

Kernel regression estimates of time delays between gravitationally lensed fluxes

Sultanah AL Otaibi¹, Peter Tiño^{1*}, Juan C. Cuevas-Tello²,
Ilya Mandel³ and Somak Raychaudhury^{4,5,3}

¹ *School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK*

² *Engineering Faculty, Autonomous University of San Luis Potosí, México*

³ *School of Physics and Astronomy, University of Birmingham, Birmingham, B15 2TT, UK*

⁴ *Inter-University Centre for Astronomy & Astrophysics, Pune 411007, India*

⁵ *Department of Physics, Presidency University, Kolkata 700073, India*

29 February 2016

ABSTRACT

Strongly lensed variable quasars can serve as precise cosmological probes, provided that time delays between the image fluxes can be accurately measured. A number of methods have been proposed to address this problem. In this paper, we explore in detail a new approach based on kernel regression estimates, which is able to estimate a single time delay given several datasets for the same quasar. We develop realistic artificial data sets in order to carry out controlled experiments to test of performance of this new approach. We also test our method on real data from strongly lensed quasar Q0957+561 and compare our estimates against existing results.

Key words: Gravitational lensing, quasars, Q0957+561A,B, Time series, kernel regression, statistical analysis, Multiobjective algorithms

1 INTRODUCTION

Time delays between images of strongly-lensed distant variable sources can serve as a valuable tool for cosmography, providing an alternative to other tools, such as CMB measurements and distance measures based on standard candles (e.g., Refsdal 1964; Linder 2011; Suyu et al. 2013; Greene et al. 2013; Treu et al. 2013). Actively studied strong quasars with time-delay measurements include RXJ1131-1231 (e.g., Suyu et al. 2013; Tewes et al. 2013) and B1608+656 (e.g., Fassnacht et al. 2002; Suyu et al. 2010; Greene et al. 2013); Q0957+561 (e.g., Oguri 2007; Fadely et al. 2010; Hainline et al. 2012); SDSS J1650+4251 and HE 0435-1223 (e.g., Kochanek et al. 2006; Vuissoz et al. 2007; Courbin et al. 2011); SDSS J1029+2623 (e.g., Fohlmeister et al. 2013); and SDSS J1001+5027 (e.g., Rathna Kumar et al. 2013). These have been used to infer Hubble constant measurements with competitive accuracies.

However, time delays are difficult to measure because of the unknown intrinsic source variability, the

limited observational cadence, and the measurement noise. A number of methods have been developed to accurately estimate time delays. These include the dispersion spectra method (Pelt et al. 1998; Vuissoz et al. 2007; Courbin et al. 2011); the polynomial and curve-fitting methods (Vuissoz et al. 2008; Eulaers et al. 2013); the free-knot spline, variability of regression differences (based on Gaussian process regression), and dispersion minimisation (Tewes, Courbin & Meylan 2013); Gaussian process modelling (e.g., Hojjati, Kim & Linder 2013); and the combined method based on the PRH approach (Hirv, Olsperg & Pelt 2011). However, this remains an active area of research, especially in view of the upcoming surveys such as LSST which will provide unprecedented data sets with strongly lensed distant quasars (e.g., Treu et al. 2013) (and the recent mock data challenge Dobler et al. 2015; Liao et al. 2015).

Some of the authors of the present work previously proposed a kernel-based method with variable width (K-V) for time delay estimation (Cuevas-Tello, Tiño & Raychaudhury 2006). This was combined with an evolutionary algorithm (EA) for parameter optimisation (Cuevas-Tello et al. 2009). However, the computational time

* P.Tino@cs.bham.ac.uk

complexity of EA method is $O(n^6)$ (Cuevas-Tello 2007). This restriction makes it inadequate for handling long time series, e.g. Schild & Thomson (1997) data¹. This complexity is due to matrix inversion in kernel-based methods for weights estimation. Automatic methods for time delay estimation have been proposed to speed up algorithms in order to deal with long time series, based on Artificial Neural Networks (Gonzalez-Grimaldo & Cuevas-Tello 2008); these can be parallelized (Cuevas-Tello et al. 2012). Alternatively, a simple hill-climbing optimisation has been proposed (Cuevas-Tello & Perez-Gonzalez 2011).

The main contribution of the present paper is a new probabilistic method that is efficient, robust to observational gaps, capable of directly incorporating measured noise levels reported for individual flux measurement, and able to estimate *a single time delay given several datasets* for the same quasar. We also carefully construct synthetic data sets within the framework of multiobjective optimization to reproduce realistic flux variability, observational gaps, and noise levels. This allows us to test our proposed kernel regression estimate method on synthetic as well as real data, in order to measure the bias and variance of the method.

The paper is organised as follows. Section §2 presents the Nadaraya-Watson estimator with known noise levels (henceforth *NWE*) and in §3 we extend it to a linear noise model with unknown noise (henceforth *NWE++*). In section §4, we discuss two previously proposed delay methods, cross correlation and dispersion spectra, to compare to the new approach. Section §5 shows the real datasets studied in this paper, and presents our procedure for generating synthetic data. The results are given in §6, and we conclude with a summary in §7.

2 THE MODEL

We consider a distant point source (e.g., a quasar) with two strongly lensed images², referred to as *A* and *B*, and one or more time series of flux measurements, possibly taken by different instruments and/or at different frequencies. The entire data collection $D = \{D^1, D^2, \dots, D^L\}$ consists of L data sets D^ℓ , $\ell \in [1, L]$, each corresponding to a sequence of measurements taken with a given instrument and at a given frequency. Data sets D^ℓ consist of flux measurements of both images, y_A^ℓ and y_B^ℓ , taken at a non-uniform sequence of N^ℓ observational times times $t_1^\ell, t_2^\ell, \dots, t_{N^\ell}^\ell$.

Formally, each set D^ℓ contains N^ℓ 3-tuples

$$(t_k^\ell, y_{A,k}^\ell, y_{B,k}^\ell), k = 1, 2, \dots, N^\ell,$$

$$D^\ell = \{(t_1^\ell, y_{A,1}^\ell, y_{B,1}^\ell), (t_2^\ell, y_{A,2}^\ell, y_{B,2}^\ell), \dots, (t_{N^\ell}^\ell, y_{A,N^\ell}^\ell, y_{B,N^\ell}^\ell)\},$$

where $y_{A,k}^\ell$ and $y_{B,k}^\ell$ denote the observed fluxes of image *A* and *B*, respectively, in D^ℓ at time t_k^ℓ . We also assume

that the standard errors $\sigma_{A,k}^\ell$ and $\sigma_{B,k}^\ell$ are known for each observation $y_{A,k}^\ell$ and $y_{B,k}^\ell$, respectively.

The fluxes corresponding to the two images *A* and *B* are collected in sets

$$D_A^\ell = \{(t_1^\ell, y_{A,1}^\ell), (t_2^\ell, y_{A,2}^\ell), \dots, (t_{N^\ell}^\ell, y_{A,N^\ell}^\ell)\}$$

and

$$D_B^\ell = \{(t_1^\ell, y_{B,1}^\ell), (t_2^\ell, y_{B,2}^\ell), \dots, (t_{N^\ell}^\ell, y_{B,N^\ell}^\ell)\}.$$

For observations at frequencies above a few tens of MHz, dispersion yields sub-hour arrival time differences, and is not significant relative to typical time-delay measurement accuracy. We therefore assume that the time delay between gravitationally lensed fluxes does not depend on the wavelength at which the observations are taken. We also assume stationarity of the lensing object (e.g., a galaxy) in the sense that the delay does not change in time; in particular, we ignore micro-lensing contributions.

2.1 Nadaraya-Watson Estimator with Known Noise Levels (NWE)

Given a delay Δ , we seek to find a probabilistic model $p(D|\Delta)$ that explains³ D . Assuming independence of the observation sets D^ℓ , we obtain

$$p(D|\Delta) = \prod_{\ell=1}^L p(D^\ell|\Delta).$$

Assuming independent observations at distinct measurement times, we get

$$p(D^\ell|\Delta) = \prod_{k=1}^{N^\ell} p(y_{A,k}^\ell, y_{B,k}^\ell | t_k^\ell, \Delta)$$

and further assumption of independence of measurement noise in images *A* and *B* leads to

$$p(y_{A,k}^\ell, y_{B,k}^\ell | t_k^\ell, \Delta) = p_A(y_{A,k}^\ell | t_k^\ell, \Delta) p_B(y_{B,k}^\ell | t_k^\ell, \Delta).$$

2.1.1 Modelling the source using image *A*

It is typically assumed that the measurement uncertainties on fluxes D_A^ℓ and D_B^ℓ are normally distributed, with zero mean Gaussian noise of known standard deviation $\sigma_{A,k}^\ell$ and $\sigma_{B,k}^\ell$ associated with noisy observations $y_{A,k}^\ell$ and $y_{B,k}^\ell$, respectively. We model the mean of image *A* using Nadaraya-Watson kernel regression (Nadaraya 1964), (Watson 1964),

$$f_A^\ell(t) = \sum_{k=1}^{N^\ell} y_{A,k}^\ell \frac{K(t, t_k^\ell; h^\ell)}{\sum_{j=1}^{N^\ell} K(t, t_j^\ell; h^\ell)}, \quad (1)$$

³ We slightly abuse mathematical notation as we are actually building conditional models of flux values, given the observation times.

¹ <http://cfa-www.harvard.edu/~rschild/fulldata2.txt>

² generalisation to four images is straightforward

where $f^\ell(t)$ is the predicted flux at time t and $K(t, t_j^\ell; h^\ell)$ is a kernel positioned at t_j^ℓ with bandwidth parameter h^ℓ . We use the Gaussian kernel

$$K(t, t_k; h) = \exp \left\{ -\frac{(t - t_k)^2}{\kappa^2(t_k)} \right\},$$

where the kernel scale $\kappa(t_k)$ at position t_k is defined as the distance spanned by the h neighbours (to the left and to the right) of t_k , i.e. $\kappa(t_k) = t_{k+h} - t_{k-h}$. This approach to modelling the noise should work when the autocorrelation length of the observed flux is much longer than any gaps in the data during which the flux is modelled via the Nadaraya-Watson kernel regression estimator. If the autocorrelation length of the observed flux, which can be estimated from a time interval when the observations are relatively closely spaced, is comparable to or larger than a data gap, this approach (or any other approach that does not incorporate a physically accurate flux model) cannot be trusted.

To respect the nature of gravitationally lensed data, we impose that the mean model for image B follows exactly that for image A, up to scaling by a constant⁴ $M > 0$ and time shift by Δ :

$$f_B^\ell(t; \Delta) = M f_A^\ell(t - \Delta).$$

Since the shift Δ plays no role in modelling image A, we write

$$p(y_{A,k}^\ell, y_{B,k}^\ell | t_k^\ell, \Delta) = p_A(y_{A,k}^\ell | t_k^\ell) p_B(y_{B,k}^\ell | t_k^\ell, \Delta), \quad (2)$$

where

$$p_A(y_{A,k}^\ell | t_k^\ell) = \frac{1}{\sqrt{2\pi} \sigma_{A,k}^\ell} \exp \left\{ -\frac{1}{2} \frac{(y_{A,k}^\ell - f_A^\ell(t_k^\ell))^2}{(\sigma_{A,k}^\ell)^2} \right\} \quad (3)$$

and

$$p_B(y_{B,k}^\ell | t_k^\ell, \Delta) = \frac{1}{\sqrt{2\pi} \sigma_{B,k}^\ell} \cdot \exp \left\{ -\frac{1}{2} \frac{(y_{B,k}^\ell - M f_A^\ell(t_k^\ell - \Delta))^2}{(\sigma_{B,k}^\ell)^2} \right\}. \quad (4)$$

Note that given Δ , the only free parameter of $p(y_{A,k}^\ell, y_{B,k}^\ell | t_k^\ell, \Delta)$ is the kernel width parameter h^ℓ in the formulation of the mean model (1).

Ignoring constant terms and scaling, the negative log likelihood, $-\log p(D^\ell | \Delta)$, forms the approximation error for the set D^ℓ ,

$$E_A^\ell(h^\ell; \Delta) = \sum_{k=1}^{N^\ell} \left\{ \frac{(y_{A,k}^\ell - f_A^\ell(t_k^\ell))^2}{(\sigma_{A,k}^\ell)^2} + \frac{(y_{B,k}^\ell - M f_A^\ell(t_k^\ell - \Delta))^2}{(\sigma_{B,k}^\ell)^2} \right\}. \quad (5)$$

Writing down the negative log likelihood for the whole data, $-\log p(D | \Delta)$, and ignoring scaling and constant terms leads to the total approximation error

$$E_A(\mathbf{h}; \Delta) = \sum_{\ell=1}^L E_A^\ell(h^\ell; \Delta),$$

⁴ assumed known, or easily estimated in a preprocessing stage using the means of the fluxes in D_A^ℓ and D_B^ℓ

where $\mathbf{h} = (h^1, h^2, \dots, h^L)$ is a vector that collects kernel width parameters for all datasets D^1, D^2, \dots, D^L in D .

2.1.2 Modelling the source using image B

One can, of course, start by building a mean flux model $f_B^\ell(t)$ for image B via Nadaraya-Watson kernel regression,

$$f_B^\ell(t) = \sum_{k=1}^{N^\ell} y_{B,k}^\ell \frac{K(t, t_k^\ell; h^\ell)}{\sum_{j=1}^{N^\ell} K(t, t_j^\ell; h^\ell)}, \quad (6)$$

imposing that the mean model of image A is

$$f_A^\ell(t; \Delta) = \frac{1}{M} f_B^\ell(t + \Delta).$$

Crucially, since both images A and B come from the same source, we require that the kernel width h^ℓ for the mean models $f_A^\ell(t)$ and $f_B^\ell(t)$ (and hence for $f_A^\ell(t; \Delta)$ and $f_B^\ell(t; \Delta)$ as well) be the same for the whole dataset D^ℓ .

Using the same reasoning as in section §2.1.1, we obtain an approximation error for the set D^ℓ :

$$E_B^\ell(h^\ell; \Delta) = \sum_{k=1}^{N^\ell} \left\{ \frac{(y_{A,k}^\ell - \frac{1}{M} f_B^\ell(t_k^\ell + \Delta))^2}{(\sigma_{A,k}^\ell)^2} + \frac{(y_{B,k}^\ell - f_B^\ell(t_k^\ell))^2}{(\sigma_{B,k}^\ell)^2} \right\}$$

leading to the total approximation error

$$E_B(\mathbf{h}; \Delta) = \sum_{\ell=1}^L E_B^\ell(h^\ell; \Delta).$$

2.1.3 Estimating the Unique Time Delay across D

Since there is no a-priori reason to prefer one image over the other, we aim to find the unique delay Δ that minimises both the errors $E_A(\mathbf{h}; \Delta)$ and $E_B(\mathbf{h}; \Delta)$ with the same ‘level of importance’. In other words, we are looking for Δ and the set of kernel width parameters $\mathbf{h} = (h^1, h^2, \dots, h^L)$, one for each dataset D^ℓ in D , that minimise the error

$$E(\mathbf{h}; \Delta) = E_A(\mathbf{h}; \Delta) + E_B(\mathbf{h}; \Delta).$$

Note that the imposition that there is a unique delay Δ for the whole data D and that the kernel widths are the same throughout each set D^ℓ for all the corresponding mean models $f_A^\ell(t)$, $f_B^\ell(t)$, $f_A^\ell(t; \Delta)$ and $f_B^\ell(t; \Delta)$, not only makes sense from the point of view of underlying physics, but is also a stabilising factor in the analysis and modelling of D .

The structure of our problem enables us to use an efficient and practical approach to finding the optimal time delay Δ_* . The error $E(\mathbf{h}; \Delta)$ to be minimised can be rewritten as

$$E(\mathbf{h}; \Delta) = \sum_{\ell=1}^L E^\ell(h^\ell; \Delta), \quad (7)$$

where

$$E^\ell(h^\ell; \Delta) = E_A^\ell(h^\ell; \Delta) + E_B^\ell(h^\ell; \Delta).$$

For every test value Δ we can separately optimise $E^\ell(h^\ell; \Delta)$ for h^ℓ within each set D^ℓ . Note that this boils down into a set of L one-dimensional optimisations of bandwidths h^1, h^2, \dots, h^L . In addition, because of the

nature of the mean models, the errors $E^\ell(h^\ell; \Delta)$ will behave ‘reasonably’ with changes in h^ℓ , i.e. the changes will be smooth and we can expect a roughly unimodal shape of cross-validated $E^\ell(h^\ell; \Delta)$. That enables us to use further speed-up tricks (such as halving) in the 1-dimensional optimisations. The estimated time delay is the one with the minimal overall $E(\mathbf{h}; \Delta)$ for the (cross-validation) optimised kernel width parameters \mathbf{h} .

3 NADARAYA-WATSON ESTIMATOR WITH LINEAR NOISE MODEL (NWE++)

In section §2 only the mean fluxes were modelled, the standard errors on observations were assumed known. Our approach can be extended to full probabilistic modelling by assuming a model for the relationship between the noise level and the observed fluxes. Here, we consider a simple model in which the standard error on the measured flux depends linearly on the observed flux value y , i.e., $\sigma(y) = \nu \cdot y$, where the proportionality constant ν depends on the wavelength at which the flux is measured (e.g., ν could be 1% and 0.1% for radio and optical data, respectively). Assuming that the mean models for dataset D^ℓ are fitted reasonably well, so that $y_{I,k}^\ell \approx f_I^\ell(t_k^\ell)$, $I \in \{A, B\}$, then to lowest order $\sigma(y_{I,k}^\ell) \approx \nu^\ell \cdot f_I^\ell(t_k^\ell)$.

Most of the material developed in sections 2 will stay unchanged, modifications are required only in the formulation of the noise models (3) and (4):

$$p_A(y_{A,k}^\ell | t_k^\ell) = \frac{1}{\nu^\ell \sqrt{2\pi} f_A^\ell(t_k^\ell)} \exp \left\{ \frac{-1}{2(\nu^\ell)^2} \left[\frac{y_{A,k}^\ell}{f_A^\ell(t_k^\ell)} - 1 \right]^2 \right\} \quad (8)$$

and

$$p_B(y_{B,k}^\ell | t_k^\ell, \Delta) = \frac{1}{M \nu^\ell \sqrt{2\pi} f_A^\ell(t_k^\ell - \Delta)} \cdot \exp \left\{ \frac{-1}{2(\nu^\ell)^2} \left[\frac{y_{B,k}^\ell}{M f_A^\ell(t_k^\ell - \Delta)} - 1 \right]^2 \right\}. \quad (9)$$

This time, however, we can write a full probabilistic model for any time point t and evaluate the likelihood within our model given any observation pair $(y_A^\ell(t), y_B^\ell(t))$ that could have been measured at time t :

$$p_A(y_A^\ell(t)) = \frac{1}{\nu^\ell \sqrt{2\pi} f_A^\ell(t)} \exp \left\{ \frac{-1}{(\nu^\ell)^2} \left[\frac{y_A^\ell(t)}{f_A^\ell(t)} - 1 \right]^2 \right\} \quad (10)$$

and

$$p_B(y_B^\ell(t) | \Delta) = \frac{1}{M \nu^\ell \sqrt{2\pi} f_A^\ell(t - \Delta)} \cdot \exp \left\{ \frac{-1}{(\nu^\ell)^2} \left[\frac{y_B^\ell(t)}{M f_A^\ell(t - \Delta)} - 1 \right]^2 \right\}. \quad (11)$$

The approximation error $E_A^\ell(h^\ell; \Delta)$ to be minimised by the choice of kernel width h^ℓ now reads:

$$E_A^\ell(h^\ell; \Delta) = \frac{1}{(\nu^\ell)^2} \sum_{k=1}^{N^\ell} \left\{ \left[\frac{y_{A,k}^\ell}{f_A^\ell(t_k^\ell)} - 1 \right]^2 + \left[\frac{y_{B,k}^\ell}{M f_A^\ell(t_k^\ell - \Delta)} - 1 \right]^2 \right\}.$$

Following analogous arguments for the case of modelling the source using image B, we have

$$p_A(y_A^\ell(t) | \Delta) = \frac{M}{\nu^\ell \sqrt{2\pi} f_B^\ell(t + \Delta)} \cdot \exp \left\{ \frac{-1}{(\nu^\ell)^2} \left[\frac{M y_A^\ell(t)}{f_B^\ell(t + \Delta)} - 1 \right]^2 \right\} \quad (12)$$

and

$$p_B(y_B^\ell(t)) = \frac{1}{\nu^\ell \sqrt{2\pi} f_B^\ell(t)} \exp \left\{ \frac{-1}{(\nu^\ell)^2} \left[\frac{y_B^\ell(t)}{f_B^\ell(t)} - 1 \right]^2 \right\},$$

which leads to the approximation error

$$E_B^\ell(h^\ell; \Delta) = \frac{1}{(\nu^\ell)^2} \sum_{k=1}^{N^\ell} \left\{ \left[\frac{M y_{A,k}^\ell}{f_B^\ell(t_k^\ell + \Delta)} - 1 \right]^2 + \left[\frac{y_{B,k}^\ell}{f_B^\ell(t_k^\ell)} - 1 \right]^2 \right\}.$$

Again, the final cost to be minimised is

$$E(\mathbf{h}; \Delta) = \sum_{\ell=1}^L E^\ell(h^\ell; \Delta), \quad (13)$$

where

$$E^\ell(h^\ell; \Delta) = E_A^\ell(h^\ell; \Delta) + E_B^\ell(h^\ell; \Delta).$$

4 PREVIOUS WORK

4.1 Cross Correlation

There are two versions of the methods based on cross correlation: the Discrete Correlation Function (DCF) (Edelson & Krolik 1988) and its variant, the Locally Normalised Discrete Correlation Function (LNDCF) (Lehar et al. 1992). Both calculate correlations directly on discrete pairs of light curves. These methods avoid interpolation in the observational gaps. They are also the simplest and quickest time delay estimation methods.

First, time differences (lags), $\Delta t_{ij} = t_j - t_i$, between all pairs of observations are binned into discrete bins. Given a bin size $\Delta\tau$, the bin centred at lag τ is the time interval $I_\tau = [\tau - \Delta\tau/2, \tau + \Delta\tau/2]$. The DCF at lag τ is given by

$$DCF(\tau) = \frac{1}{P(\tau)} \sum_{i,j}^{t_i, t_j \in I_\tau} \frac{(y_A(t_i) - \bar{a})(y_B(t_j) - \bar{b})}{\sqrt{(\sigma_a^2 - \sigma_A^2(t_i))(\sigma_b^2 - \sigma_B^2(t_j))}}, \quad (14)$$

where $P(\tau)$ is the number of observational pairs in the bin centred at τ , \bar{a} and \bar{b} are the means of the observed data, $y_A(t_i)$ and $y_B(t_j)$, and their variances are σ_a^2 and σ_b^2 , respectively.

Likewise,

$$LNDCF(\tau) = \frac{1}{P(\tau)} \sum_{i,j}^{t_i, t_j \in I_\tau} \frac{(y_A(t_i) - \bar{a}(\tau))(y_B(t_j) - \bar{b}(\tau))}{\sqrt{(\sigma_a^2(\tau) - \sigma_A^2(t_i))(\sigma_b^2(\tau) - \sigma_B^2(t_j))}}, \quad (15)$$

where $\bar{a}(\tau)$, $\bar{b}(\tau)$, $\sigma_a^2(\tau)$ and $\sigma_b^2(\tau)$ are the lag means and variances in the bin centred at τ .

The time delay Δ is found when $DCF(\tau)$ and $LNDCF(\tau)$, given by equations (14) and (15), are greatest; i.e., at the best correlation (Edelson & Krolik 1988; Lehar et al. 1992).

4.2 Dispersion Spectra

The Dispersion Spectra method (Pelt et al. 1996, 1998) measures the dispersion of time series of two light curves $y_A(t_i)$ and $y_B(t_j)$ by combining them (given a trial time delay Δ and ratio M) into a single signal, $y(t_k)$, $k = 1, 2, \dots, 2N$. In other words, given the delay Δ , the observed values of signal A , $\{y_A(t_i)\}_{i=1}^N$, and (delayed and rescaled) signal B , $\{\tilde{y}_B(t_i)\}_{i=1}^N$, where $\tilde{y}_B(t) = My_B(t - \Delta)$, are joined together and re-ordered in time, forming a joint signal $\{y(t_k)\}_{k=1}^{2N}$ of length $2N$. We employ two versions of this method (Pelt et al. 1998):

$$DS_1^2(\Delta) = \min_M \frac{\sum_{a=1}^{2N-1} w_a (y(t_{a+1}) - y(t_a))^2}{2 \sum_{a=1}^{2N-1} w_a} \quad (16)$$

and

$$DS_{2,4}^2(\Delta) = \min_M \frac{\sum_{a=1}^{2N-1} \sum_{c=a+1}^{2N} H_{a,c} W_{a,c} G_{a,c} (y(t_a) - y(t_c))^2}{2 \sum_{a=1}^{2N-1} \sum_{c=a+1}^{2N} H_{a,c} W_{a,c} G_{a,c}}, \quad (17)$$

where

$$w_a = \frac{1}{\sigma^2(t_{a+1}) + \sigma^2(t_a)}, \quad W_{a,c} = \frac{1}{\sigma^2(t_a) + \sigma^2(t_c)} \quad (18)$$

are the statistical weights taking in account the measurement errors, where $G_{a,c} = 1$ only when $y(t_a)$ and $y(t_c)$ are from different images, and $G_{a,c} = 0$ otherwise, and

$$H_{a,c} = \begin{cases} 1 - \frac{|t_a - t_c|}{\delta}, & \text{if } |t_a - t_c| \leq \delta \\ 0, & \text{otherwise.} \end{cases} \quad (19)$$

Compared with DS_1^2 , the $DS_{2,4}^2$ method has an additional parameter, the *decorrelation length* δ , which signifies the maximum distance between observations that we are willing to consider when calculating the correlations (Pelt et al. 1996).

The estimated time delay Δ is found by minimising DS^2 over a range of time delay trials Δ , as above.

5 DATA

We employ six different datasets from the same quasar Q0957+561, $L = 6$. The details are in Table 1 and the plots in Figure 1. The column labelled N^ℓ in Table 1

corresponds to the number of observations per dataset. The *Data* column shows whether the data are optical or radio and the *Type* column shows the filter and the frequency used to obtain the data. The Q0957+561 is a two-image quasar, so there is either an optical magnitude offset or a flux ratio between images A and B. D^1 was provided by R. Schild (Schild & Thomson 1997), private communication.

In order to consistently compare the performance of different time delay estimation methods in a controlled experimental setting, we also construct synthetic data on the basis of known gravitationally lensed fluxes in the optical and radio ranges, with the given observational noise and gaps structure. The ‘‘ground truth’’ - the delay - is imposed by us so that the statistics of different delay estimators can be consistently evaluated and compared.

5.1 Synthetic Data - Realistic Experimental Setting

In this section we construct synthetic signals on which we will test the proposed and some of the existing approaches to gravitational delay estimation in the presence of observational noise and gaps. We constructed synthetic fluxes in the optical range on the basis of D^1 (real r-band optical data of (Schild & Thomson 1997)) spanning roughly 10 and half years). In particular, we used D^1 to fit a distribution of possible fluxes ‘compatible’ with the data (formulated as a Gaussian process) and then sampled from this distribution synthetic fluxes of 3,500 observations.

Gaussian process (GP) represents a distribution over functions

$$f(t) \sim GP(\mu_{gp}(t), K_{gp}(t, t')), \quad (20)$$

with mean and covariance functions $\mu_{gp}(t)$ and $K_{gp}(t, t')$, respectively. Any sample from the GP corresponding to a finite set of observational times t_1, t_2, \dots, t_N is Gaussian distributed with mean $\mu_{gp}(t_1), \mu_{gp}(t_2), \dots, \mu_{gp}(t_N)$ and covariance matrix

$$K_{gp} = \begin{pmatrix} K_{gp}(t_1, t_1) & K_{gp}(t_1, t_2) & \dots & K_{gp}(t_1, t_N) \\ K_{gp}(t_2, t_1) & K_{gp}(t_2, t_2) & \dots & K_{gp}(t_2, t_N) \\ \vdots & \vdots & \ddots & \vdots \\ K_{gp}(t_N, t_1) & K_{gp}(t_N, t_2) & \dots & K_{gp}(t_N, t_N) \end{pmatrix}. \quad (21)$$

For our purposes, we imposed zero mean (the mean of observations in D^1 was shifted to zero) and used the ‘squared exponential’ kernel function

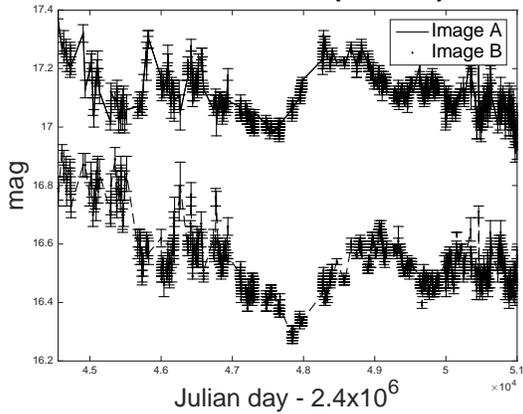
$$K_{gp}(t, t') = \exp \left\{ -\frac{(t - t')^2}{h_{gp}^2} \right\}, \quad (22)$$

with scale parameter h_{gp} set using cross validation on D^1 .

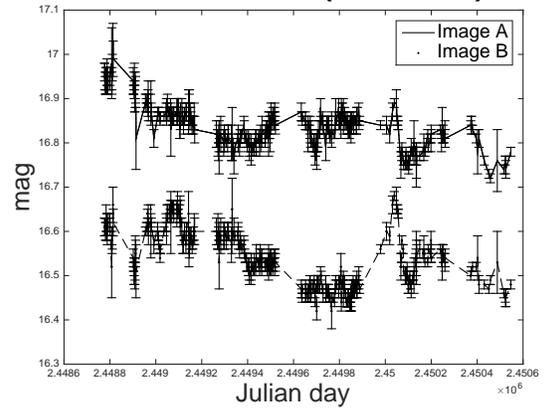
A vector $(\mathbf{y}, \mathbf{y}_*)^T$ of observations sampled at observation times \mathbf{t} and \mathbf{t}_* from the Gaussian process is distributed as

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y}_* \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} K_{gp} & K_{gp*} \\ K_{gp*}^T & K_{gp**} \end{pmatrix} \right), \quad (23)$$

Q0957+561 at r-band (Schild) n=1232 **Q0957+561 at r-band (Ovaldsen) n=422**

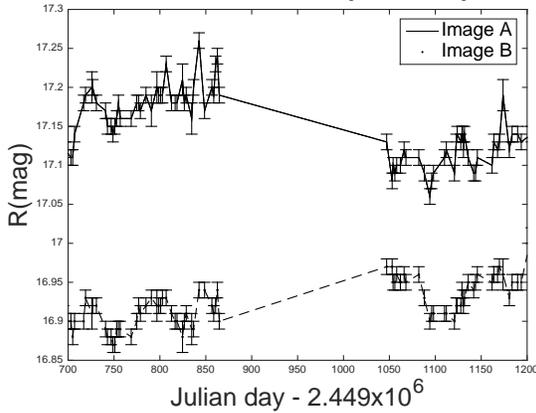


(a) D^1



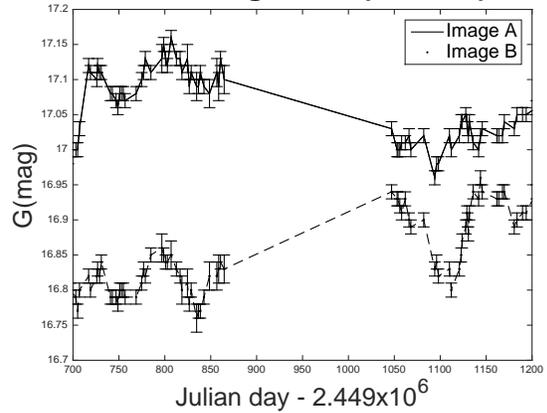
(b) D^2

Q0957+561 at r-band (Kundic) n=100



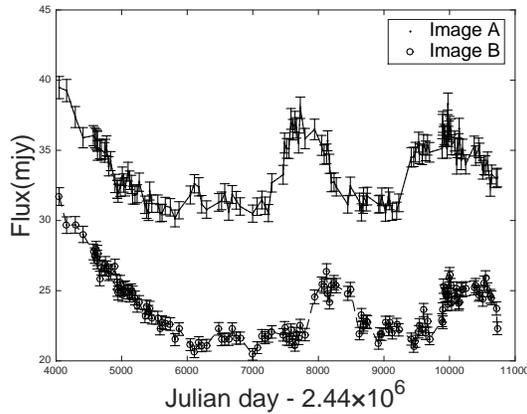
(c) D^3

Q0957+561 at g-band (Kundic) n=97



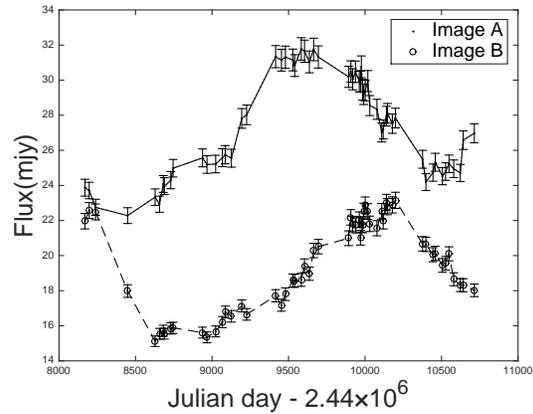
(d) D^4

Q0957+561 at 6cm n=143



(e) D^5

Q0957+561 at 4cm n=58

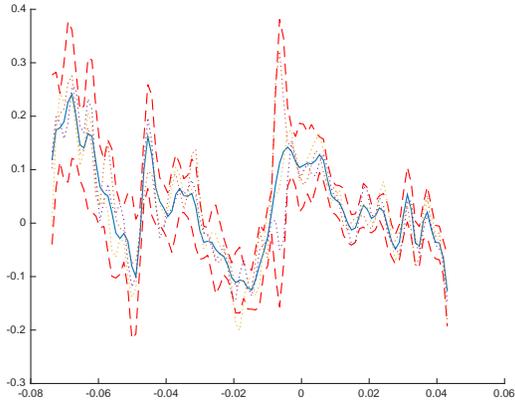


(f) D^6

Figure 1. Data set Q0957+561. Image A from D^1 is shifted up by 0.6 magnitudes for clarity; image A from D^2 is shifted up by 0.25 magnitudes; image A from D^4 is shifted up by 0.05 magnitudes. For more details on these datasets see Table 1.

Table 1. Datasets: Q0957+561

Id	N^ℓ	Data	Type	Ratio/Offset	Monitoring Range	Ref
D^1	1232	optical	r-band	0.05	16/11/1979 – 4/7/1998	(Schild & Thomson 1997)
D^2	422	optical	r-band	0.076	2/6/1992 – 8/4/1997	(Ovaldsen et al. 2003a)
D^3	100	optical	r-band	0.21	3/12/1994 – 6/7/1996	(Kundic et al. 1997)
D^4	97	optical	g-band	0.117	3/12/1994 – 6/7/1996	(Kundic et al. 1997)
D^5	143	radio	6cm	1/1.43	23/6/1979 – 6-Oct-1997	(Haarsma et al. 1999)
D^6	58	radio	4cm	1/1.44	4/10/1990 – 22/9/1997	(Haarsma et al. 1999)


Figure 2. Three Gaussian process posterior samples (dotted) based on D^1 (solid). Dashed curves signify ± 2 standard deviations.

where K_{gp} , K_{gp*} and K_{gp**} are kernel matrices corresponding to time instances $\mathbf{t} \times \mathbf{t}$, $\mathbf{t} \times \mathbf{t}_*$ and $\mathbf{t}_* \times \mathbf{t}_*$, respectively. However, given observations \mathbf{y} at times \mathbf{t} , the conditional distribution of \mathbf{y}_* at times \mathbf{t}_* is given by

$$p(\mathbf{y}_* | \mathbf{t}_*, \mathbf{y}, \mathbf{t}) = N(\mathbf{y}_* | \mu_*, \Sigma_*) \quad (24)$$

with

$$\mu_* = K_{gp*}^T K_{gp}^{-1} \mathbf{y} \quad (25)$$

and

$$\Sigma_* = K_{gp**} - K_{gp*}^T K_{gp}^{-1} K_{gp*}. \quad (26)$$

We sampled signals \mathbf{y}_* from the Gaussian process based on D^1 on a regular grid of 3,500 time stamps covering the temporal range of D^1 . As an example, we show three such signals in Figure 2. Dashed curves signify ± 2 standard deviations. To create a pair of time shifted signals A and B, the smooth long signal (signal A) $y_A = \mathbf{y}_*$ was shifted in time by a delay $\Delta = 200$ days to obtain signal B,

$$y_B(t) = y_A(t - \Delta). \quad (27)$$

Finally, (as explained in greater detail in the following sections), we added observational noise independently to both signals A and B, and imposed observational gaps.

5.1.1 Observational Noise

Based on D^1 data, we first calculated the empirical distribution $p(\rho)$ of the ratio ρ of the reported flux levels y_k and their associated standard errors σ_k : $\rho_k = \sigma_k / y_k$. For each observation $y(t)$ in the synthetic stream we generated an additive observational noise from a zero mean Gaussian distribution with standard deviation $\sigma(t)$, where $\sigma(t) = \rho(t)y(t)$, with $\rho(t)$ generated randomly i.i.d. from the empirical distribution $p(\rho)$.

5.1.2 Observational Gaps

Real data are irregularly sampled due to practical considerations such as weather conditions, equipment availability, object visibility, etc. (Eigenbrod et al. 2005; Cuevas-Tello 2007). Gaps in real data are characterised by two important quantities: gap size and gap position. The histogram in Figure 3(a) shows the empirical gap size distribution in D^1 . Shorter gaps of 1–5 days are more frequent than longer ones (more than 6 days).

To make the synthetic data more realistic, we would like to respect constraints given by the gap size and inter-gap distance distributions for dominant gap sizes (up to 10 days). Gaps were imposed on the synthetic data by generating their sizes and positions through a multiobjective optimisation algorithm. The algorithm incorporated three constraints: (1) closeness of the generated and empirical gap size distributions; (2) closeness of the generated and empirical inter-gap interval distributions for gaps of 1-5 days; (3) closeness of the generated and empirical inter-gap interval distributions for gaps of 6-10 days.

The particular algorithm we used was the computationally efficient *Random Weighted Genetic Algorithm (RWGA)* (Ghosh & Dehuri 2004; Konak, Coit & Smith 2006; Murata & Ishibuchi 1995; Murata, Ishibuchi & Tanaka 1996; Zitzler, Deb & Thiele 2000). It uses a weighted average of normalised objectives for fitness assignment (for diversity imposition the weights are randomized). The procedure is outlined in Algorithm 1.

The genome of each individual contains a suggestion for start positions and sizes of observational gaps. The design of individuals allows for a variable number of gaps and ensures that the gaps are not overlapping. Figure 3 shows the results of applying the multiobjective genetic algorithm RWGA based on D^1 . Each objective corresponds to a row of two plots in Figure 3,

Algorithm 1 RWGA

S = external archive to store non-dominated solutions found during the search so far;

n_S = number of elitist solutions immigrating from S to the population of potential solutions X_ι in each generation ι .

Step 1: Generate a initial random population X_1 , set $\iota = 1$.

Step 2: Assign a fitness value to each individual solution $\chi \in X_\iota$ by performing the following steps:

Step 2.1: Calculate the fitness $z_o(\chi)$ for each objective $o = 1, \dots, O$.

Step 2.2: Generate a random number u_o in $[0, 1]$ for each objective $o = 1, \dots, O$

Step 2.3: Calculate the random weight of each objective o as $w_o = \frac{1}{u_o} \sum_{i=1}^o u_i$.

Step 2.4: Update the overall fitness of the solution χ as $F(\chi) = \sum_{o=1}^O w_o z_o(\chi)$.

Step 3: Calculate the selection probability $p_s(\chi)$ of each solution $\chi \in X_\iota$ as follows:

$$p_s(\chi) = \frac{\sum_{\Upsilon \in P_\iota} (F(\Upsilon) - F^{min})}{F(\chi) - F^{min}},$$

where $F^{min} = \min \{F(\chi) \mid \chi \in X_\iota\}$.

Step 4: Select parents using the selection probabilities calculated in Step 3. Mutate offspring with a predefined mutation rate. Copy all offspring to $X_{\iota+1}$.

Step 5: Randomly remove n_S solutions from $X_{\iota+1}$ and add the same number of solutions from S to $X_{\iota+1}$.

Step 6: If the stopping condition is not satisfied, set $\iota = \iota + 1$ and go to Step 2. Otherwise, return to S .

left and right plots showing empirical normalized histograms from the real and synthetic data, respectively.

Generation of synthetic 'radio' data proceeded in the same way as described in the previous section for optical data, this time based on data D^5 .

5.2 Synthetic Data - Controlled Experimental Setting

Generation of synthetic fluxes described above was motivated by the desire to preserve realistic gap and noise distributions. We will refer to this approach as the '*realistic experimental setting*' (RS). For comparing delay estimation algorithms in a large-scale controlled setting, we also considered an alternative specification of gap and noise distributions. The synthetic fluxes were first generated from the Gaussian Process model fitted to D^1 , as described in the previous section. The fluxes were then corrupted with observational gaps and noise. The gap sizes g were generated as realisations from a mixture distribution $P_M(g) = \alpha \cdot P_B(g; \mu_g) + (1 - \alpha) \cdot P_U(g; L_g, U_g)$, where $P_B(g; \mu_g)$ is the Binomial distribution with mean μ_g and $P_U(g; L_g, U_g)$ is the uniform distribution over $[L_g, U_g]$. We used the following settings: $\alpha = 0.95$, $\mu_g = 4, 6, 8$ days, $L_g = 20$ and $U_g = 80$. The gap positions were randomised, subject to the con-

straint of minimum inter-gap distance of 2 days. The allowed range for gap size was 1 to 80 days. For the additive Gaussian zero mean 'observational' noise we considered three settings for the standard deviation: 0.1%, 0.2% and 0.3% of the flux level. We will refer to this approach as the '*controlled experimental setting*' (CS).

6 EXPERIMENTAL RESULTS

We performed experiments on synthetic datasets described in section §5, as well as on real gravitationally lensed fluxes in the radio and optical ranges. In the experiments we compared our methods NWE and NWE++, introduced in sections 2 and 3, respectively, with two dispersion spectra approaches, namely DS_1^2 and $DS_{2,4}^2$ and two cross correlation approaches DCF and LNDCF.

6.1 Experiments on Synthetic Data

As mentioned above, we set the 'true' time delay in the synthetic data to 200 days. The results of all approaches are based on testing time delay values in the range of 175 to 225 days (1 day increment).

It was found that the best setting for decorrelation length δ in the $DS_{2,4}^2$ method was 3 days. For NWE and NWE++ the kernel width h was estimated as variable kernel width with $h = 2$ neighbours⁵. For DCF and LNDCF, the bin size is set to 5 days. (see (Cuevas-Tello, Tiño & Raychaudhury 2006)).

For each method we show the mean (bias) μ and standard deviation σ of the maximum-likelihood delay estimates across experiments. In all plots, the true delay is represented by the horizontal line at $\mu = 200$.

6.1.1 Realistic Experimental Setting (RS)

For synthetic experiments in the realistic setting we generated 500 base signals from the Gaussian process fitted to the optical data set D^1 , as described in section §5.1. We then ran the RWGA algorithm to generate 500 pareto front solutions for observational gap positions and sizes (see 5.1.2). Each base signal thus had a corresponding observational gap structure imposed on it. Finally, the signals were corrupted by observational noise (see 5.1.1). The same procedure was applied for generating 500 datasets in the radio range.

Summary results for the RS experiments on the 500 optical and radio data sets are presented in Tables 2 and 3, respectively. We report the mean (μ) and standard deviation (σ) of the delay estimates $\hat{\Delta}_i$, $i = 1, 2, \dots, 500$, the mean absolute error (MAE) of the delay estimates (MAE = $\sum_{i=1}^{500} |\hat{\Delta}_i - 200|/500$), and the 95% Credibility

⁵ 2 neighbours came consistently as the favourite option when cross-validating the number of neighbours on several initial datasets.

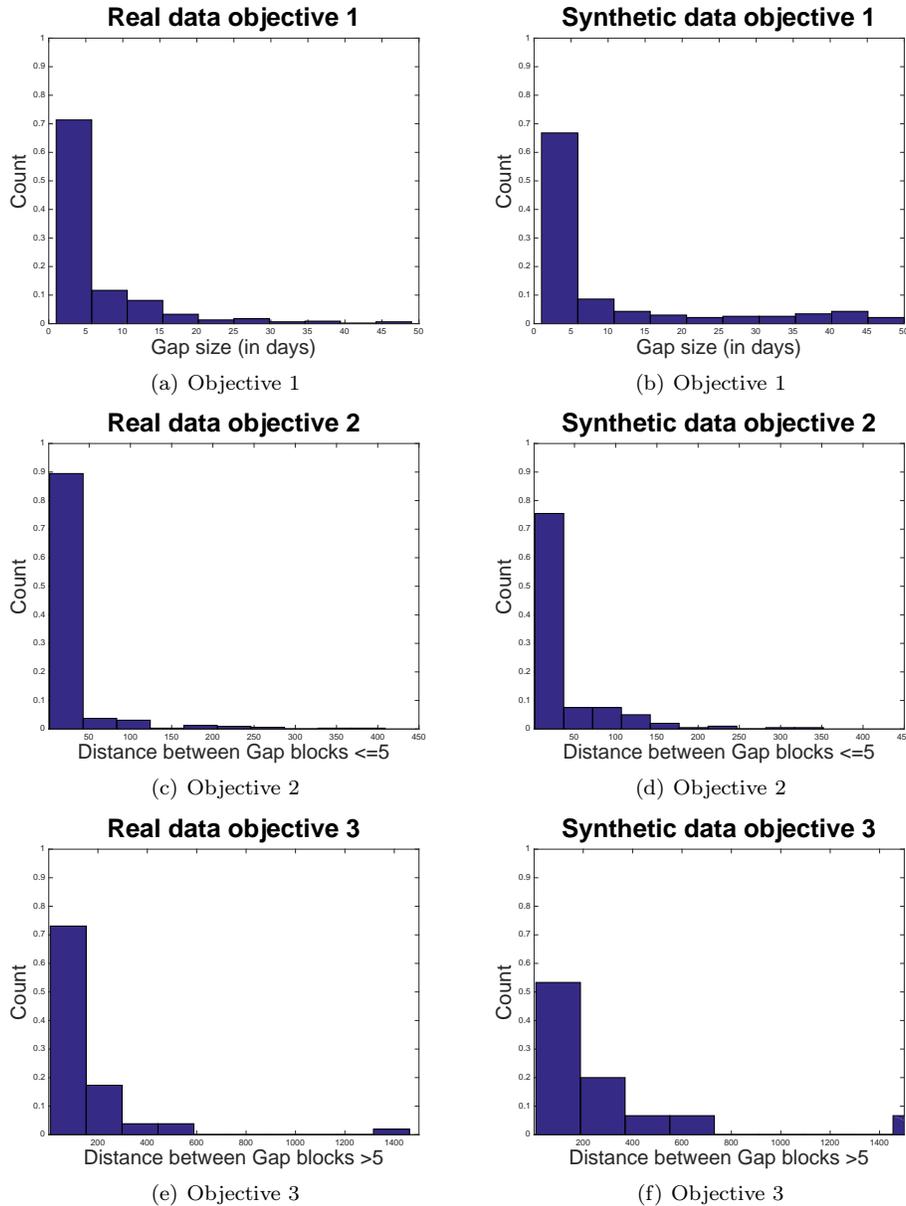


Figure 3. Empirical distributions of gap size (a),(b), inter-gap interval for gaps of 1–5 days (c),(d) and inter-gap interval for gaps of 6–10 days (e),(f). Each objective of RWGA corresponds to a row of two plots, left and right plots showing empirical normalized histograms from the real (D^1) and synthetic data, respectively.

Interval (CI). The overall performance of the methods is also shown in Figure 4. On smaller and noisier radio data the NWE is the best performing method, followed closely by NWE++. On optical data, the best performing method is $D_{2,4}^2$. It is important, however, to note that, in contrast to NWE methods, the dispersion spectra methods (DS) have parameters that are difficult to set objectively based on the given data only. In the experiments, we found the best DS parameter settings by imposing the true delay $\Delta = 200$, which obviously biases the DS results towards over-optimistic better performance levels.

6.1.2 Controlled Experimental Setting (CS)

For each setting of the Binomial gap distribution $\mu_g = 4, 6, 8$ days and for every noise level ratio from 0.1%, 0.2%, 0.3% we generated 100 base signals from the underlying Gaussian process fitted on D^1 . We thus obtained 900 datasets. The length of the time series (after applying observational gaps) varied from 800 to 3000 observations.

An analogous procedure was used to generate 900 datasets in the radio range. For each setting of the Binomial gap distribution $\mu_g = 4, 6, 8$ days and for every

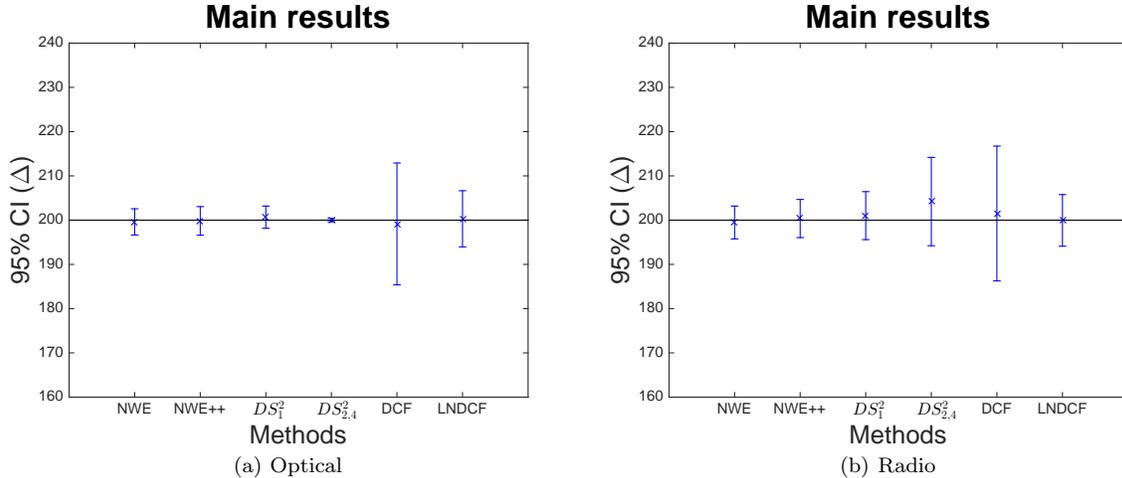


Figure 4. RS results for optical and radio data

Table 2. RS Results for optical range

Method	$\mu \pm \sigma$	MAE	CI range	95% CI
NWE	199.60 ± 2.97	2.19	0.26	[199.34, 199.86]
NWE++	199.83 ± 3.23	2.37	0.28	[199.55, 200.11]
DS_1^2	200.67 ± 2.51	1.05	0.22	[200.45, 200.89]
$DS_{2,4}^2$	200.02 ± 0.40	0.16	0.04	[199.98, 200.06]
DCF	199.14 ± 13.77	11.61	1.21	[197.93, 200.35]
LNDCF	200.30 ± 6.34	4.47	0.56	[199.74, 200.86]

noise level ratio from 1%, 2%, 3% we generated 100 base signals from the underlying Gaussian process fitted on D^5 . The overall results across all CS optical and radio datasets are summarized in Tables 4 and 5, respectively. Figures 5 and 6 present the results in greater detail, grouped by noise level and gap size.

The kernel-based methods lead to more stable time delay estimates. NWE is the best performing method with respect to all performance measures, followed by NWE++. It is interesting to note that while in general larger noise level ratio corresponds to larger standard deviation of the delay estimates, the DCF method seems to be more robust to increased noise levels. For low noise levels and with correlations between time-shifted data streams close to unity, the DCF method is, by construction, relatively insensitive to the level of the noise. However, it is still clearly outperformed by other techniques for the range of noise levels explored in this paper (see Figs. 4 and 5).

6.2 Experiments on Real Data

In this section, we present results of methods studied in this paper on real data - see Table 1 and Figure 1. Since for real data the noise levels related to observations are available, the NWE++ method was not used.

We have $L = 6$ datasets $D^1 - D^6$ and for all methods, we test values for time delay on the range of $\Delta = [400, 450]$ (increments of 1 day). The NWE cost to

be minimised is $E(\mathbf{h}; \Delta)$ (eq. (7)), with cross-validated kernel scale parameters $\mathbf{h} = (3, 2, 2, 2, 2, 2)$.

For DCF and LNDCF, the bin size $\Delta\tau$ was set to 5, 5, 5, 5, 45, and 30 for D^1, D^2, D^3, D^4, D^5 , and D^6 , respectively. As mentioned before, unlike in NWE, there is no objective way of setting such parameters based on the data only and we used the setting giving most robust results in the test range of delays 400-450 days. For a fixed delay Δ , the (LN)DCF function values at lag Δ are averaged across the 6 datasets $D^1 - D^6$ and the combined delay estimate is obtained at the maximum of the averaged (LN)DCF curve.

For the Dispersion spectra method $DS_{2,4}^2$, as argued above, the value of the decorrelation length parameter cannot be resolved in a principled manner based on the data and hence it was set to $\delta = 3$, since at this value DS_1^2 and $DS_{2,4}^2$ have more agreement. Again, for a fixed delay Δ , the $DS_1^2(\Delta)$ and $DS_{2,4}^2(\Delta)$ values are averaged across the 6 datasets and the combined delay estimate is obtained at the minimum of such averaged curves. The results (unique time delay across Q0957+561) are presented in Table 6.

To measure the uncertainty of time delay estimations, following (Haarsma et al. 1999; Oscoz et al. 1997, 2001; Ovaldsen et al. 2003b), we also performed Monte Carlo simulations by adding white noise generated according to the reported errors to each observation⁶. For

⁶ Note that this effectively adds noise to already noisy ob-

Table 3. RS Results for radio range

Method	$\mu \pm \sigma$	MAE	CI range	95% CI
NWE	199.47 \pm 3.71	2.95	0.32	[199.15,199.79]
NWE++	200.37 \pm 4.31	3.38	0.38	[199.99,200.75]
DS_1^2	201.02 \pm 5.42	4.42	0.47	[200.55,201.49]
$DS_{2,4}^2$	204.20 \pm 9.98	8.73	0.87	[203.33,205.07]
DCF	201.50 \pm 15.23	13.10	1.33	[200.17,202.83]
LNDCF	199.94 \pm 5.83	4.73	0.51	[199.43,200.45]

Table 4. Overall CS Results across all observational gap and noise settings for optical range.

Method	$\mu \pm \sigma$	MAE	CI range	95% CI
NWE	199.69 \pm 4.91	3.76	0.32	[199.37,200.01]
NWE++	199.69 \pm 5.78	4.41	0.38	[199.31,200.07]
DS_1^2	200.61 \pm 9.86	7.62	0.64	[199.97,201.25]
$DS_{2,4}^2$	199.97 \pm 14.10	11.98	0.92	[199.05,200.89]
DCF	202.71 \pm 16.26	14.22	1.06	[201.65,203.77]
LNDCF	200.63 \pm 10.56	8.37	0.69	[199.94,201.32]

each data set we generated 500 randomized Monte Carlo realisations. The results (mean and std dev across the 500 delay estimates) are presented in Table 7.

Although we cannot compare these results against a known true value, it is apparent that time delay estimates obtained with different methods are not mutually consistent, unlike estimates on synthetic data. For example, DS_1^2 and DCF estimates appear to lie more than 50σ apart. Moreover, we find that estimates using different frequency estimates on Q0957+561 data appear to be inconsistent even when the same method is used. This suggests that the claimed measurement errors on the data are significantly under-estimated. Alternatively, there may be unmodelled systematics (e.g., micro-lensing) that lead to varied biases for different analysis techniques.

7 CONCLUSIONS

We have introduced a new probabilistic efficient model-based methodology for estimating time delays between two gravitationally-lensed images of the same variable point source. The method enables one to use directly the noise levels reported for individual flux measurements. It is more robust to observational gaps than purely "unmodelled" techniques, since the imposition of an identical smooth model behind multiple lensed fluxes effectively regularizes the overall model fit, and consequently, the time delay estimate itself. Methods such as these will be useful in the automated search for time-delay systems as well as in the accurate measurement

of delays in targeted systems in future very large time-domain surveys such as those planned for the Large Synoptic Survey telescope (LSST) (e.g. (Hojjati & Linder 2015; Liao et al. 2015)).

The methods were tested and compared in two experimental settings. In the realistic setting the synthetic data were generated so that multiple aspects of the real data were preserved: noise-to-observed flux ratio, observational gap size distribution and the inter-gap interval distributions. The core synthetic signals were generated from a Gaussian process fitted to the real data. In the larger controlled experimental setting the signals generated from the Gaussian process were subject to controlled levels of observational noise and gap sizes. Our method, while being computationally efficient, showed robustness with respect to noise levels and observational gap sizes.

We also applied our method to real observed optical and radio fluxes from quasar Q0957+561 as a combined dataset. Of course, with real data one can estimate the variance of the estimator estimations, but never the bias, since the true time delay for Q0957+561 is not known. Our NWE estimator on the combined optical and radio data suggests a delay of approximately 420 days; however, we find that different estimators produce inconsistent results, indicating the presence of statistical or systematic measurement errors in the data in excess of the claimed measurement uncertainty. In particular, the impact of microlensing corrections was not accounted for in the present work, and needs to be quantified in the future.

servations, resulting in a different noise distribution. For example, assuming the original noise is Gaussian, and adding random Gaussian noise from the same distribution, the standard deviation of the noise distribution in this Monte Carlo data will be $\sqrt{2}$ larger than the original one.

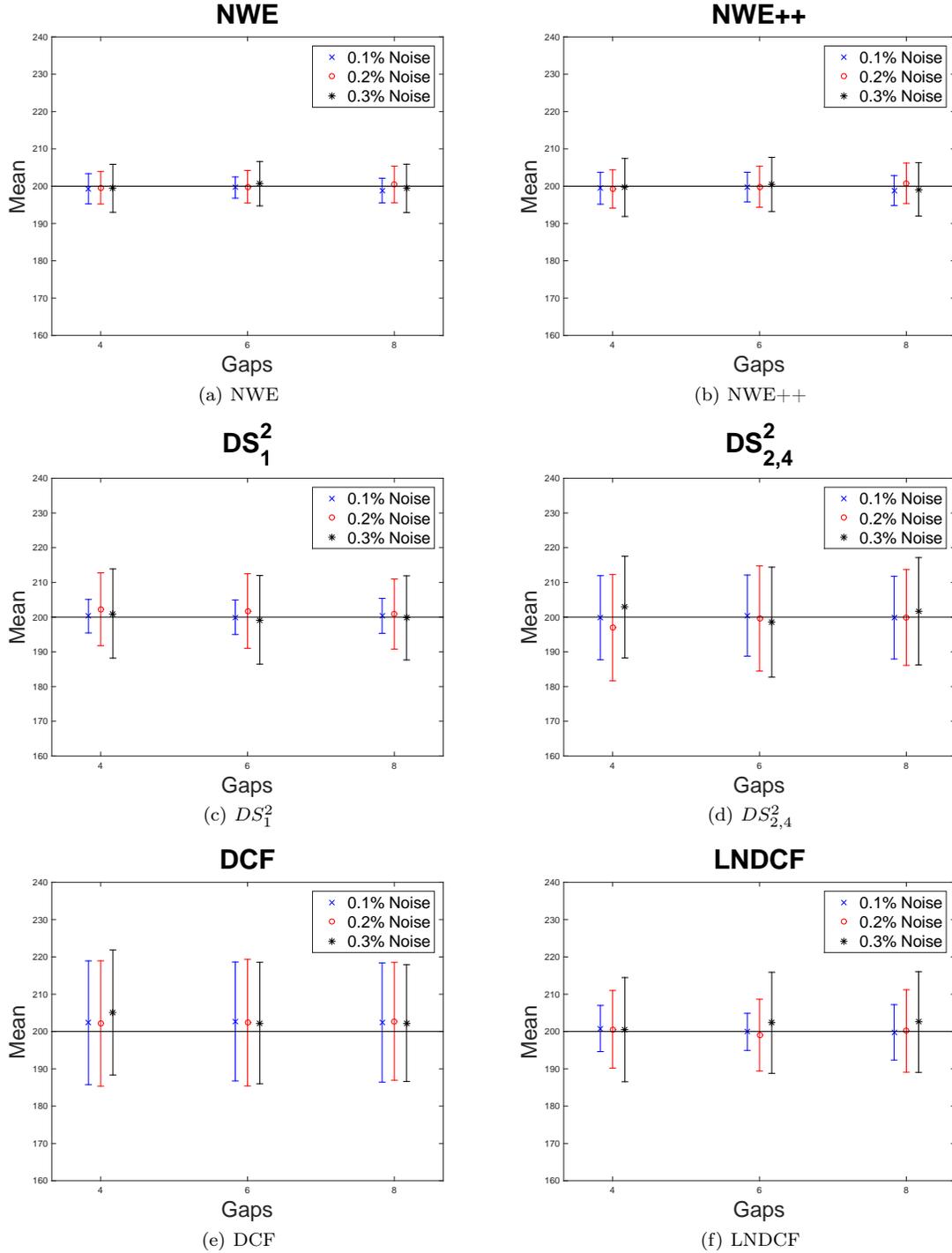


Figure 5. CS optical range results for NWE, NWE++, DS_1^2 , $DS_{2,4}^2$, DCF and LNDCF methods (plots (a), (b), (c), (d), (e) and (f), respectively) shown as functions of $\mu_g = 4, 6, 8$ days (mean of the binomial gap size distribution) and observational noise level. In each case we present the mean and std dev of the delay estimates for the corresponding 100 data sets.

Table 5. Overall CS Results across all observational gap and noise settings for radio range.

Method	$\mu \pm \sigma$	MAE	CI range	95% CI
NWE	199.70 \pm 4.23	3.24	0.28	[199.42,199.98]
NWE++	199.89 \pm 5.07	3.90	0.33	[199.56,200.22]
D_1^2	200.49 \pm 7.79	5.92	0.51	[199.98,201.00]
$D_{4,2}^2$	201.31 \pm 11.70	9.36	0.76	[200.57,202.09]
DCF	201.13 \pm 15.70	13.45	1.03	[200.10,202.16]
LNDCF	200.90 \pm 7.92	5.96	0.52	[200.38,201.42]

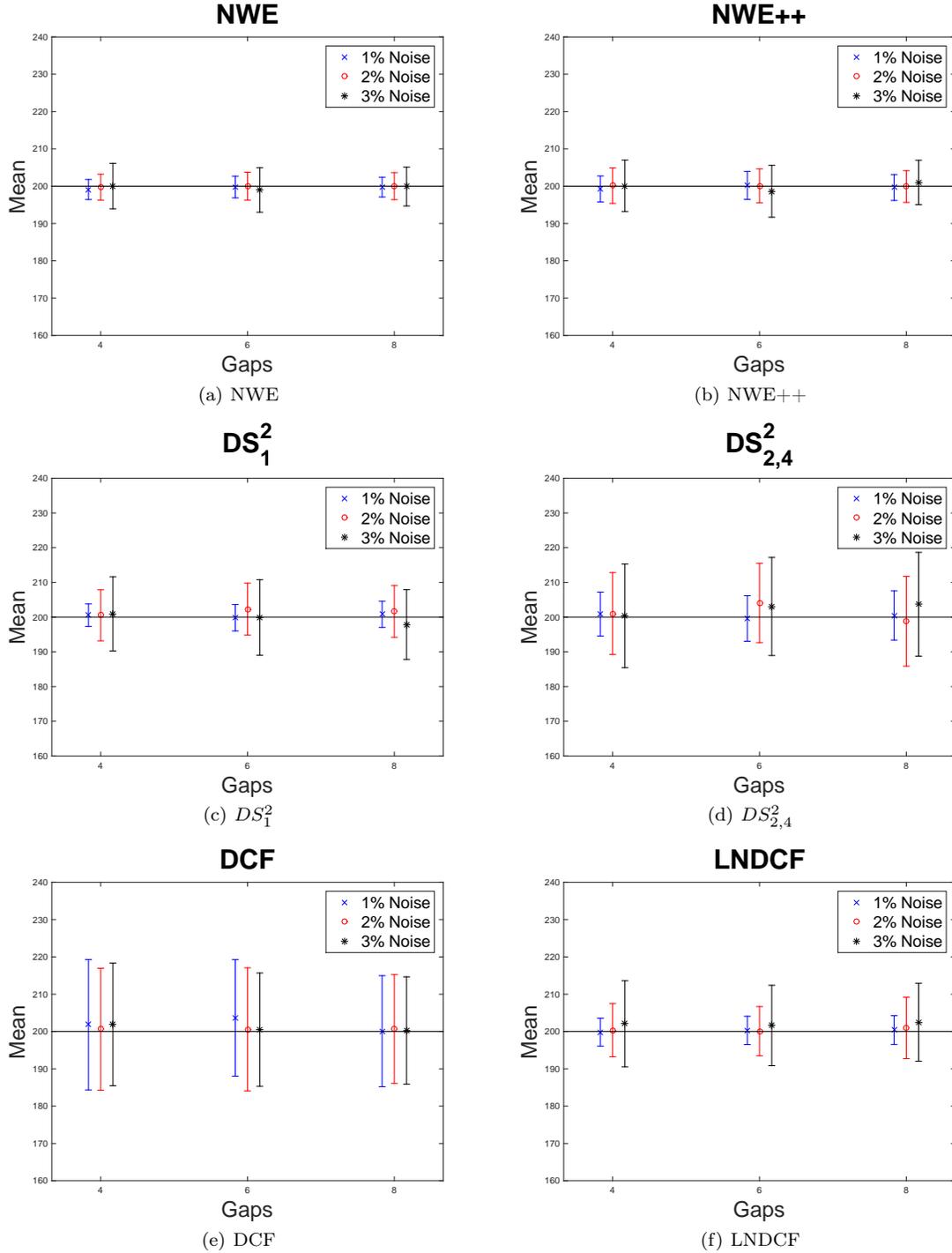


Figure 6. CS radio rang results for NWE, NWE++, DS_1^2 , $DS_{2,4}^2$, DCF and LNDCF methods (plots (a), (b), (c), (d), (e) and (f), respectively) shown as functions of $\mu_g = 4, 6, 8$ days (mean of the binomial gap size distribution) and observational noise level. In each case we present the mean and std dev of the delay estimates for the corresponding 100 data sets.

Table 6. The unique time delay across Q0957+561

Method	μ (days)
NWE	420
DS_1^2	435
$DS_{2,4}^2$	435
DCF	408.78
LNDCF	426.31

Table 7. Results of 500 Monte Carlo simulations:
Q0957+561

Method	μ (d)	σ (d)
NWE	418.65	0.49
DS_1^2	434.98	0.22
$DS_{2,4}^2$	434.92	1.08
DCF	408.77	0.42
LNDCF	431.09	15.04

ACKNOWLEDGEMENTS

The authors are grateful to Sherry Suyu, Ioana Oprea and to the anonymous reviewer for many helpful comments. Peter Tino was supported by the EPSRC grant EP/L000296/1.

REFERENCES

- Courbin F. et al., 2011, *Astronomy & Astrophysics*, 536
- Cuevas-Tello J., 2007, PhD thesis, School of Computer Science, University of Birmingham, <http://etheses.bham.ac.uk/88/>
- Cuevas-Tello J., Gonzalez-Grimaldo R., Rodriguez-Gonzalez O., Perez-Gonzalez H., Vital-Ochoa O., 2012, *Journal of Applied Research and Technology*, 10, 162
- Cuevas-Tello J., Perez-Gonzalez H., 2011, *Image*, 17, 17
- Cuevas-Tello J. C., Tino P., Raychaudhury S., Yao X., Harva M., 2009, *Pattern Recognition*, 43, 1165
- Cuevas-Tello J. C., Tiño P., Raychaudhury S., 2006, *Astronomy & Astrophysics*, 454, 695
- Dobler G., Fassnacht C. D., Treu T., Marshall P., Liao K., Hojjati A., Linder E., Rumbaugh N., 2015, *Astrophysical Journal*, 799, 168
- Edelson R., Krolik J., 1988, *The Astrophysical Journal*, 333, 646
- Eigenbrod A., Courbin F., Vuissoz C., Meylan G., Saha P., Dye S., 2005, *Astronomy & Astrophysics*, 436, 25
- Eulaers E. et al., 2013, *Astronomy & Astrophysics*, 553
- Fadely R., Keeton C. R., Nakajima R., Bernstein G. M., 2010, *Astrophysical Journal*, 711, 246
- Fassnacht C. D., Xanthopoulos E., Koopmans L. V. E., Rusin D., 2002, *Astrophysical Journal*, 581, 823
- Fohlmeister J., Kochanek C. S., Falco E. E., Wambsganss J., Oguri M., Dai X., 2013, *The Astrophysical Journal*, 764, 186
- Ghosh A., Dehuri S., 2004, *International Journal of Computing & Information Sciences*, 2, 38
- Gonzalez-Grimaldo R., Cuevas-Tello J., 2008, in *Proceedings of 7th Mexican International Conference on Artificial Intelligence (MICAI)*, IEEE Computer Society, pp. 131–137
- Greene Z. S. et al., 2013, *The Astrophysical Journal*, 768, 39
- Haarsma D., Hewitt J., Lehar J., Burke B., 1999, *The Astrophysical Journal*, 510, 64
- Hainline L. J. et al., 2012, *The Astrophysical Journal*, 744, 104
- Hirv A., Olsper N., Pelt J., 2011, *ArXiv e-prints*
- Hojjati A., Kim A. G., Linder E. V., 2013, *Phys. Rev. D*, 87, 123512
- Hojjati A., Linder E., 2015, *Physical Review D*, 90, 123501
- Kochanek C. S., Morgan N. D., Falco E. E., McLeod B. A., Winn J. N., Dembicky J., Ketzeback B., 2006, *Astrophysical Journal*, 640, 47
- Konak A., Coit D. W., Smith A. E., 2006, *Reliability Engineering & System Safety*, 91, 992
- Kundic T. et al., 1997, *The Astrophysical Journal*, 482, 75
- Lehar J., Hewitt J., Roberts D., Burke B., 1992, *The Astrophysical Journal*, 384, 453
- Liao K. et al., 2015, *Astrophysical Journal*, 800, 11
- Linder E. V., 2011, *Physical Review D*, 84, 123529
- Murata T., Ishibuchi H., 1995, in *Evolutionary Computation, 1995.*, IEEE International Conference on, Vol. 1, p. 289
- Murata T., Ishibuchi H., Tanaka H., 1996, *Computers & Industrial Engineering*, 30, 957
- Nadaraya E., 1964, *Theory of Probability and its Applications*, 9, 141
- Oguri M., 2007, *The Astrophysical Journal*, 660, 1
- Oscoz A. et al., 2001, *The Astrophysical Journal*, 552, 81
- Oscoz A., Mediavilla E., Goicoechea L. J., Serra-Ricart M., Buitrago J., 1997, *The Astrophysical Journal Letters*, 479, L89
- Ovaldsen J., Teuber J., Schild R., Stabell R., 2003a, *Astronomy & Astrophysics*, 402, 891
- Ovaldsen J.-E., Teuber J., Schild R., Stabell R., 2003b, *Astron. Astrophys.*, 402, 891
- Pelt J., Kayser R., Refsdal S., Schramm T., 1996, *Astronomy & Astrophysics*, 305, 97
- Pelt J., Schild R., Refsdal S., Stabell R., 1998, *Astronomy & Astrophysics*, 336, 829
- Rathna Kumar S. et al., 2013, *Astronomy & Astrophysics*, 553
- Refsdal S., 1964, *MNRAS*, 128, 307
- Schild R., Thomson D., 1997, *The Astronomical Journal*, 113, 130
- Suyu S. H. et al., 2013, *The Astrophysical Journal*, 766, 70
- Suyu S. H., Marshall P. J., Auger M. W., Hilbert S., Blandford R. D., Koopmans L. V. E., Fassnacht C. D., Treu T., 2010, *Astrophysical Journal*, 711, 201
- Tewes M., Courbin F., Meylan G., 2013, *Astronomy & Astrophysics*, 553
- Tewes M. et al., 2013, *Astronomy & Astrophysics*, 556
- Treu T. et al., 2013, *ArXiv e-prints*
- Vuissoz C. et al., 2008, *Astronomy & Astrophysics*, 488, 481
- Vuissoz C. et al., 2007, *Astronomy & Astrophysics*, 464, 845
- Watson G., 1964, *Sankhya: The Indian Journal of Statistics, Series A*, 26, 359
- Zitzler E., Deb K., Thiele L., 2000, *Evolutionary Computation*, 8, 173