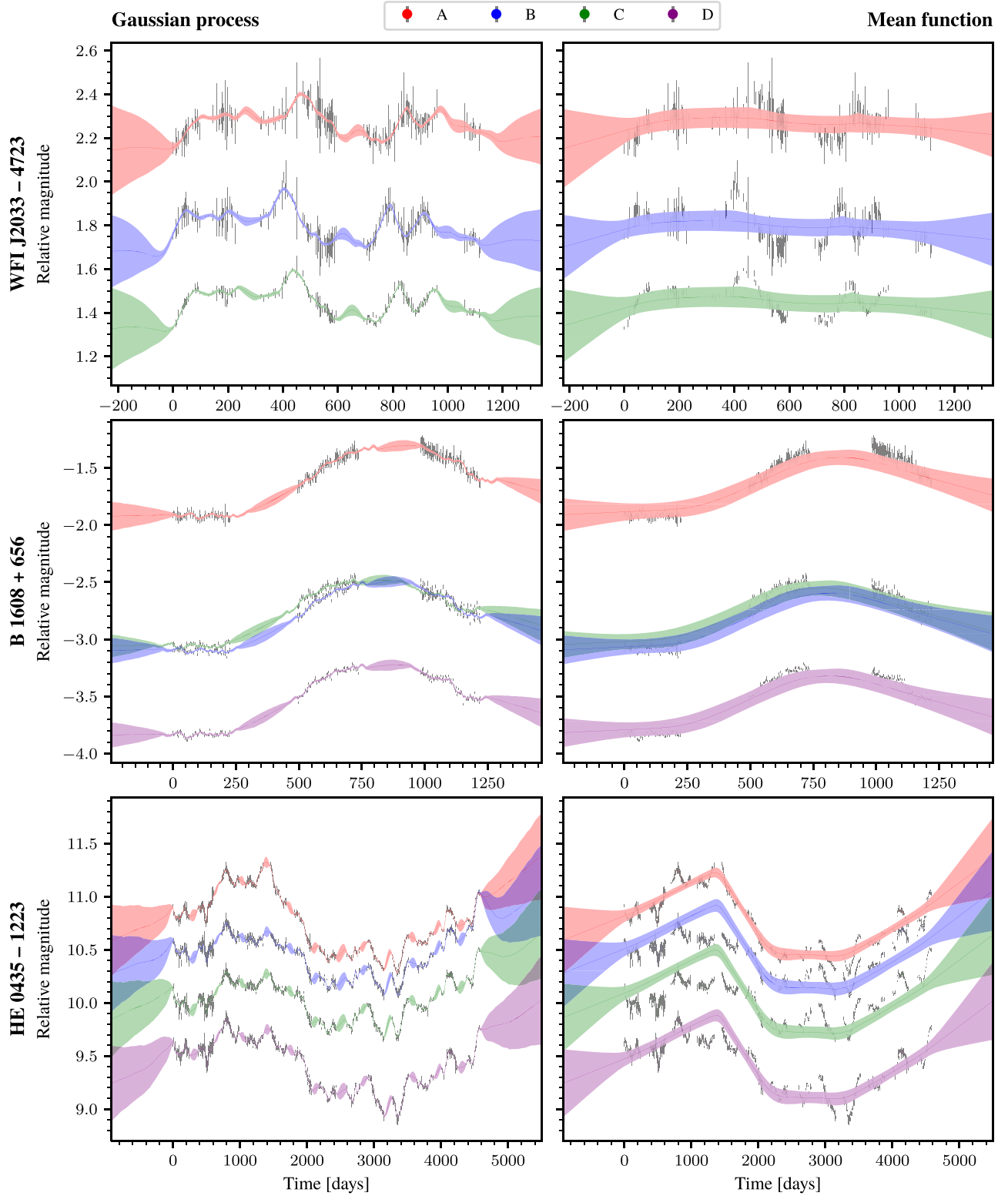


FIG. 2. GP (left column) and mean function (right column) posterior predictive distributions for the three quasar light curve datasets, each consisting of a number of light curves; A, B, C and (if existing) D. The GP is the sum of the deterministic mean function and the stochastic contribution from the GP kernel, which are fit jointly to the data. Plotting just the mean function therefore extracts the nonstationarity of the GP. The fits are marginalized over all choices in Fig. 1. Visually, the mean function of WFI J2033–4723 is consistent with a constant, whereas for B 1608 + 656 and HE 0435–1223, the mean function is clearly nonstationary. Note that the full GP on the right reverts to the mean function outside the observation window, i.e., extrapolation is dictated by the mean function. The time axis is set to zero at the first observation.

M32 kernel indicates that the underlying stochastic process of the light curve is non-Markovian, i.e., there is a memory of previous values.

To test this property more precisely, we restricted the HE 0435–1223 data to some subset $t_1, \dots, t_{n_{\text{subset}}}$ and computed the Bayes factor between the M32 and E kernel for increasing n_{subset} . The flux variability model was chosen to be a constant mean function model. This showed that the Bayes factor increases with observation duration n_{subset} and suggests that features in the light curves at longer timescales contribute significantly to the kernel posterior and that the Markov property of the E kernel fails to hold on longer timescales.

Finally, we note that the Matérn kernels can be further generalized to yield the continuous autoregressive moving average (CARMA) kernels [28]. However, as our results show, the complexity penalty is already sufficient to substantially decrease the Bayesian evidence for $\nu = \frac{5}{2}$. SM kernels are generally disfavored as well, so we do not consider this extension here. The power spectral density of a CARMA kernel is a weighted sum of Lorentzians instead of Gaussians. It remains to be seen in future work whether the heavier tails of such a power spectral density yield substantially different results.

2. Mean function inference

Considering the inferred mean function, Fig. 1 shows that the evidence continues to increase with the number of knots for B 1608 + 656 and HE 0435–1223 and therefore clearly show a preference to a larger number of knots. The mean function posterior in Fig. 2 agrees with this, showing mean functions which follow the average fluctuation of the data.

We ascribe the mean function preferences to the overall structure of the light curves on longer timescales. In particular, the light curves of B 1608 + 656 and HE 0435–1223 visibly contain variations of large amplitude on longer timescales, superimposed on smaller variations on shorter timescales. The mean function effectively separates out the long timescale variations, leaving the fluctuations around the mean to be fit by the GP kernel. Moreover, this implies that the short timescale variations are well-described by the E and M32 kernel for B 1608 + 656 and HE 0435–1223, respectively. As it could be possible for kernels with multiple length scales (RQ and SM kernels) to fit the longer scale variations with one length scale and shorter scale variations with another, the fact that single length scale kernels (E and M32 kernels) are preferred indicates strongly that the variation modeled by the flexknot is indeed a deterministic drift in the light curve.

In contrast, visual inspection of Fig. 2 shows that the mean function of WFI J2033–4723 is approximately constant and agrees with the empirical mean of each light curve. We would therefore expect that the model evidence decreases as the number of knots is increased. However, the

number of knots remains unconstrained in Fig. 1, i.e., the model posterior is approximately flat as the number of knots is increased, for the most probable kernel M32. Furthermore, $\langle \log L \rangle_P$ and D_{KL} both increase at the same rate with increasing number of knots. Overall, this implies that higher order flexknots significantly contribute to the mean function posterior, but exhibit only little variation around a constant mean.

To explain the behaviour of the evidence, we first recall that we use a sorted-uniform prior on the t -values of the flexknot. Therefore, as we add more knots, each knot can only access an interval of width $\frac{\Delta t_{\text{range}}}{n_{\text{fk}}}$ on the t -axis, on average, times a fixed interval along the y -axis. That is, the accessible parameter space per knot is a rectangle, on average. Clearly, the width of this rectangle goes to zero as the number of knots becomes large. In this sense, the flexknot is self-constraining, namely adding more parameters does not exponentially increase the accessible parameter space. To offer a more precise explanation, the total parameter space accessible to the t -positions of the knots is $\frac{\Delta t_{\text{range}}^{n_{\text{fk}}}}{n_{\text{fk}}!}$ [90,93], which goes to zero as the number of knots n_{fk} becomes large. For such models in which the prior becomes more narrow with an increasing number of parameters, it is possible that the evidence flattens out beyond a certain threshold value of the model complexity [99]. The genericity of the above argument is supported by the same phenomenon observed on flexknots in a different context [53,89]. We leave the question of whether this problem is resolved by a change of prior (for example a sparsity promoting prior) to future work.

3. Goodness of fit and model complexity

Figure 1 further shows the posterior-averaged log-likelihood $\langle \log L \rangle_P$ as a measure of the goodness of fit and the KL divergence D_{KL} as a measure of model complexity. These are related to the evidence by $\log Z = \langle \log L \rangle_P - D_{\text{KL}}$ [100]. The model complexity increases with the number of knots, as expected. We also observe an increase with the order of the Matérn kernel and the Spectral Mixture kernel. The former corresponds to increasing smoothness of the corresponding GP. That is, the model complexity roughly increases from top to bottom on our kernel axis. This ordering of kernels in terms of complexity is consistent with previous results for GPs [46].

The KL divergence of HE 0435–1223 has little effect on the model evidence. That is, the fit is primarily driven by the goodness of fit. In contrast, the KL divergence enacts a notable regularization for WFI J2033–4723 and B 1608 + 656. This behavior is expected as the dataset of HE 0435–1223 is significantly larger. More precisely, we expect the posterior-averaged log-likelihood and KL divergence to scale approximately linearly and logarithmically with the number of data points, respectively, so that the latter becomes subdominant [101].

4. Preference to the reference light curve

We note that the mean function of HE 0435–1223 appears to fit light curve A visually better than the others. Specifically, the magnitude of light curve A increases more strongly around $t = 1400$ days compared to the others, and the mean function follows this increase. This is certainly unexpected, as the penalty from the misfit to the three other light curves should outweigh the goodness of fit to light curve A. The underlying problem here is that the light curve model treats light curve A as a reference light curve (Sec. II B 1). That is, other light curves are modeled as shifted and distorted versions of light curve A. Moreover, the priors (Table I) are set by the properties of light curve A, for instance the range of the flexknot knot t -values. We confirmed that the mean function fits the reference light curve better by refitting with a different choice of reference light curve on synthetic data.

Clearly, the choice of light curve A as the reference light curve is artificial and ideally the light curve model should be permutation-invariant to the relabeling of light curves and show no affinity to a particular light curve. An *a posteriori* solution is therefore to perform multiple fits for each choice of reference light curve and subsequently average over this choice following Bayes theorem. More precisely, it is possible to treat the choice of the reference as a further parameter on which we could set a uniform prior, calculate the evidence for each choice, which in turn is obtained from the evidences of all flexknot and kernel choices, and finally marginalize over these choices. We do not proceed with this solution due to the increase in computational cost and the fact that the mean functions at least visually recover the correct trend for all light curves. Instead, the formulation of a permutation-invariant flux variability model is left to future work.

B. Stationarity

From the results described in Sec. IV A 2, it follows that the stochastic process of the WFI J2033–4723 light curves is consistent with stationarity, whereas the B 1608 + 656 and HE 0435–1223 light curves are nonstationary. Note however that the mean functions of B 1608 + 656 and HE 0435–1223 do not exhibit an overall drift to larger or smaller magnitudes but instead first increase and subsequently decrease in an oscillatory manner. It is therefore still possible that the light curve is stationary but exhibits stochastic variations on multiple length scales when measured for a much longer observation window. In this case, we would expect that the evidence values shift to favor the RQ or SM kernels, which contain multiple length scales, compared to the E and M32 kernel, only contain a single length scale. For the given data, we therefore conclude that the light curves of B 1608 + 656 and HE 0435–1223 are

indeed nonstationary within the adopted flux variability model.

C. Time delays

Our fits directly yield joint samples of the time delays Δt_{AB} , Δt_{AC} and Δt_{AD} measured relative to light curve A. This allows us to calculate samples of the other time delays, Δt_{BC} , Δt_{BD} and Δt_{CD} , by taking appropriate differences. The resulting joint posteriors with mean and standard deviation are shown in Fig. 3 and Table II.

We compare with the results in Hojjati *et al.* [22], in which the inference was performed with a constant mean function and an E (damped random walk) or M32 kernel. To ensure a direct comparison under matched modeling assumptions, we use the subset of our model grid which exactly matches their models (number of knots $n_{fk} = 1$, and either an E or M32 kernel). No additional fitting is required for this comparison as these are NS runs already included in our analysis, and are distinct from the marginalized results reported in Table II. We find that while the mean time delays change, they agree within the error, i.e., the time delay does not strongly depend on the choice of the kernel. Beyond this, our results show that time delay inference is broadly robust under the choice of mean function. Overall, we do not observe large tensions between individual model choices.

For WFI J2033–4723, the error bars of our fits increase. We rule out that this is due to marginalization over different models. Indeed, by inspecting our results purely for the E and M32 kernel with a single knot in the mean function, we validate that these time delays are even larger than the marginalized time delays. Marginalization therefore reduces the error in this case. For the E kernel, the length scale posterior shows a substantial degeneracy in parameter space. That is, the posterior is flat above a certain length scale. However, we rule out that this causes the larger time delay error bars since the length scale posterior of the M32 kernel, which has a higher evidence, is properly constrained. It is therefore possible that the difference may arise from the lack of convergence of the sampling method used in [22], which we cannot reproduce here and therefore do not further investigate. Finally, we note that WFI J2033–4723 generally has a broad time delay posterior, as discussed in Sec. III, which we ascribe to the small dataset size relative to the other two datasets.

For B 1608 + 656 and HE 0435–1223, the size of the error bars is smaller than and lies in between those of the E and M32 kernel, respectively. The error bars obtained from our time delay posteriors for only the E and M32 kernels is of comparable size to the marginalized ones. As above, it is possible that this discrepancy arises from the convergence of the sampling methods used.

It is noteworthy that we use an updated version of the HE 0435–1223 dataset, which is approximately twice as

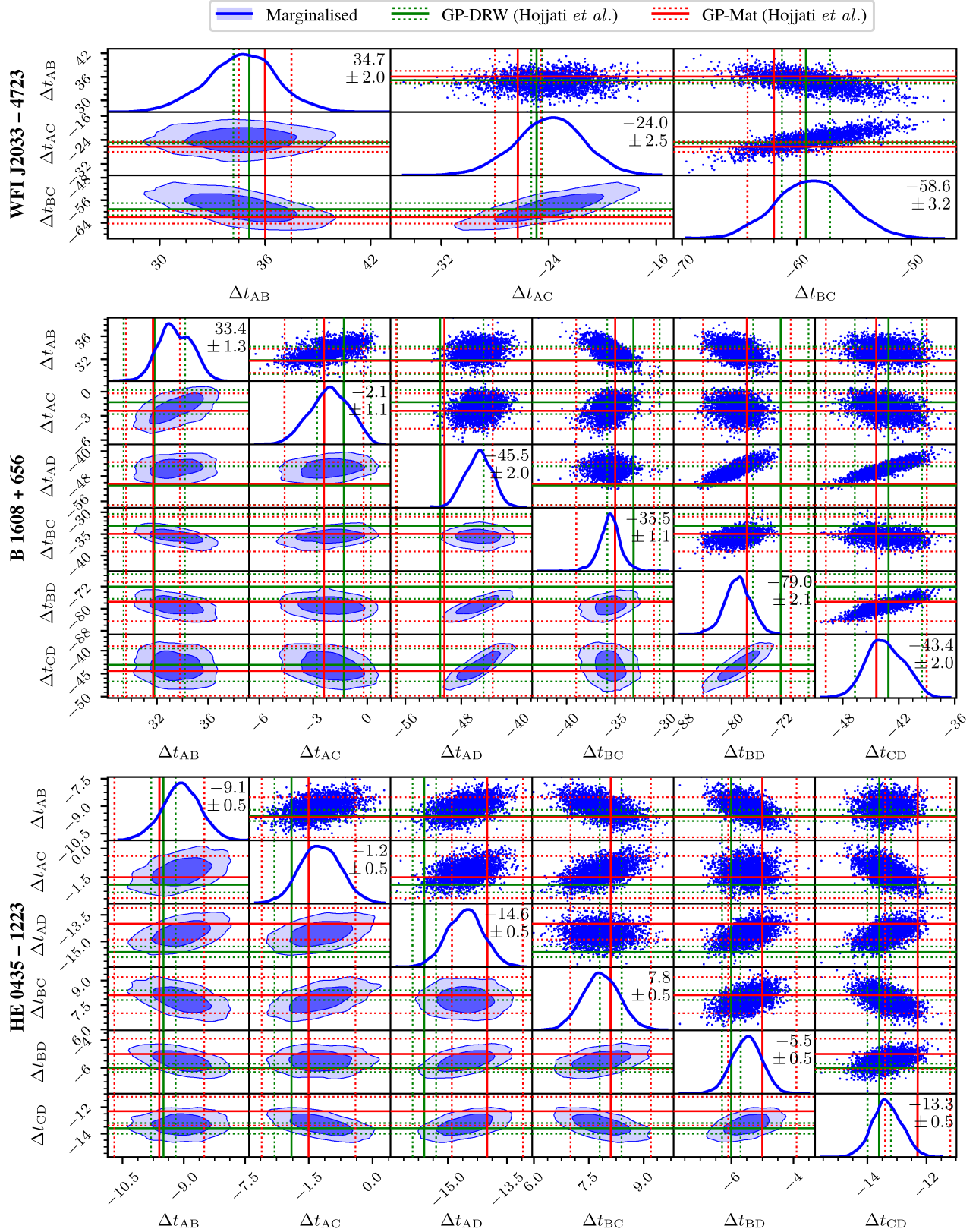


FIG. 3. Corner plots of the time delay posteriors for the three datasets. The posteriors are marginalized over all model choices in Fig. 2. For comparison, the inferred values (mean and standard deviation) from Hojjati *et al.* [22] are overplotted. All posteriors are unimodal and well-constrained, with the exception of Δt_{AB} of B 1608 + 656, which displays a weakly protruding secondary peak. The numerical means and standard deviations are also shown in Table II.

TABLE II. Comparison of the time delay posterior values for the quasars WFI J2033–4723, B 1608 + 656 and HE 0435–1223 (Fig. 3) with those from Hojjati *et al.* [22]. The values were calculated by marginalizing over the model choices in Fig. 2. The results in Hojjati *et al.* [22] were obtained for the exponential (DRW) and Matérn-3/2 kernels. All time delays are in units of days.

System and method	Δt_{AB}	Δt_{AC}	Δt_{AD}	Δt_{BC}	Δt_{BD}	Δt_{CD}
WFI J2033–4723 marginalized	34.7 ± 1.9	-24.0 ± 2.5	...	-58.7 ± 3.1	...	...
WFI J2033–4723 GP-DRW [22]	35.1 ± 0.9	-24.9 ± 0.4	...	-59.2 ± 2.1	...	...
WFI J2033–4723 GP-Mat [22]	36.0 ± 1.5	-26.3 ± 1.7	...	-62.0 ± 2.3	...	...
B1608 + 656 marginalized	33.4 ± 1.3	-2.1 ± 1.1	-45.4 ± 2.0	-35.5 ± 1.1	-78.8 ± 2.1	-43.3 ± 2.0
B1608 + 656 GP-DRW [22]	31.8 ± 2.4	-1.3 ± 1.5	-51.0 ± 6.2	-33.1 ± 2.7	-72.0 ± 4.5	-43.1 ± 3.6
B1608 + 656 GP-Mat [22]	31.7 ± 2.1	-2.4 ± 2.2	-50.4 ± 6.9	-35.0 ± 4.0	-77.5 ± 7.1	-44.4 ± 5.4
HE 0435–1223 marginalized	-9.1 ± 0.5	-1.2 ± 0.5	-14.5 ± 0.5	7.8 ± 0.5	-5.5 ± 0.5	-13.3 ± 0.5
HE 0435–1223 GP-DRW [22]	-9.5 ± 0.3	-1.9 ± 0.4	-15.6 ± 0.3	8.1 ± 0.3	-6.0 ± 0.3	-13.6 ± 0.4
HE 0435–1223 GP-Mat [22]	-9.6 ± 1.1	-1.5 ± 1.1	-14.0 ± 0.9	8.1 ± 1.1	-5.0 ± 1.1	-12.3 ± 1.1

large than the earlier one used in Hojjati *et al.* [22]. Despite this increase in dataset size, the time delays only change within the error. It is therefore possible that future measurements may not significantly influence the time delays further but instead predominantly inform the modeling of the light curve itself. However, a rigorous investigation of the influence of more data is beyond the scope of this work.

D. Scatter around the mean

For a stationary GP $y(t)$, we expect the histogram of the observed y -values to converge to a Gaussian distribution. This is an immediate consequence of the fact that stationarity, by definition, implies that the marginal distribution $p(y)$ is the same for all times t . However, an important assumption here is that the observations of the GP are representative of the true underlying distribution. This is only guaranteed if the observation time becomes sufficiently long so that ergodicity applies [102,103]. The constraint of ergodicity is only satisfiable by measuring a light curve for a long time and can therefore strictly not be resolved by any method of inference.

Nevertheless, we can subtract the inferred nonstationary mean function from the data to render our GP model stationary and inspect the resulting histogram of residuals. Historically [37,38], the observation that the flux distribution is log-normal, i.e., that magnitudes are Gaussian, provided one piece of evidence that the underlying processes in the accretion disc follow a multiplicative stochastic process instead of an additive one. Confirmation that the residuals with respect to our model possess a Gaussian distribution in units of magnitude therefore poses an important consistency test of our modeling assumptions.

Since we do not have a single mean function but instead a posterior distribution over mean functions, we can sample a mean function, calculate the residuals, plot a kernel density estimate (KDE) and repeat this for every posterior sample. Thus, we obtain a distribution over KDEs, which

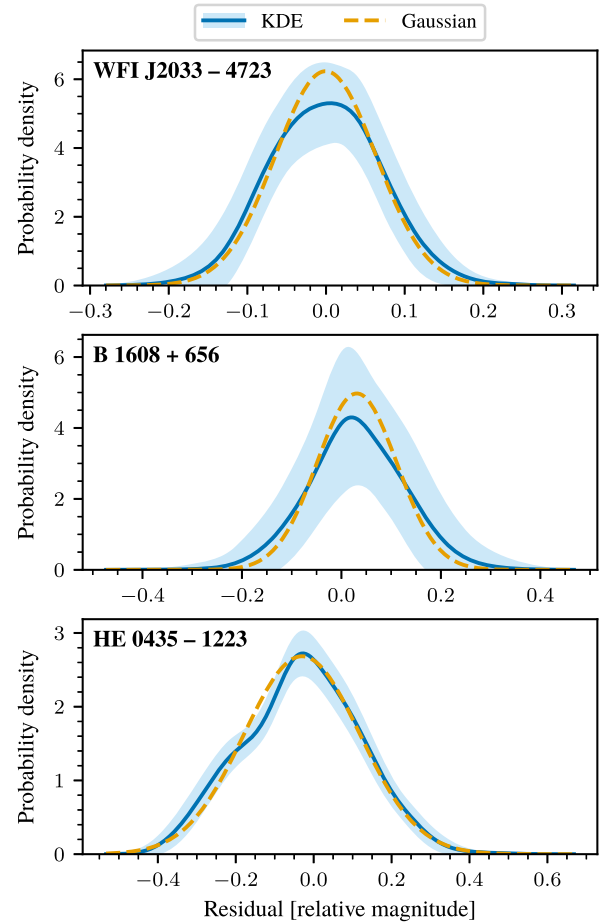


FIG. 4. Scatter of the light curve data points around the mean function of the GP fit in units of magnitude. The distribution of the residuals is expected to follow a normal distribution. For comparison, a Gaussian distribution with the same mean and standard deviation is shown as well. For WFI J2033–4723 and B 1608 + 656, the agreement in the tails is better than in the bulk. For HE 0435–1223, the observed deviation from the exact Gaussian on the left half is attributed to a misfit resulting from lack of permutation-invariance of the GP model.

we can plot by calculating the mean and error bar over KDEs.

Figure 4 shows the mean and error bar of the KDE for WFI J2033–4723, B 1608 + 656 and HE 0435–1223. For comparison, we also show a normal distribution with mean and standard deviation set by the mean and standard deviation of the residual samples. Visually, all three residual distributions are consistent with normal distributions, within the error. Deviations around the peak of the normal distribution from the mean KDE are larger, and the tails of the normal distribution are lighter. These deviations are within one standard deviation.

For HE 0435–1223, the error bars around the KDE are much narrower than for the other two datasets, which we attribute to the larger size of the dataset. Similarly, the error bars on the time delays are smallest for HE 0435–1223. While the right tail agrees well with the normal distribution, the left tail does not decay smoothly but displays an inflection point which causes larger deviation from the tail of the normal distribution. This deviation is consistent with the misfit of the mean function discussed in Sec. IV A 4, visible in Fig. 2. That is, we expect the misfit of the mean function at around $t = 1400$ days for light curves B, C and D to lead to a distortion in the distribution of the residuals.

Evidently, the uncertainty around the kernel density estimate is much larger for B 1608 + 656 compared to WFI J2033–4723. Considering Fig. 2, this is consistent with the fact that the uncertainty on the mean function is comparable to the fluctuations of the data around the mean. In contrast, the uncertainty of the mean function is much smaller for WFI J2033–4723 and HE 0435–1223, compared to the fluctuation of the data around the mean. The KDE of residuals for these two datasets therefore fluctuates less.

E. Discussion of nested sampling instabilities

We conclude the presentation of results by offering further convergence diagnostics for the nested sampling instabilities in the context of time delay inference, in addition to the convergence protocol in Sec. III B. We first stress that the evidence values of the unconverged and converged runs can differ by significantly more than five sigma, so ensuring convergence is crucial for an accurate model posterior. We also observe that unconverged runs visually appear as discontinuities in the $\log Z$, $\langle \log L \rangle_p$ and D_{KL} plots and proper convergence causes these values to change more smoothly when plotted as a function of model complexity, such as in Fig. 1. This serves as a further heuristic for assessing convergence, beyond monitoring the time delay. Finally, we have compared the marginalized time delays for both converged (obtained by our protocol) and unconverged

(obtained with default PolyChord settings) runs for the WFI J2033–4723 dataset and the resulting shift in the time delays lies well within one sigma. This robustness is due to the marginalization over models, which protects the inference from spurious unconverged values (which have low evidence). We conjecture that the mechanism behind this is that, if the dominant peak in the posterior is not found by nested sampling, the evidence automatically decreases, so that the posterior probability of that model becomes smaller and therefore contributes less to the final marginalized inference. This mechanism is only helpful if there are few unconverged runs among many properly converged runs, unlike the inference for HE 0435–1223, for which no runs had initially converged (though adoption of the convergence protocol lead to proper convergence thereafter).

Finally, we stress that the NS convergence problems fundamentally arise from the use of an insufficient number of live points, which itself is limited by the computational cost. Advances in sampling on GPUs [104,105] were shown to allow for a substantial increase in the number of live points at negligible additional cost, therefore offering a promising venue for future work.

V. CONCLUSIONS

In this work, we have introduced a robust Bayesian framework for analyzing gravitationally lensed quasar light curves by marginalizing over a flexible set of GP models. To disentangle long-term, nonstationary trends from intrinsic stochastic variability, we modeled the underlying light curve with a deterministic mean function, parametrized by a piecewise linear spline known as a flexknot, and a stationary stochastic component described by kernels from the Matérn and Spectral Mixture families. This hierarchical approach allows for the inference of cosmological time delays that are robust to assumptions about the underlying variability model.

Applying this framework to the lensed quasars WFI J2033–4723, B 1608 + 656, and HE 0435–1223, our analysis reveals distinct variability properties for each system. We find strong evidence that the light curves of both B 1608 + 656 and HE 0435–1223 are nonstationary, requiring a flexible mean function to model their long-term deterministic drifts. However, their underlying stochastic processes differ. B 1608 + 656 strongly prefers a Markovian Exponential kernel, while HE 0435–1223 is best described by a non-Markovian Matérn- $\frac{3}{2}$ kernel. In contrast, the WFI J2033–4723 light curves are consistent with stationarity, yet their stochastic component also shows a strong preference for the non-Markovian Matérn- $\frac{3}{2}$ kernel. Our analysis also uncovered a model artifact, wherein the fit shows a preferential dependence on the arbitrarily chosen reference light curve.

The time delays inferred from our marginalized posteriors are consistent with previous findings, demonstrating the robustness of these measurements, but with the advantage of systematically accounting for uncertainty in the choice of the GP kernel and mean function. Furthermore, by subtracting the inferred nonstationary drift, we show that the residuals of the light curve data are consistent with a Gaussian distribution, thereby validating a core assumption of the GP model.

Methodologically, we highlight the critical importance of rigorous convergence testing for the nested sampling algorithm to ensure accurate Bayesian evidence calculation. We note, however, that the process of marginalizing over a wide range of models provides a degree of robustness to the final time delay inference, mitigating the impact of unconverged results from any single model.

Looking forward, a natural extension of this work is to formulate a permutation-invariant GP model to eliminate the artifact associated with the reference light curve. Another promising avenue is the extension to multiwavelength data, where observations that trace bulk flows in the accretion disc could provide stronger physical constraints on the mean function, improving both the model's fidelity and its extrapolative power. We envision that this work lays a solid foundation for a more comprehensive and fully Bayesian approach to time delay inference, which, when paired with advances in GP computation such as sparse GPs, will pave the way for highly robust analyses of quasar light curves.

ACKNOWLEDGMENTS

N. K. gratefully acknowledges Andrew Fabian for discussions on stationarity in quasar light curves, Eric V. Linder and Alex G. Kim for assistance in reproducing GP inference results, Christopher D. Fassnacht for providing the WFI J2033–4723 data and Gábor Csányi and Christopher C. Lovell for helpful discussions. N. K. was supported by the Harding Distinguished Postgraduate Scholarship. This work was performed using the Cambridge Service for Data Driven Discovery (CSD3), part of which is operated by the University of Cambridge Research Computing on behalf of the STFC DiRAC HPC Facility [106]. The DiRAC component of CSD3 was funded by BEIS capital funding via STFC capital Grants No. ST/P002307/1 and No. ST/R002452/1 and STFC operations grant ST/R00689X/1. DiRAC is part of the National e-Infrastructure.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they are owned by a third party and the terms of use prevent public distribution. The data are available from the authors upon reasonable request.

APPENDIX: KERNEL CHOICES

The Matérn- ν kernel [44], indexed by ν , is defined by

$$k_\nu(\tau) = A^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} \tau \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}}{\ell} \tau \right), \quad (\text{A1})$$

which has two hyperparameters, the length scale ℓ and amplitude A . Here, the kernel is only a function of the time difference $\tau = |t - t'|$ as the kernel is stationary. K_ν is a modified Bessel function of the second kind. For half-integer $\nu = k + \frac{1}{2}$, with integer k , the GP is k -times differentiable in the mean-square sense and thus becomes more smooth as ν increases. The Matérn-1/2 kernel is also known as the E kernel and the limit $\nu \rightarrow \infty$ gives the SE kernel.

Equivalently, the GP induced by the Matérn kernel may be understood as the solution to the following stochastic differential equation [107,108]:

$$\left(\frac{d}{dt} + \frac{\sqrt{2\nu}}{\ell} \right)^{k+1} f(t) = \epsilon(t). \quad (\text{A2})$$

Here, ϵ is a white noise process with mean zero and autocorrelation $\langle \epsilon(t)\epsilon(t') \rangle = q\delta(t - t')$ driving the stochastic process f . The noise strength q is related to ℓ and A by

$$q = 2A^2\pi^{1/2} \frac{\Gamma(k+1)}{\Gamma(k+1/2)} \left(\frac{\sqrt{2\nu}}{\ell} \right)^{2k+1}. \quad (\text{A3})$$

The Matérn kernel is then the autocorrelation of the solution f in steady-state, i.e., in the $t \rightarrow \infty$ limit. This is consistent with the fact that we are not guaranteed to infer the true underlying kernel from given data if the observation time interval is short, regardless how fine the data is sampled in time, but only if it is sufficiently large compared to the length scale.

From Eq. (A2), we see that increasing the index $\nu = k + \frac{1}{2}$ introduces higher-order derivatives of f . Intuitively, this increases the smoothness of f if f is roughly thought of as the result of repeatedly integrating the white noise ϵ . Moreover, as the length scale ℓ is increased, variations in f occur over longer timescales as the weight of lower-order derivative terms decreases. Increasing the amplitude A , the strength of the white noise process driving f increases, and therefore the variance of the distribution of f increases as well, in agreement with Eq. (A1) since $k_\nu(0)$ is equal to the variance.

As a final comment, the inclusion of second- or higher-order derivatives in Eq. (A2) introduces memory into the stochastic process. This is readily seen from the counterpart in discrete time wherein the derivatives are replaced by time-shifts. In fact, for $\nu = \frac{1}{2}$, we obtain the damped random walk, which is the unique stationary Markov Gaussian process [109]. Larger values of ν are therefore non-Markov.

Another common choice for light curve modeling is the RQ kernel [9,110,111], which is derived as a length scale mixture of SE kernels,

$$k_{\text{RQ}}(\tau) = A^2 \left(1 + \frac{\tau^2}{2\alpha\ell^2} \right)^{-\alpha}, \quad (\text{A4})$$

which has three hyperparameters: the amplitude A , length scale ℓ and α , which controls the heaviness of the kernel tail. As $\alpha \rightarrow \infty$, the SE kernel is recovered.

Since the Matérn kernel only has a single length scale but quasar light curves were shown to exhibit multiple length scales or periodicity, we also consider the SM_Q kernel, indexed by a positive integer Q [112]. The motivation here is that any power spectral density can be approximated

arbitrarily well by a sufficiently large mixture of Gaussians. The Fourier transform of such a power spectral density yields the kernel

$$k_Q(\tau) = \sum_{q=1}^Q w_q e^{-2\pi^2 \tau^2 / \ell_q^2} \cos(2\pi f_q \tau), \quad (\text{A5})$$

where the weights w_q , length scales ℓ_q and frequencies f_q constitute $3Q$ hyperparameters in total. Further motivation for this kernel lies in the fact that the multiple length scales and periodicity may help to extrapolate better, thus addressing the extrapolation problem [47].

In practice, we will use the E, M32, M52, M72, RQ and SE kernels and SM_Q kernels up to $Q = 5$.

-
- [1] D. G. York, J. Adelman, J. E. Anderson, Jr., S. F. Anderson, J. Annis, N. A. Bahcall, J. Bakken, R. Barkhouser, S. Bastian, E. Berman *et al.*, The sloan digital sky survey: Technical summary, *Astron. J.* **120**, 1579 (2000).
 - [2] Y. Shen, G. T. Richards, M. A. Strauss, P. B. Hall, D. P. Schneider, S. Snedden, D. Bizyaev, H. Brewington, V. Malanushenko, E. Malanushenko *et al.*, A catalog of quasar properties from sloan digital sky survey data release 7, *Astrophys. J. Suppl. Ser.* **194**, 45 (2011).
 - [3] M.-H. Ulrich, L. Maraschi, and C. M. Urry, Variability of active galactic nuclei, *Annu. Rev. Astron. Astrophys.* **35**, 445 (1997).
 - [4] S. Vaughan, Random time series in astronomy, *Phil. Trans. R. Soc. A* **371**, 20110549 (2013).
 - [5] A. Özdönmez and M. Tekkeşinoğlu, Multi-band optical variability on diverse timescales of blazar 1e 1458.8 + 2249, *Pub. Astron. Soc. Aust.* **41**, e052 (2024).
 - [6] B. C. Kelly, J. Bechtold, and A. Siemiginowska, Are the variations in quasar optical flux driven by thermal fluctuations?, *Astrophys. J.* **698**, 895 (2009).
 - [7] P. Padovani, D. Alexander, R. Assef, B. De Marco, P. Giommi, R. Hickox, G. Richards, V. Smolčić, E. Hatziminaoglou, V. Mainieri *et al.*, Active galactic nuclei: What's in a name?, *Astron. Astrophys. Rev.* **25**, 1 (2017).
 - [8] B. M. Peterson, Variability of active galactic nuclei, in *Advanced Lectures on the Starburst-AGN Connection* (World Scientific, Singapore, 2001), pp. 3–68.
 - [9] S. Covino, F. Tobar, and A. Treves, Detecting the periodicity of highly irregularly sampled light curves with Gaussian processes: The case of SDSS j025214. 67-002813.7, *Mon. Not. R. Astron. Soc.* **513**, 2841 (2022).
 - [10] Y.-J. Chen, S. Zhai, J.-R. Liu, W.-J. Guo, Y.-C. Peng, Y.-R. Li, Y.-Y. Songsheng, P. Du, C. Hu, and J.-M. Wang, Searching for quasar candidates with periodic variations from the Zwicky Transient Facility: Results and implications, *Mon. Not. R. Astron. Soc.* **527**, 12154 (2024).
 - [11] C. L. MacLeod, Ž. Ivezić, C. Kochanek, S. Kozłowski, B. Kelly, E. Bullock, A. Kimball, B. Sesar, D. Westman, K. Brooks *et al.*, Modeling the time variability of SDSS stripe 82 quasars as a damped random walk, *Astrophys. J.* **721**, 1014 (2010).
 - [12] Y. Luo, Y. Shen, and Q. Yang, Characterization of optical light curves of extreme variability quasars over a 16-yr baseline, *Mon. Not. R. Astron. Soc.* **494**, 3686 (2020).
 - [13] Z. Stone, Y. Shen, C. J. Burke, Y.-C. Chen, Q. Yang, X. Liu, R. Gruendl, M. Adamów, F. Andrade-Oliveira, J. Annis *et al.*, Optical variability of quasars with 20-yr photometric light curves, *Mon. Not. R. Astron. Soc.* **514**, 164 (2022).
 - [14] S. Refsdal, On the possibility of determining Hubble's parameter and the masses of galaxies from the gravitational lens effect, *Mon. Not. R. Astron. Soc.* **128**, 307 (1964).
 - [15] T. Treu and P. J. Marshall, Time delay cosmography, *Astron. Astrophys. Rev.* **24**, 1 (2016).
 - [16] P. Shah, P. Lemos, and O. Lahav, A buyer's guide to the Hubble constant, *Astron. Astrophys. Rev.* **29**, 1 (2021).
 - [17] E. Abdalla, G. F. Abellán, A. Aboubrahim, A. Agnello, Ö. Akarsu, Y. Akrami, G. Alestas, D. Aloni, L. Amendola, L. A. Anchordoqui *et al.*, Cosmology intertwined: A review of the particle physics, astrophysics, and cosmology associated with the cosmological tensions and anomalies, *J. High Energy Astrophys.* **34**, 49 (2022).
 - [18] S. Suyu, M. Auger, S. Hilbert, P. Marshall, M. Tewes, T. Treu, C. Fassnacht, L. Koopmans, D. Sluse, R. Blandford *et al.*, Two accurate time-delay distances from strong lensing: Implications for cosmology, *Astrophys. J.* **766**, 70 (2013).
 - [19] K. C. Wong, S. H. Suyu, G. C. Chen, C. E. Rusu, M. Millon, D. Sluse, V. Bonvin, C. D. Fassnacht, S. Taubenberger, M. W. Auger *et al.*, H0licow–XIII. A 2.4 per cent measurement of H0 from lensed quasars: 5.3 σ tension between early- and late-universe probes, *Mon. Not. R. Astron. Soc.* **498**, 1420 (2020).

- [20] M. Millon, A. Galan, F. Courbin, T. Treu, S. Suyu, X. Ding, S. Birrer, G.-F. Chen, A. Shajib, D. Sluse *et al.*, Tdcosmo-i. an exploration of systematic uncertainties in the inference of h_0 from time-delay cosmography, *Astron. Astrophys.* **639**, A101 (2020).
- [21] S. Birrer, A. Shajib, A. Galan, M. Millon, T. Treu, A. Agnello, M. Auger, G.-F. Chen, L. Christensen, T. Collett *et al.*, Tdcosmo-IV. Hierarchical time-delay cosmography–joint inference of the Hubble constant and galaxy density profiles, *Astron. Astrophys.* **643**, A165 (2020).
- [22] A. Hojjati, A. G. Kim, and E. V. Linder, Robust strong lensing time delay estimation, *Phys. Rev. D* **87**, 123512 (2013).
- [23] A. Hojjati and E. V. Linder, Next generation strong lensing time delay estimation with Gaussian processes, *Phys. Rev. D* **90**, 123501 (2014).
- [24] M. Tewes, F. Courbin, and G. Meylan, Cosmograil: The cosmological monitoring of gravitational lenses-XI. Techniques for time delay measurement in presence of microlensing, *Astron. Astrophys.* **553**, A120 (2013).
- [25] R. Blandford and C. F. McKee, Reverberation mapping of the emission line regions of Seyfert galaxies and quasars, *Astrophys. J.* **255**, 419 (1982).
- [26] B. M. Peterson, Reverberation mapping of active galactic nuclei, *Publ. Astron. Soc. Pac.* **105**, 247 (1993).
- [27] Y. Zu, C. Kochanek, and B. M. Peterson, An alternative approach to measuring reverberation lags in active galactic nuclei, *Astrophys. J.* **735**, 80 (2011).
- [28] B. C. Kelly, A. C. Becker, M. Sobolewska, A. Siemiginowska, and P. Uttley, Flexible and scalable methods for quantifying stochastic variability in the era of massive time-domain astronomical data sets, *Astrophys. J.* **788**, 33 (2014).
- [29] H. McDougall, B. J. Pope, and T. M. Davis, LITMUS: Bayesian lag recovery in reverberation mapping with fast differentiable models, *Pub. Astron. Soc. Aust.* **43**, 1 (2025).
- [30] B. A. Vaughan and M. A. Nowak, X-ray variability coherence: How to compute it, what it means, and how it constrains models of GX 339–4 and Cygnus X-1, *Astrophys. J.* **474**, L43 (1997).
- [31] P. Uttley, I. McHardy, and I. Papadakis, Measuring the broad-band power spectra of active galactic nuclei with RXTE, *Mon. Not. R. Astron. Soc.* **332**, 231 (2002).
- [32] A. Epitropakis and I. Papadakis, Statistical properties of fourier-based time-lag estimates, *Astron. Astrophys.* **591**, A113 (2016).
- [33] J. D. Scargle, J. P. Norris, B. Jackson, and J. Chiang, Studies in astronomical time series analysis. VI. Bayesian block representations, *Astrophys. J.* **764**, 167 (2013).
- [34] Y. Tachibana, M. J. Graham, N. Kawai, S. Djorgovski, A. J. Drake, A. A. Mahabal, and D. Stern, Deep modeling of quasar variability, *Astrophys. J.* **903**, 54 (2020).
- [35] J. Fagin, J. W. Park, H. Best, J. H. Chan, K. S. Ford, M. J. Graham, V. A. Villar, S. Ho, and M. O’Dowd, Latent stochastic differential equations for modeling quasar variability and inferring black hole properties, *Astrophys. J.* **965**, 104 (2024).
- [36] S. Kozłowski, C. S. Kochanek, A. Udalski, I. Soszyński, M. Szymański, M. Kubiak, G. Pietrzyński, O. Szewczyk, K. Ulaczyk, R. Poleski *et al.*, Quantifying quasar variability as part of a general approach to classifying continuously varying sources, *Astrophys. J.* **708**, 927 (2009).
- [37] P. Uttley and I. M. McHardy, The flux-dependent amplitude of broadband noise variability in X-ray binaries and active galaxies, *Mon. Not. R. Astron. Soc.* **323**, L26 (2001).
- [38] P. Uttley, I. McHardy, and S. Vaughan, Non-linear X-ray variability in X-ray binaries and active galaxies, *Mon. Not. R. Astron. Soc.* **359**, 345 (2005).
- [39] R. F. Mushotzky, R. Edelson, W. Baumgartner, and P. Gandhi, Kepler observations of rapid optical variability in active galactic nuclei, *Astrophys. J. Lett.* **743**, L12 (2011).
- [40] Y. Zu, C. Kochanek, S. Kozłowski, and A. Udalski, Is quasar optical variability a damped random walk?, *Astrophys. J.* **765**, 106 (2013).
- [41] W. Alston, A. Fabian, D. Buisson, E. Kara, M. Parker, A. Lohfink, P. Uttley, D. Wilkins, C. Pinto, B. De Marco *et al.*, The remarkable X-ray variability of IRAS 13224–3809–I. The variability process, *Mon. Not. R. Astron. Soc.* **482**, 2088 (2019).
- [42] W. Alston, Non-stationary variability in accreting compact objects, *Mon. Not. R. Astron. Soc.* **485**, 260 (2019).
- [43] S. Bhattacharyya, R. Ghosh, R. Chatterjee, and N. Das, Blazar variability: A study of nonstationarity and the flux-RMS relation, *Astrophys. J.* **897**, 25 (2020).
- [44] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA, 2005).
- [45] S. Roberts, M. Osborne, M. Ebdon, S. Reece, N. Gibson, and S. Aigrain, Gaussian processes for time-series modelling, *Phil. Trans. R. Soc. A* **371**, 20110550 (2013).
- [46] N. Kroupa, D. Yallup, W. Handley, and M. Hobson, Kernel-, mean-, and noise-marginalized Gaussian processes for exoplanet transits and H_0 inference, *Mon. Not. R. Astron. Soc.* **528**, 1232 (2024).
- [47] N. Kroupa and W. Handley, The global structure of the time delay likelihood, [arXiv:2602.22307](https://arxiv.org/abs/2602.22307).
- [48] S. E. Said and D. A. Dickey, Testing for unit roots in autoregressive-moving average models of unknown order, *Biometrika* **71**, 599 (1984).
- [49] J. J. Warnes and B. D. Ripley, Problems with likelihood estimation of covariance functions of spatial Gaussian processes, *Biometrika* **74**, 640 (1987).
- [50] V. Lalchand and C. E. Rasmussen, Approximate inference for fully Bayesian Gaussian process regression, in *Symposium on Advances in Approximate Bayesian Inference* (PMLR, 2020), pp. 1–12, <https://proceedings.mlr.press/v118/lalchand20a.html>.
- [51] J. A. Vazquez, M. Bridges, M. Hobson, and A. Lasenby, Reconstruction of the dark energy equation of state, *J. Cosmol. Astropart. Phys.* **09** (2012) 020.
- [52] S. Hee, W. Handley, M. P. Hobson, and A. N. Lasenby, Bayesian model selection without evidences: Application to the dark energy equation-of-state, *Mon. Not. R. Astron. Soc.* **455**, 2461 (2016).
- [53] A. Ormondroyd, W. Handley, M. Hobson, and A. Lasenby, Nonparametric reconstructions of dynamical dark energy via flexknots, *Mon. Not. R. Astron. Soc.* **541**, 3388 (2025).
- [54] J. D. Scargle, Studies in astronomical time-series analysis. VII. An enquiry concerning nonlinearity, the RMS–mean

- flux relation, and lognormal flux distributions, *Astrophys. J.* **895**, 90 (2020).
- [55] H. Zhang, Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, *J. Am. Stat. Assoc.* **99**, 250 (2004).
- [56] D. Velandia, F. Bachoc, M. Bevilacqua, X. Gendre, and J.-M. Loubes, Maximum likelihood estimation for a bivariate Gaussian process under fixed domain asymptotics, *Electron. J. Stat.* **11**, 2978 (2017).
- [57] C. Vuissoz, F. Courbin, D. Sluse, G. Meylan, V. Chantry, E. Eulaers, C. Morgan, M. Eyler, C. Kochanek, J. Coles *et al.*, Cosmograil: The cosmological monitoring of gravitational lenses-VII. Time delays and the Hubble constant from WFI J2033–4723, *Astron. Astrophys.* **488**, 481 (2008).
- [58] C. Fassnacht, E. Xanthopoulos, L. Koopmans, and D. Rusin, A determination of H_0 with the class gravitational lens B1608+ 656. III. A significant improvement in the precision of the time delay measurements, *Astrophys. J.* **581**, 823 (2002).
- [59] V. Bonvin, F. Courbin, S. H. Suyu, P. Marshall, C. Rusu, D. Sluse, M. Tewes, K. Wong, T. Collett, C. Fassnacht *et al.*, H0licow–V. New cosmograil time delays of He 0435–1223: H_0 to 3.8 per cent precision from strong lensing in a flat λ CDM model, *Mon. Not. R. Astron. Soc.* **465**, 4914 (2017).
- [60] N. D. Morgan, J. A. Caldwell, P. L. Schechter, A. Dressler, E. Egami, and H.-W. Rix, WFI J2026– 4536 and WFI J2033– 4723: Two new quadruple gravitational lenses, *Astron. J.* **127**, 2617 (2004).
- [61] J. A. Blackburne, D. Pooley, S. Rappaport, and P. L. Schechter, Sizes and temperature profiles of quasar accretion disks from chromatic microlensing, *Astrophys. J.* **729**, 34 (2011).
- [62] M. Millon, F. Courbin, V. Bonvin, E. Paic, G. Meylan, M. Tewes, D. Sluse, P. Magain, J. Chan, A. Galan *et al.*, Cosmograil-XIX. Time delays in 18 strongly lensed quasars from 15 years of optical monitoring, *Astron. Astrophys.* **640**, A105 (2020).
- [63] S. T. Myers, C. Fassnacht, S. Djorgovski, R. Blandford, K. Matthews, G. Neugebauer, T. Pearson, A. Readhead, J. Smith, D. Thompson *et al.*, 1608 + 656: A quadruple-lens system found in the class gravitational lens survey, *Astrophys. J.* **447**, L5 (1995).
- [64] L. Koopmans, T. Treu, C. Fassnacht, R. Blandford, and G. Surpi, The Hubble constant from the gravitational lens B1608+ 656, *Astrophys. J.* **599**, 70 (2003).
- [65] C. Kochanek and N. Dalal, Tests for substructure in gravitational lenses, *Astrophys. J.* **610**, 69 (2004).
- [66] S. Suyu, P. Marshall, M. Auger, S. Hilbert, R. Blandford, L. Koopmans, C. Fassnacht, and T. Treu, Dissecting the gravitational lens B1608+ 656. II. Precision measurements of the Hubble constant, spatial curvature, and the dark energy equation of state, *Astrophys. J.* **711**, 201 (2010).
- [67] L. Wisotzki, P. L. Schechter, H. V. Bradt, J. Heinmueller, and D. Reimers, He 0435–1223: A wide separation quadruple QSO and gravitational lens, *Astron. Astrophys.* **395**, 17 (2002).
- [68] S. H. Suyu, V. Bonvin, F. Courbin, C. D. Fassnacht, C. E. Rusu, D. Sluse, T. Treu, K. Wong, M. W. Auger, X. Ding *et al.*, H0licow–I. H_0 lenses in cosmograil’s wellspring: Program overview, *Mon. Not. R. Astron. Soc.* **468**, 2590 (2017).
- [69] D. Sluse, A. Sonnenfeld, N. Rumbaugh, C. E. Rusu, C. D. Fassnacht, T. Treu, S. H. Suyu, K. Wong, M. W. Auger, V. Bonvin *et al.*, H0licow–II. Spectroscopic survey and galaxy-group identification of the strong gravitational lens system He 0435– 1223, *Mon. Not. R. Astron. Soc.* **470**, 4838 (2017).
- [70] C. E. Rusu, C. D. Fassnacht, D. Sluse, S. Hilbert, K. C. Wong, K.-H. Huang, S. H. Suyu, T. E. Collett, P. J. Marshall, T. Treu *et al.*, H0licow–III. Quantifying the effect of mass along the line of sight to the gravitational lens He 0435– 1223 through weighted galaxy counts, *Mon. Not. R. Astron. Soc.* **467**, 4220 (2017).
- [71] K. C. Wong, S. H. Suyu, M. W. Auger, V. Bonvin, F. Courbin, C. D. Fassnacht, A. Halkola, C. E. Rusu, D. Sluse, A. Sonnenfeld *et al.*, H0licow–IV. Lens mass model of He 0435– 1223 and blind measurement of its time-delay distance for cosmology, *Mon. Not. R. Astron. Soc.* **465**, 4895 (2017).
- [72] V. Bonvin, F. Courbin, S. H. Suyu, P. Marshall, C. Rusu, D. Sluse, M. Tewes, K. Wong, T. Collett, C. Fassnacht *et al.*, H0licow–V. New cosmograil time delays of He 0435–1223: H_0 to 3.8 per cent precision from strong lensing in a flat λ cdm model, *Mon. Not. R. Astron. Soc.* **465**, 4914 (2017).
- [73] F. Courbin, V. Chantry, Y. Revaz, D. Sluse, C. Faure, M. Tewes, E. Eulaers, M. Koleva, I. Asfandiyarov, S. Dye *et al.*, Cosmograil: The cosmological monitoring of gravitational lenses-IX. Time delays, lens dynamics and baryonic fraction in He 0435–1223, *Astron. Astrophys.* **536**, A53 (2011).
- [74] M. Millon, M. Tewes, V. Bonvin, B. Lengen, and F. Courbin, Pycs3: A Python toolbox for time-delay measurements in lensed quasars, *J. Open Source Software* **5**, 2654 (2020).
- [75] G. Wahba, Improper priors, spline smoothing and the problem of guarding against model errors in regression, *J. R. Stat. Soc. Ser. B* **40**, 364 (1978).
- [76] I. Schoenberg, Spline interpolation and best quadrature formulae, in *IJ Schoenberg Selected Papers* (Springer, New York, 1988), pp. 391–396.
- [77] D. J. C. MacKay, Introduction to Gaussian processes, in *Neural Networks and Machine Learning*, edited by C. M. Bishop (Springer, Berlin, Heidelberg, 1998), pp. 133–166.
- [78] W. J. Handley, A. N. Lasenby, H. V. Peiris, and M. P. Hobson, Bayesian inflationary reconstructions from Planck 2018 data, *Phys. Rev. D* **100**, 103511 (2019).
- [79] J. A. Vazquez, M. Bridges, M. Hobson, and A. Lasenby, Model selection applied to reconstruction of the primordial power spectrum, *J. Cosmol. Astropart. Phys.* **06** (2012) 006.
- [80] G. Aslanyan, L. C. Price, K. N. Abazajian, and R. Easther, The Knotted sky I: Planck constraints on the primordial power spectrum, *J. Cosmol. Astropart. Phys.* **08** (2014) 052.
- [81] F. Finelli, M. Bucher, A. Achúcarro, M. Ballardini, N. Bartolo, D. Baumann, S. Clesse, J. Errard, W. Handley,

- M. Hindmarsh *et al.*, Exploring cosmic origins with core: Inflation, *J. Cosmol. Astropart. Phys.* **04** (2018) 016.
- [82] P. A. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, R. Barreiro *et al.*, Planck 2013 results. XXII. Constraints on inflation, *Astron. Astrophys.* **571**, A22 (2014).
- [83] P. Ade, N. Aghanim, M. Arnaud, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. Banday, R. Barreiro *et al.*, Planck 2015 results-XX. Constraints on inflation, *Astron. Astrophys.* **594**, A20 (2016).
- [84] M. Millea and F. Bouchet, Cosmic microwave background constraints in light of priors over reionization histories, *Astron. Astrophys.* **617**, A96 (2018).
- [85] S. Heimersheim, N. S. Sartorio, A. Fialkov, and D. R. Lorimer, What it takes to measure reionization with fast radio bursts, *Astrophys. J.* **933**, 57 (2022).
- [86] M. Olamaie, M. P. Hobson, F. Feroz, K. J. Grainge, A. Lasenby, Y. C. Perrott, C. Rumsey, and R. D. Saunders, Free-form modelling of galaxy clusters: A Bayesian and data-driven approach, *Mon. Not. R. Astron. Soc.* **481**, 3853 (2018).
- [87] S. Heimersheim, L. Rønneberg, H. Linton, F. Pagani, and A. Fialkov, Flexknot and Gaussian process for 21 cm global signal analysis and foreground separation, *Mon. Not. R. Astron. Soc.* **527**, 11404 (2024).
- [88] E. Shen, D. Anstey, E. de Lera Acedo, and A. Fialkov, Flexknot as a generalized model of the sky-averaged 21-cm signal at $z=6-30$ in the presence of systematics, *Mon. Not. R. Astron. Soc.* **529**, 1642 (2024).
- [89] A. Ormondroyd, W. Handley, M. Hobson, and A. Lasenby, Comparison of dynamical dark energy with Λ CDM in light of DESI DR2, *arXiv:2503.17342*.
- [90] R. Busicchio, E. Roebber, J. M. Goldstein, and C. J. Moore, Label switching problem in Bayesian analysis for gravitational wave astronomy, *Phys. Rev. D* **100**, 084041 (2019).
- [91] W. J. Handley, M. P. Hobson, and A. N. Lasenby, PolyChord: Next-generation nested sampling, *Mon. Not. R. Astron. Soc.* **453**, 4384 (2015).
- [92] F. Simpson, V. Lachand, and C. E. Rasmussen, Marginalised Gaussian processes with nested sampling, *Adv. Neural Inf. Process. Syst.* **34**, 13613 (2021); *arXiv:2010.16344*.
- [93] W. Handley, M. Hobson, and A. Lasenby, PolyChord: Nested sampling for cosmology, *Mon. Not. R. Astron. Soc.* **450**, L61 (2015).
- [94] A. Petrosyan and W. Handley, SuperNest: Accelerated nested sampling applied to astrophysics and cosmology, *Phys. Sci. Forum* **5**, 51 (2022).
- [95] P. Lemos, N. Malkin, W. Handley, Y. Bengio, Y. Hezaveh, and L. Perreault-Levasseur, Improving gradient-guided nested sampling for posterior inference, in *International Conference on Machine Learning* (PMLR, 2024), pp. 27230–27253, <https://proceedings.mlr.press/v235/lemos24a.html>.
- [96] W. Handley, anesthetic: Nested sampling visualisation, *J. Open Source Software* **4**, 1414 (2019).
- [97] D. Foreman-Mackey, dfm/tinygp: The tiniest of Gaussian Process libraries (2023).
- [98] J. Skilling, Nested sampling for general Bayesian computation, *Bayesian Anal.* **1**, 833 (2006).
- [99] C. Rasmussen and Z. Ghahramani, Occam’s razor, *Adv. Neural Inf. Process. Syst.* **13** (2000), https://papers.nips.cc/paper_files/paper/2000/hash/0950ca92a4dcf426067cfd2246bb5ff3-Abstract.html.
- [100] L. T. Hergt, W. J. Handley, M. P. Hobson, and A. N. Lasenby, Bayesian evidence for the tensor-to-scalar ratio r and neutrino masses m_ν : Effects of uniform versus logarithmic priors, *Phys. Rev. D* **103**, 123511 (2021).
- [101] R. Trotta, Bayes in the sky: Bayesian inference and model selection in cosmology, *Contemp. Phys.* **49**, 71 (2008).
- [102] G. A. Pavliotis, *Stochastic Processes and Applications*, Texts in Applied Mathematics (Springer, New York, NY, 2014), 10.1007/978-1-4939-1323-7.
- [103] G. Grimmett and D. Stirzaker, *Probability and Random Processes* (Oxford University Press, New York, 2020).
- [104] D. Yallup, N. Kroupa, and W. Handley, Nested slice sampling, in *Frontiers in Probabilistic Inference: Learning meets Sampling* (2025), <https://openreview.net/forum?id=ekbkMSuPo4>.
- [105] D. Yallup, N. Kroupa, and W. Handley, Nested slice sampling: Vectorized nested sampling for GPU-accelerated inference, *arXiv:2601.23252*.
- [106] <http://www.dirac.ac.uk>
- [107] J. Hartikainen and S. Särkkä, Kalman filtering and smoothing solutions to temporal Gaussian process regression models, in *2010 IEEE International Workshop on Machine Learning for Signal Processing* (IEEE, 2010), pp. 379–384.
- [108] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations* (Cambridge University Press, Cambridge, England, 2019), Vol. 10.
- [109] J. L. Doob, The Brownian movement and stochastic equations, *Ann. Math.* **43**, 351 (1942).
- [110] D. Wilkins, Low-frequency X-ray timing with Gaussian processes and reverberation in the radio-loud AGN 3C 120, *Mon. Not. R. Astron. Soc.* **489**, 1957 (2019).
- [111] R.-R. Griffiths, J. Jiang, D. J. Buisson, D. Wilkins, L. C. Gallo, A. Ingram, D. Grupe, E. Kara, M. L. Parker, W. Alston *et al.*, Modeling the multiwavelength variability of MRK 335 using Gaussian processes, *Astrophys. J.* **914**, 144 (2021).
- [112] A. Wilson and R. Adams, Gaussian process kernels for pattern discovery and extrapolation, in *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 28, edited by S. Dasgupta and D. McAllester (PMLR, Atlanta, Georgia, USA, 2013), pp. 1067–1075.