Astronomy & Astrophysics manuscript no. time¢delay¢fin
(DOI: will be inserted by hand later)

July 13, 2007

How accurate are the time delay estimates in gravitational lensing?

Juan C. Cuevas-Tello[1,3], Peter Tiňo[1], and Somak Raychaudhury[2]

[1] School of Computer Science, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom
[2] School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom
[3] Engineering Faculty, Autonomous University of San Luis Potosí, México

Received December 2005 / Accepted

Abstract. We present a novel approach to estimate the time delay between light curves of multiple images in a gravitationally lensed system, based on Kernel methods in the context of machine learning. We perform various experiments with artificially generated irregularly-sampled data sets to study the effect of the various levels of noise and the presence of gaps of various size in the monitoring data. We compare the performance of our method with various other popular methods of estimating the time delay and conclude, from experiments with artificial data, that our method is least vulnerable to missing data and irregular sampling, within reasonable bounds of Gaussian noise. Thereafter, we use our method to determine the time delays between the two images of quasar Q0957+561 from radio monitoring data at 4 cm and 6 cm, and conclude that if only the observations at epochs common to both wavelengths are used, the time delay gives consistent estimates, which can be combined to yield 408 ± 12 days. The full 6 cm dataset, which covers a longer monitoring period, yields a value which is 10% larger, but this can be attributed to differences in sampling and missing data.

Key words. Methods: statistical – Methods: data analysis – Gravitational lensing – quasars: individual: 0957+561 A,B

1. Introduction

Long before the first gravitationally lensed quasar was discovered in 1979, Refsdal suggested that time delays of source fluctuations between the multiple images could be used to measure the Universe (Refsdal 1964, 1966). This first lensed quasar, Q0957+561, is also the most studied so far (Fig. 1), and many attempts have been made to estimate the time delay between its two principal images.

The measurement of the delay between the images A and B has been the subject of sensitive controversy ever since the first claim of measurement in 1981, culminating in a "definitive" measure of a time delay of 417±3 days (Kundic et al. 1997). Haarsma et al. (1997) reviews the various measurements, showing how various delays in the range of 300 to 1000 days have been claimed, from various data sets using different methods (Kochanek & Schechter 2004). However, Ovaldsen et al. (2003) report a time delay of 424.9±1.2 days more recently from new optical photometry- given the error bars, this measurement is inconsistent with the definitive value quoted above. Other often-quoted studies like Oscoz et al. (2001), Burud et al. (2001) and Colley et al. (2003) have concluded that the delay is 422.6±0.6, 423±9 and 417±0.07 respectively (more estimates are in Table 1). Therefore, it would suffice to say that the last word has not be said regarding the time delay between the images of Q0957+561. Similar controversy exists for almost all other quasars measures reported to date.

To measure these time delays, typically ranging from a few days to a few years, between signals arriving from the same source but via different paths, one needs to frequently observe the same set of sources over long periods of time. Due to the usual methods of allocation of telescope time and the natural timescale of projects, the data obtained typically are not regularly sampled, and could be obtained by a wide range of instruments and at different frequencies, often with large gaps in the time series. Since the use of time delays in constraining cosmological parameters (e.g., Saha 2004), requires these delays to be measured to a precision and reliability that is better than afforded by current practice, it is important to look for better and more robust methods, where the dependence of the results on the incompleteness of the data is well understood.

We present a novel approach to the problem of determining the delay between noisy time-dependent signals that have been measured at irregular intervals over several years, often with large gaps in the monitoring programme. Ours is an automatic method that allows us to analyse large-scale experiments more accurately than typical methods. We study the effect of the different levels of noise against the gaps, varying its size; the light curves under analysis are always irregularly sampled. The results of this study should be taken into account before lunching an observational campaign for gravitational lens. Moreover, we

Send offprint requests to: somak@star.sr.bham.ac.uk present some results of this method for the quasar Q0957+561 observed on radio data at different wavelengths, 4 cm and 6 cm (Haarsma et al. 1999).

The remainder of this article is organised as follows: in §2, we present our method. §3 is a survey of methods to estimate the time delay presenting a detailed review of three of the most popular methods. §4 describes the artificial data generated to perform our simulations. §5 shows the results on these artificial data. In §6 we present estimates for the time delay between the two principal images of Q0957+561 from radio data at 4 cm and 6 cm, using our methods presented here, followed by a concluding summary.

2. The model

We model the observed flux at a given frequency (in the radio or optical range) from two lensed images A and B of the same distant source, as two time series

$$x_A(t_i) = h_A(t_i) + \varepsilon_A(t_i) \qquad x_B(t_i) = M \cdot h_B(t_i) + \varepsilon_B(t_i), \tag{1}$$

where M is the ratio of the fluxes of the two images, and $t_i, i = 1, 2, ..., n$ are discrete observation times. The observation errors $\varepsilon_A(t_i)$ and $\varepsilon_B(t_i)$ are modelled as zero-mean Normal distributions

$$N(0, \sigma_A(t_i)) \ \text{ and } \ N(0, \sigma_B(t_i)), \tag{2}$$

respectively. Now,

$$h_A(t_i) = \sum_{j=1}^{N} \alpha_j K(c_j, t_i) \tag{3}$$

is the "underlying" light curve that underpins image A, whereas

$$h_B(t_i) = \sum_{j=1}^{N} \alpha_j K(c_j + \quad, t_i) \tag{4}$$

is a time-delayed (by) version of $h_A(t_i)$ underpinning image B. The functions h_A and h_B are formulated within the generalised linear regression framework. Each function is a linear superposition of N kernels $K(\cdot, \cdot)$ centred at either c_j, $j = 1, 2, ..., N$ (function f_A), or $c_j +$, $j = 1, 2, ..., N$ (function f_B). The model (1)-(4) has N free parameters α_j, $j = 1, 2, ..., N$, that need to be determined by (learned from) the data. We use Gaussian kernels of width ω^2: for $c, t \in \Re$,

$$K(c, t) = \exp \frac{-|t - c|^2}{\omega_c^2}. \tag{5}$$

The kernel width $\omega_c > 0$ determines the 'degree of smoothness' of the underlying curves h_A and h_B. We describe setting of $\omega_j = \omega_{c_j}$ and regression weights α_j in the next subsections. In this study, we position kernels on all observations, i.e. $N = n$.

Finally, our aim is to estimate the time delay between the temporal light curves corresponding to images A and B. Given the observed data, the likelihood of our model reads

$$P(Data|Model) = \prod_{i=1}^{n} p(x_A(t_i), x_B(t_i) \mid \quad, \{\alpha_j\}), \tag{6}$$

where

$$\begin{aligned} p(x_A(t_i), x_B(t_i) \mid \quad, \{\alpha_j\}) = {} & \frac{1}{2\pi\sigma_A^2(t_i)\sigma_B^2(t_i)} \\ & \exp\left\{\frac{(x_A(t_i) - h_A(t_i))^2}{2\sigma_A^2(t_i)}\right\} \\ & \exp\left\{\frac{(x_B(t_i) - M \cdot h_B(t_i))^2}{2\sigma_B^2(t_i)}\right\}. \end{aligned} \tag{7}$$

The negative log-likelihood (without constant terms) simplifies to

$$Q = \sum_{i=1}^{n} \left(\frac{(x_A(t_i) - h_A(t_i))^2}{\sigma_A^2(t_i)} + \frac{(x_B(t_i) - M \cdot h_B(t_i))^2}{\sigma_B^2(t_i)} \right). \tag{8}$$

To avoid extrapolation when we apply a time delay to our underlying curve, we do not evaluate the goodness of fit over all observations:

$$Q = \sum_{u=1}^{n-b_1} \frac{(x_A(t_u) - h_A(t_u))^2}{\sigma_A^2(t_u)} + \sum_{v=b_2}^{n} \frac{(x_B(t_v) - M \cdot h_B(t_v))^2}{\sigma_B^2(t_v)}, \tag{9}$$

where b_1 is the greatest index satisfying $t_{n-b_1} \leq t_n - \quad_{max}$, and b_2 is the smallest index satisfying $t_{b_2} \geq t_1 + \quad_{max}$. Here, ${}_{max}$ is the maximum possible time delay we are willing to consider (fixed).

We determine the model parameters and evaluate Eq. (9) for a series of trial values of . The time delay is then estimated as the value of with minimal cost (9). Note that if the errors cannot be modelled as Gaussian, Eq. (9) would need to be rewritten using an appropriate noise term.

2.1. Weights

We rewrite Eq. (8) as

$$Q = \sum_{i=1}^{n} \left(\left[\frac{x_A(t_i)}{\sigma_A(t_i)} - \frac{h_A(t_i)}{\sigma_A(t_i)} \right]^2 + \left[\frac{x_B(t_i)}{\sigma_B(t_i)} - \frac{M \cdot h_B(t_i)}{\sigma_B(t_i)} \right]^2 \right). \tag{10}$$

Since we expect each of the two terms in (10) to be individually equal to zero, we impose

$$\boldsymbol{K\alpha} = \boldsymbol{x}, \tag{11}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_N)^T$,

$$\boldsymbol{K} = \left[\begin{array}{ccc} K_A(c_1, t_1) & \cdots & K_A(c_N, t_1) \\ \vdots & \ddots & \vdots \\ K_A(c_1, t_n) & \cdots & K_A(c_N, t_n) \\ \hline K_B(c_1, t_1) & \cdots & K_B(c_N, t_1) \\ \vdots & \ddots & \vdots \\ K_B(c_1, t_n) & \cdots & K_B(c_N, t_n) \end{array} \right], \qquad \boldsymbol{x} = \left[\begin{array}{c} \frac{x_A(t_1)}{\sigma_A(t_1)} \\ \vdots \\ \frac{x_A(t_n)}{\sigma_A(t_n)} \\ \hline \frac{x_B(t_1)}{\sigma_B(t_1)} \\ \vdots \\ \frac{x_B(t_n)}{\sigma_B(t_n)} \end{array} \right], \tag{12}$$

and the kernels $K_A(\cdot, \cdot)$, $K_B(\cdot, \cdot)$ have the form:

$$K_A(c, t) = \frac{K(c, t)}{\sigma_A(t)}, \qquad K_B(c, t) = \frac{M \cdot K(c + \quad, t)}{\sigma_B(t)}. \tag{13}$$

Hence,

$$\boldsymbol{\alpha} = \boldsymbol{K}^{+}\boldsymbol{x}. \tag{14}$$

We regularise the inversion in (14) through singular value decomposition (SVD).

2.2. Kernel parameters

In general, in order to use Gaussian kernels (5) in generalised linear regression (1)-(4), the kernel positions c_j, as well as kernel widths ω_j, need to be determined (Shawe-Taylor & Cristianini 2004, §9). Several approaches have been taken in the literature. For instance, those who use radial basis function (RBF) networks employ e.g. k-means clustering, or EM algorithm and Gaussian mixture modelling (Haykin 1999; Hastie et al. 2001, §6 & §8)[1]. We have explored two approaches to kernel positioning: **(i)** centres c_j uniformly distributed across the input range and **(ii)** centres c_j positioned at input samples $t_j,\ j = 1, 2, ..., n$. The latter approach lead to superior performance and the results reported in this paper were obtained using kernels centred at observation times t_j. As for the kernel widths, we propose two approaches: **(j)** fixed width ω and **(jj)** variable widths ω_j, $j = 1, 2, ..., n$. Both are described in the next subsections.

2.2.1. Fixed kernel width

The width of kernels determine the degree of smoothing for the underlying flux curves (3) and (4). Finding 'appropriate' values of smoothing parameters is one of the challenges in non- and semi-parametric regression. We use cross validation (Hastie et al. 2001, §7.10) to find the 'optimal' kernel width ω. In particular, we invoke a variant of k-fold-cross-validation: We start by dividing the data set uniformly into k blocks. In the first step, we construct a validation set as a collection of the first elements of each block. The validation set has k elements. The training set is formed by the remaining observations, i.e. the observations not included in the validation set. We fit our models on the training set and determine the mean square error (MSE) over a range of delay values   on the validation set. In the next step, we construct a new validation set as a collection of the second elements of each block. The new training set is again formed by the remaining observations. As before we fit our models on the training set and determine MSE on the validation set. We repeat this procedure r times, where r is the number of observations in each block. Finally, the mean of all such mean square errors (there is r of them), MSE$_{CV}$, is calculated. The kernel width ω selected using the cross-validation is the kernel width yielding the smallest MSE$_{CV}$. The scheme is summarised in Algorithm 1.

2.2.2. Variable kernel width

Rather than considering a fixed kernel width ω, in this section we allow variable width Gaussian kernels of the form

$$K(c_j, t_i) = \exp \frac{-|t_i - c_j|^2}{\omega_j^2}\ ;\ K(c_j + \quad, t_i) = \exp \frac{-|t_i - (c_j + \quad)|^2}{\omega_j^2}.$$

We determine each ω_j through a smoothing parameter $k \in \{1, 2, ..., k_{max}\}$. Parameter k is the number of neighbouring observations t_i on both sides of c_j (boundary conditions need to be taken into account). In particular, since we centre a kernel on each observation time, i.e. $c_j = t_j$, we have the cumulative kernel width

$$\omega_j = \sum_{d=1}^{k}(t_j - t_{j-d}) + (t_{j+d} - t_j) = \sum_{d=1}^{k}(t_{j+d} - t_{j-d}). \qquad (15)$$

The optimal value of k can be estimated using a cross-validation procedure analogous to that of section §2.2.1.

Algorithm 1: Cross validation

```
Fix $M$, $LowerBound$ and $UpperBound$
Fix $Blocks \leftarrow 5$
Fix $PointsPerBlock \leftarrow min(\{b_1, n - b_2\})/Blocks$
for $omega \leftarrow LowerBound$ to $UpperBound$ do
    for $i \leftarrow 1$ to $PointsPerBlock$ do
        Remove the $i^{th}$ observation of each block and include
        it in the validation set
        for    ←   $_{min}$ to   $_{max}$ do
            Get weights $\alpha$ on the training set (using eq. (14))
            Compute $h_A(t_u)$ and $h_B(t_v)$
            Get MSE on the validation set
            $S($  $)$ ←MSE
        $R(i) \leftarrow mean(S)$
    $Best(omega) \leftarrow mean(R)$
$\omega \leftarrow \text{argmin}_\omega(Best)$
```

3. Methods for estimating the Time delay

Table 1 contains a review, in chronological order, of the more recent time delay estimates of the quasar Q0957+561 and the methods employed. This gravitational lens is the most extensively monitored so far. Fig. 1 presents examples of the observed lightcurves across various frequency bands, from radio to optical. As is evident in Table 1, whole range of time delay estimates (with varying uncertainty bounds) for the gravitational lens are available. The problem is that *we do not know the actual time delay*. One of the aims of this paper is to study the reliability of several time delay estimation methods in a large set of controlled experiments on artificially generated data with realistically modelled observational noise and mechanisms of missing measurements. We feel that only after learning lessons from such a study does it make sense to come up with yet another batch of time delay estimation claims.

In this section, we review the principal time delay estimation methods that have been used on gravitational lens data. The **Cross correlation** method (Kundic et al. 1997; Oscoz et al. 1997), **PRH** method (Press et al. 1992) and **Dispersion** spectra, (Pelt et al. 1996), described in §3.1, §3.2 and 3.3, respectively, have been widely used in the literature. We employ them in §5 as base-line models when reporting performance of our methods (described in §2).

Of the methods mentioned in Table 1, the **Linear** method uses chi-squared (χ^2) fitting (Press et al. 1986, §14). Since the data are irregularly sampled, linear interpolation in the observational gaps is performed (Kundic et al. 1997).

The method of Subtractive Optimally Localised Averages (**SOLA**) has been proposed as a method for solving inverse

[1] Some approaches attempt to simultaneously optimize the number of kernels.

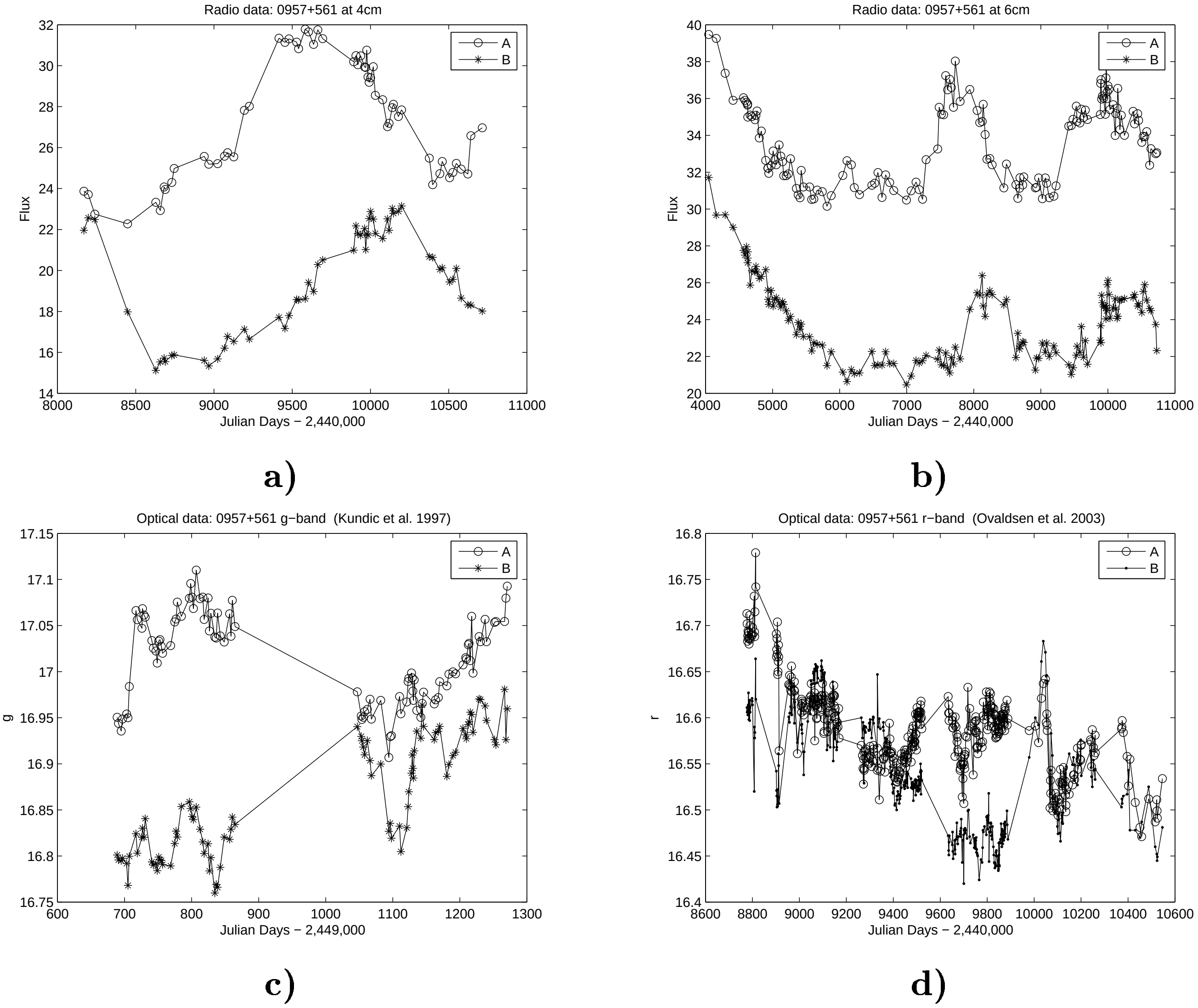


Fig. 1. The variation of flux density with time of the two gravitationally lensed images of Quasar Q0957+561. **a)** Radio data at 4 cm, **b)** radio data at 6 cm (Haarsma et al. 1999), **c)** optical data at g-band (Kundic et al. 1997), and **d)** optical data at r-band (Ovaldsen et al. 2003).

problems. The method was adopted by Pijpers (1997) who formulated time delay estimation as an inverse problem. It is worth nothing that SOLA employs kernels, called averaging kernels. However, SOLA differs from our approach in several respects: **(i)** SOLA does a symmetric treatment of the two estimated fluxes (flux A is fixed and flux B is varied to match A and vice versa), **(ii)** the reported time delay is the mean of the estimated time delays in the two symmetric cases, **(iii)** a free parameter is used to adjust the relative weighting of the errors in the variance-covariance matrix. We also note that parameter estimation in SOLA is problematic (Larsen & Hansen 1997; Rabello-Soares et al. 1999) and this method has been rarely used.

The χ^2 **algorithm** (Burud et al. 2001; Ovaldsen et al. 2003) is a χ^2-based method similar in spirit to our model in that it also uses a notion of an underlying model curve when fitting the two observed fluxes. However, the underlying model is assumed to be regularly sampled. It is regularised using a smoothing term (Burud et al. 2001, Eq. 3). Confidence intervals on the delay are estimated by performing Monte Carlo simulations (Burud et al. 2001).

In general, when Monte Carlo simulations are not performed, bootstrap techniques are used to calculate uncertainty in time delay estimates.

3.1. Cross correlation

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

Table 1. Review of time delay estimates between the two images of Q0957+561 from 1997 to 2004. The methods are reviewed in §3.

Reference	Method(s)	Time delay
Kundic et al. 1997	- Linear - Cross correlation - PRH - Dispersion	417±3
Oscoz et al. 1997	- Cross correlation - Dispersion	427±3
Pijpers 1997	- SOLA	425±17
Pelt et al. 1998b	- Dispersion	416.3±1.7
Haarsma et al. 1999	- PRH - Dispersion	409±30
Oscoz et al. 2001	- Linear - Cross correlation - Dispersion	422.6±0.6
Burud et al. 2001	- χ^2 algorithm	423±9
Colley et al. 2003	- PRH	417.09±0.07
Ovaldsen et al. 2003	- Dispersion - χ^2 algorithm	424.9±1.2

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

$$DCF(\tau) = \frac{1}{P(\tau)} \sum_{i,j} \frac{(x_A(t_i) - \bar{a})(x_B(t_j) - \bar{b})}{\sqrt{(\sigma_a^2 - \sigma_A^2(t_i))(\sigma_b^2 - \sigma_B^2(t_j))}}, \tag{16}$$

where $\bar{a}$ and $\bar{b}$ are means of the observed data fluxes $x_A(t_i)$ and $x_B(t_j)$, respectively; σ_a^2 and σ_b^2 are their variances; $\sigma_A^2(t_i)$ and $\sigma_B^2(t_j)$ are the observational errors (2).

Likewise,

$$LNDCF(\tau) = \frac{1}{P(\tau)} \sum_{i,j} \frac{(x_A(t_i) - \bar{a}(\tau))(x_B(t_j) - \bar{b}(\tau))}{\sqrt{(\sigma_a^2(\tau) - \sigma_A^2(t_i))(\sigma_b^2(\tau) - \sigma_B^2(t_j))}}, \tag{17}$$

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 centered at τ. The time delay is found when $DCF(\tau)$ and $LNDCF(\tau)$ (16)-(17) are maximum, i.e. at the best correlation.

3.2. The PRH method

This method is widely used for time delay estimation. Its fundamentals are based on the theory of stochastic processes and Wiener filtering (Press et al. 1992; Rybicki & Press 1992). Given two light curves x_A and x_B (1), the PRH method combines them into a single series $\boldsymbol{y}$ by assuming a time delay and a constant ratio M between x_A and x_B. Thus, for each of the two fluxes, we end up having a new data set of $2n$ observations; half is interpolated using the other flux. The flux ratio M is estimated as a difference between weighted means of the fluxes; the weights are derived from the quoted observational errors. The time delay, , is estimated by minimising

$$\chi^2 = \boldsymbol{y}^{\mathrm{T}} \left(\boldsymbol{A} - \frac{\boldsymbol{A}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{A}}{\boldsymbol{E}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{E}} \right) \boldsymbol{y}, \tag{18}$$

which is a measure of goodness of fit on measurements from a Gaussian process (Press et al. 1992). Here, $\boldsymbol{y}$ is the combined flux[2], $\boldsymbol{E}$ is a column vector of ones, and

$$\boldsymbol{A} = \boldsymbol{B}^{-1} \equiv \left\{ C_{ab} + \langle \sigma_a^2 \rangle \delta_{ab} \right\}^{-1} \tag{19}$$

where

$$C_{ab} = \langle y(t_a) y(t_b) \rangle \equiv C(t_a - t_b) \equiv C(\tau) \tag{20}$$

is a covariance model estimated from the data[3]; t_a, t_b, $a, b = 1, ..., 2n$, are sample times of the combined light curve. Press et al. (1992) suggest finding $C(\tau)$ through a first-order structure function $V(\tau) = \langle s^2 \rangle - C(\tau)$, where s is the clean data from $\boldsymbol{y}$. Then, the structure function $V(\tau)$ is computed from the data, single image, by determining lags

$$\tau_{ij} \equiv |t_i - t_j| \tag{21}$$

and values

$$v_{ij} \equiv (x_{\{A,B\}}(t_i) - x_{\{A,B\}}(t_j))^2 - \sigma^2_{\{A,B\}}(t_i) - \sigma^2_{\{A,B\}}(t_i). \tag{22}$$

where $\{A, B\}$ denotes that it comes from either image A or image B (1)-(2).

All pairs (τ_{ij}, v_{ij}) are sorted with respect τ_{ij} and binned into 100 bins (Press et al. 1992, pg. 407). The values of τ_{ij} and v_{ij} in each bin are averaged and finally a power-law model is built to fit the binned list,

$$V(\tau) = c_1 \tau^{c_2}. \tag{23}$$

Note that this model is linear in log scale,

$$V(\ln(\tau)) = \ln(c_1) + c_2 \ln(\tau). \tag{24}$$

Parameters c_1 and c_2 of the structure function can be determined using a simple line fitting algorithm[4]. Note that $V(\tau)$ is estimated on a single flow and one would naturally expect that estimates of $V(\tau)$ on flux A would be similar to those on flux B. However, this is often not the case. Press et al. (1992) claim that it does not matter which image is chosen for the $V(\tau)$ estimation as the time delay calculations are sufficiently robust to variations in the $V(\tau)$ estimates. Our experience suggests that this may be an overoptimistic expectation. Moreover, matrix $\boldsymbol{B}$ (19) is often ill conditioned and we regularise the inversion operation through SVD.

3.3. Dispersion spectra

Dispersion is a weighted sum of squared differences between $x_A(t_i)$ and $x_B(t_i)$ (Pelt et al. 1996, 1998b,a, 2002). The method is similar to those based on DCF (see §3.1). However, it models the time series of two light curves in a different way by

[2] Note that Press et al. (1992) refer to $\boldsymbol{y}$ as a component rather than combined components, image A and image B. The same occurs with the matrices $\boldsymbol{A}$,$\boldsymbol{B}$ and $\boldsymbol{C}$ in equations 19 and 20.

[3] Angle brackets denote the expectation operator.

[4] We have noticed that in some cases a negative slope c_2 is found. Also be aware that a negative c_1, y-intercept, in Eq. (24), and $\tau = 0$ leads to numerical overflow. In such cases we apply a shift up in Eq. (22), and we set τ to a very small positive number.

combing them (given a time delay   and ratio M) into a single flux flow, $\boldsymbol{y}$, as in the PRH method (§3.2). We worked with two versions of this method (see Pelt et al. 1998b):

$$D_1^2(\ ) = \min_M \frac{\sum_{a=1}^{2n-1} w_a \left(y(t_{a+1}) - y(t_a)\right)^2}{2\sum_{a=1}^{2n-1} w_a} \qquad (25)$$

and

$$D_{4,2}^2(\ ) = \min_M \frac{\sum_{a=1}^{2n-1}\sum_{c=a+1}^{2n} S_{a,c}^{(2)} W_{a,c} G_{a,c} \left(y(t_a) - y(t_c)\right)^2}{2\sum_{a=1}^{2n-1}\sum_{c=a+1}^{2n} S_{a,c}^{(2)} W_{a,c} G_{a,c}}, \qquad (26)$$

where

$$w_a = \frac{1}{\sigma^2(t_{a+1}) + \sigma^2(t_a)}, \qquad W_{a,c} = \frac{1}{\sigma^2(t_a) + \sigma^2(t_c)} \qquad (27)$$

are the statistical weights taking in account the measurement errors (2). $G_{a,c} = 1$ only when $y(t_a)$ and $y(t_c)$ are from different images, and $G_{a,c} = 0$ otherwise.

$$S_{a,c}^{(2)} = \begin{cases} 1 - \frac{|t_a - t_c|}{\delta}, & \text{if } |t_a - t_c| \le \delta \\ 0, & \text{otherwise.} \end{cases} \qquad (28)$$

The estimated time delay, , is found by minimizing D^2 over a range of time delay trials.

Compared with D_1^2, the $D_{4,2}^2$ method has an additional parameter, *decorrelation length* δ, that signifies the maximum distance between observations we are willing to consider when calculating the correlations (Pelt et al. 1996).

4. Constructing Artificial data sets

We use artificial data sets to perform a set of *controlled* large-scale experiments in order to measure the accuracy of time delay estimation techniques on gravitational lens systems. We generate simulated data sets with different levels of noise and varying sizes/locations of observational gaps.

The basic signal is constructed by superimposing $G = 20$ Gaussian functions with centres and widths generated randomly. Then, two artificial fluxes are created by scaling and shifting the basic signal in the flux density and time domains, respectively. The amplitude and flux densities are similar to radio data, 4 cm (Haarsma et al. 1999). The flux ratio was set to $M = 1/1.44$ and the temporal shift was equal to $= 500$ days. The time goes from 0 to $T \cdot$ days with s_1 samples per   days ($T = 10$ and $s_1 = 5$), i.e. if the samples were regularly sampled, we would have a separation of $z =$ $/s_1$ days between samples. To irregularly sample, we disturb the regular observation times with a random variable uniformly distributed in $[-P \cdot z, +P \cdot z]$, $P = 0.49$. Moreover, we simulate continuous gaps in observations by imposing $g = 5$ blocks of missing data. The blocks are located randomly with at least one sample between them. We worked with block lengths from 1 to $s_2 = 5$ (see Table 2).

Three levels of noise were used to contaminate the flux signal: 1%, 2% and 3% of the flux; these represent our measurement errors $\sigma_A(t_i)$ and $\sigma_B(t_i)$, which are standard deviations of the flux distribution at each observation time (see Eqs. (1) and (2)). Fig. 2 shows an example of a couple of scaled and shifted artificial fluxes[5].

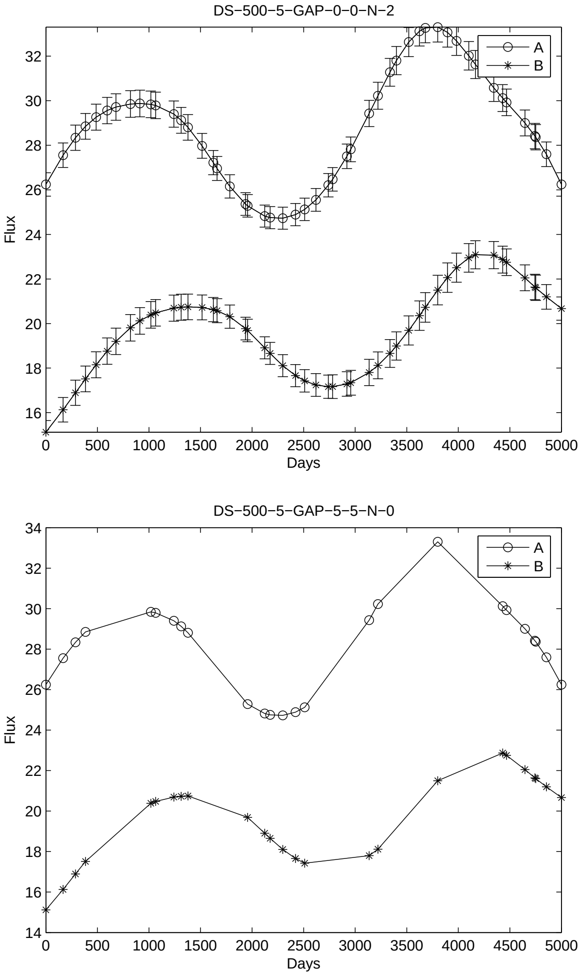


Fig. 2. Artificial flux data generated to simulate a couple of scaled and shifted fluxes coming from a quasar through a gravitational lens. The top plot shows the underlying function without observational gaps. Also shown are the error bars of 2% of the flux value. Below are the same noise-free fluxes with imposed observational gaps of length 5.

We used 20 different underlying functions (basic signals). For each underlying function, we generated 100 realisations for each noise level by adding a Gaussian noise to the underlying function as in equations (1) and (2). For each such data set, we performed 10 realisations of missing observational blocks. Overall, we employed 307 020 different data sets, 15 351 data sets per underlying function (see Table 2).

5. Experiments with artificial data

In this section, we test our methods of §2.2.1 and §2.2.2 on artificial data sets described in §4 and compare them the existing methods described in §3.1, §3.2 and §3.3. Figures 3–10 show two kinds of curves. The top panel shows curves representing the mean estimated time delay $_\mu$ versus the gap size for dif-

[5] More plots can be found at http://www.cs.bham.ac.uk/~jcc/artificial/

Table 2. Artificial data sets under analysis

	Gap size s_2					
Noise	**0**	**1**	**2**	**3**	**4**	**5**
0%	1	10	10	10	10	10
1%	100	1 000	1 000	1 000	1 000	1 000
2%	100	1 000	1 000	1 000	1 000	1 000
3%	100	1 000	1 000	1 000	1 000	1 000
Sub-Total	301	3 010	3 010	3 010	3 010	3 010

Total = 15 351 data sets per underlying function.
20 underlying functions yield 307 020 data sets.

ferent noise levels, while lower curves represent standard deviations $_{\sigma}$ of the estimated time delay. The quantities of data sets involved in this analysis are shown in Table 2.

In all experiments reported in this section, the following parameter settings were used: $M = 1/1.44$, $_{min} = 400$ and $_{max} = 600$ with increments of 1 day. We used a threshold of 0.001 (found empirically) to regularise inversion in Eq. 14 through SVD, discarding singular values less than the threshold (Press et al. 1986, §2).

Results for the fixed kernel width technique (Algorithm 1) are shown in Fig. 3. Here, $LowerBound = 900$ and $UpperBound = 1200$ with increments of 10. Fig. 4 shows results of the variable kernel width technique. We fixed the number of neighbors to $k = 3$, which was estimated through cross-validation (see Algorithm 1) with $LowerBound = 1$ and $UpperBound = 15$ with increments of 1.

Figures 5 and 6 contain results for DCF and LNDCF, respectively. For both methods, bin size of 100 days (which is close to the lag average) was used and a search was performed for the maximum correlation on bins in the range of 0 to 2 days (= 500 for artificial data).

Figure 7 displays results of the PRH method using image A to estimate the structure function (23), while Fig. 8 shows results obtained using structure function estimated on image B. When estimating the structure function, we use bins in the range 100 – 700 days. Robust linear regression (MATLAB Statistics Toolbox) was used to estimate parameters c_1 and c_2 in (24). If $\tau = 0$ in equation (20), we take the minimum lag from v_{ij} (see Eqs. (21) and (22)).

The results of Dispersion spectra method are in Figs. 9 and 10 for D_1^2 and $D_{4,2}^2$ respectively. We set $\delta = 100$ as decorrelation length [6] for $D_{4,2}^2$.

We point out that the results in Figs. 3 to 10 were obtained on the same collection of artificial data sets and are plotted with the same scale on the y-axis. Compared to the existing methods (DCF, LNDCF, PRH, Dispersion spectra), our methods are more accurate and robust with respect to the increasing gap size and noise level. In general, for all methods, there is (an obvious) tendency of increased uncertainty as the gap size increases. Increasing noise levels in the data result in increased uncertainty of the time delay estimates.

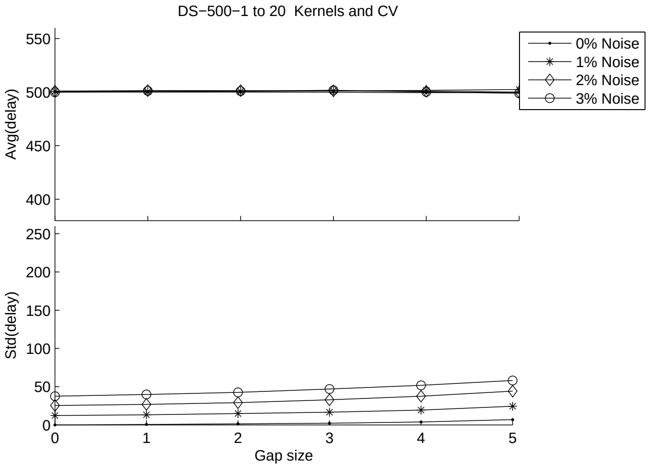


Fig. 3. Results of the application of our Kernel method with fixed width (in §2.2.1) on all artificial data sets (see §4). Details in §5.

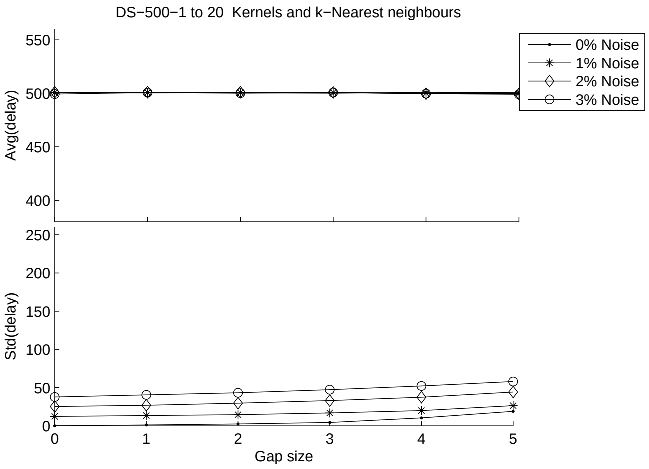


Fig. 4. Results of the application of our Kernel method with variable width (in §2.2.2) on all artificial data sets (see §4). Details are in §5.

6. The gravitational lens Q0957+561: radio observations

In this section we apply the tools developed in this paper to estimate the time delay for the much studied quasar Q0957+561. We use radio monitoring data at 4 cm and 6 cm wavelengths. For the 6 cm data set, we use the light curve with four points from Spring 1990 removed, as in (see Haarsma et al. 1999), [7]. These radio data sets are plotted at the top in Fig. 1. Our results are presented in Table 3.

To estimate the time delay for this quasar, we use both the fixed kernel width and variable kernel width approaches outlined in §2. We employ flux ratios $M = 1/1.44$ and $M = 1/1.43$ for the 4 cm and 6 cm data, respectively (the most likely values given our models). We tested time delays between $_{min} = 300$

[6] This value of decorrelation length gives the best resolution on artificial data. Pelt et al. (1996) and Haarsma et al. (1999) used $\delta = 60$ for radio data.

[7] Data from `http://space.mit.edu/RADIO/papers.html`. Note that the 6 cm data set has a record not included in the published papers and the observation on 11th April 1994 is recorded a day earlier in previous studies.

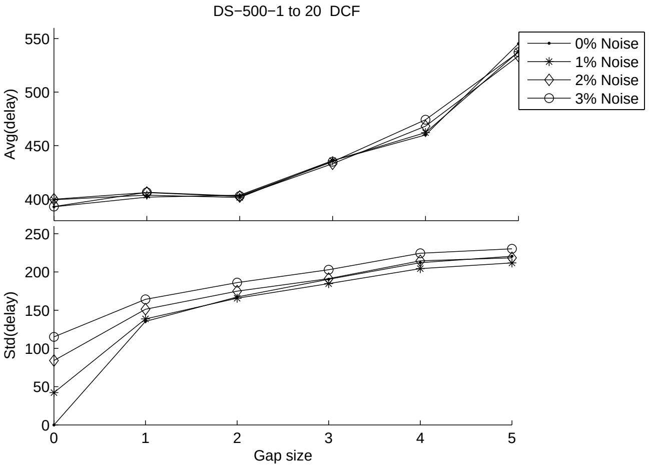


Fig. 5. Results of the application of the DCF method (in §3.1) on all artificial data sets (see §4). Details are in §5.

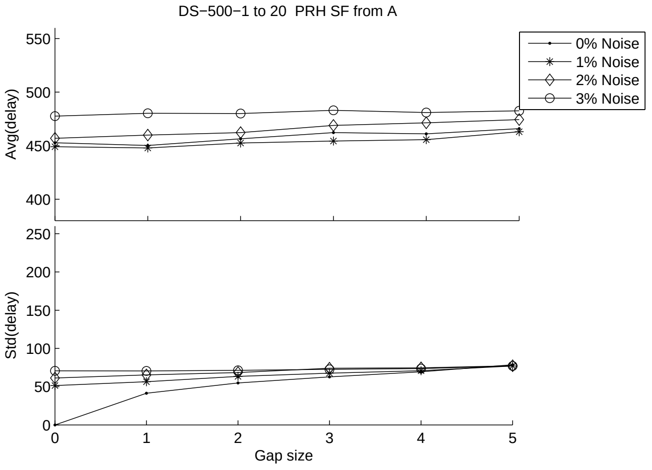


Fig. 7. Results of PRH method with structure function from A image (in §3.2) on all artificial data sets (see §4). Details in §5.

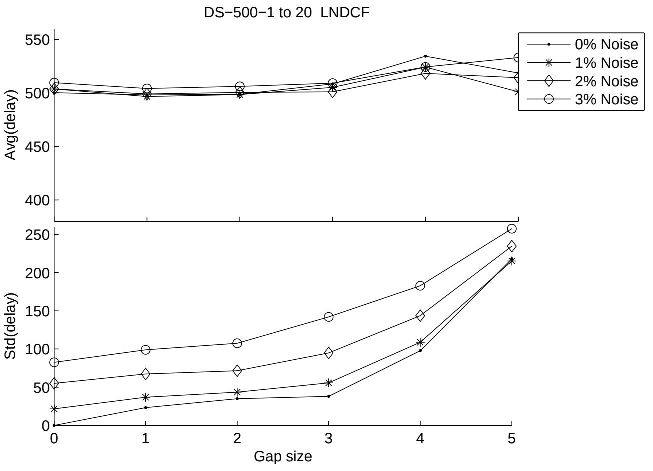


Fig. 6. Results of the application of the LNDCF method (in §3.1) on all artificial data sets (see §4). Details are in §5.

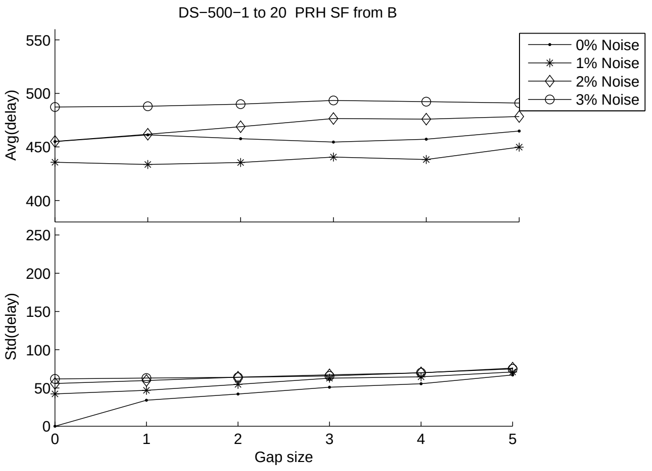


Fig. 8. Results of the application of the PRH method with structure function from B image (in §3.2) on all artificial data sets (see §4). Details are in §5.

and $_{max}$ = 500, with increments of 1 day. As in the previous section, we use a threshold of 0.001 when regularising matrix inversion through SVD. The noise model is assumed to be zero mean i.i.d. Gaussians with standard deviation of 2% of the observed flux value.

For the fixed kernel width technique (§2.2.1), we use Algorithm 1 with the following parameters: $LowerBound$ = 100 and $UpperBound$ = 1200 with increments of 1 day, The selected kernel widths (ω) were 481 and 488 days, and the estimated time delays were 409 days and 459 days for the 4 cm and 6 cm bands, respectively. To calculate confidence intervals on our time delay estimates, we performed 500 Monte Carlo simulations by adding noise realisations to the observed data. Confidence intervals were determined as standard deviations of time delay estimates across the Monte Carlo samples. We found delays of 408±10 days and 460±18 days for 4 cm and 6 cm respectively. Flux reconstructions with these time delays are shown in in Figs. 11 a) and 11 b).

For the variable kernel width method (§2.2.2), the number of neighbors k determining local kernel widths was estimated by Algorithm 1 (ω is replaced by k) with $LowerBound = 1$ and $UpperBound = 15$ (increments of 1). We obtained $k = 3$ for 4 cm, and the estimated time delay was 405 days. Confidence interval computed on 500 Monte Carlo samples was 404.8±11. Flux reconstructions with this time delay are shown in Fig. 11 c). For 6 cm data, we found $k = 3$, and the delay of 450 days. The 500 Monte Carlo samples gave us a time delay of 451.1±30 days. Flux reconstructions are presented in Fig. 11 d).

Using the PRH method, Haarsma et al. (1999) report time delays of 397±12 and 452^{+14}_{-15} days for the 4 cm and 6 cm data, respectively, and 409±30 on the combined 4+6 cm data set. They also report results of the Dispersion spectra method: 383^{+15}_{-19} and 416^{+22}_{-24} days for the 4 cm and 6 cm data, respectively, and 395^{+13}_{-15} days on the combined 4+6 cm data set.

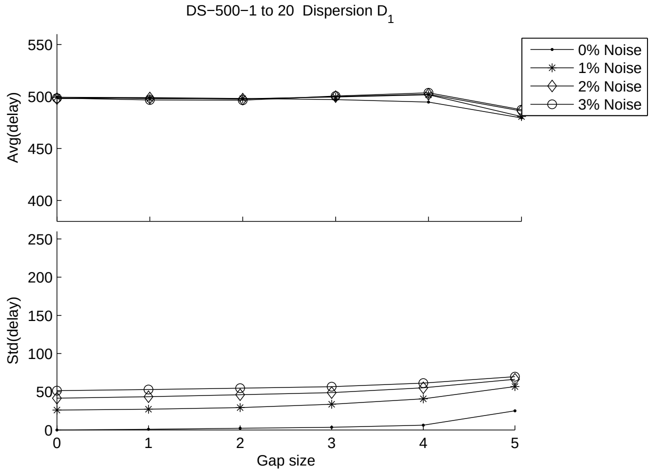


Fig. 9. Results of the application of the Dispersion spectra method D_1^2 (in §3.3) on all artificial data sets (see §4). Details are in §5.

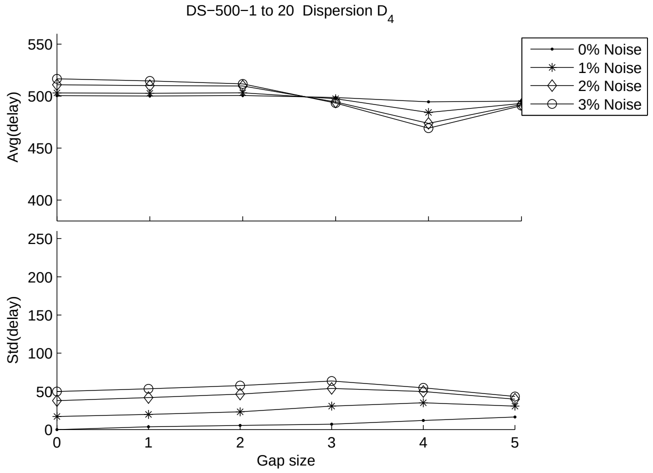


Fig. 10. Results of the application of the Dispersion spectra method D_4^2 (in §3.3) on all artificial data sets (see §4). Details are in §5.

There has been a great deal of concern about the difference in time delay estimates from the two different wavelengths, since gravitational lensing is achromatic. Inspired by the results from our experimentation with artificial data, where the uncertainty of time delay estimates increases as the gap size increases, we have generated a new data set, 6 cm*, in order to avoid the effect of the different gap sizes for different wavelengths. The 6 cm* data set contains 6 cm observations sampled only at observation times of the 4 cm dataset. In other words, we keep a 6 cm observation at time t if there is a 4 cm observation at the same time t. The 4 cm and 6 cm* data sets both contain 58 observations.

Time delay estimates obtained by our methods on 500 Monte Carlo samples based on the 6 cm* data set are presented in Table 3. The 'optimal' kernel parameters, $\omega = 528$ and $k = 5$, are obtained following the procedure described above. The estimated time delays are 405 days and 412 days for the fixed kernel width and variable kernel width methods, respectively. The resulting flux reconstructions are shown in Figs. 11 e) and 11 f).

Table 3. The time delay between Q0957+561 A & B estimated from radio "light" curves at 4 cm and 6 cm.

Kernel method:	fixed width	variable width
4 cm	408.3±10	404.8±11
6 cm	459.9±18	451.1±30
6 cm*	405.3±29	412.6±35

Note: The time delays are in days.
The construction of the **6 cm*** sample, which contains only the 6 cm observations that have a corresponding 4 cm observation at the same epoch, is described in §6.

Table 4. Results from experiments with artificial data sets.

Method	**Figure**
Kernel method with fixed width	Fig. 3
Kernel method with variable width	Fig. 4
DCF and LNDCF	Figs. 5 and 6
PRH method, Structure function from A	Fig. 7
PRH method, Structure function from B	Fig. 8
Dispersion spectra (D_1^2 and $D_{4,2}^2$)	Figs. 9 and 10

It is evident that the large difference in the estimate in time delay seen in several analyses of the same observed data sets that we analyse in this paper is due to the presence of the gaps in the monitoring at the two wavelengths. Such gaps are unavoidable in realistic long-term observing programmes. However, this is the first attempt at quantifying the effect such gaps have on the time delay estimates, leading to unacceptably deviant time delays (in this case, too large by more than 10%). On comparing the 4 cm and 6 cm* samples, which are pairs of the observations at the same epoch and thus have identical gaps in the time series, we find a consistent value for the estimated time delay.

7. Conclusions

We have introduced a novel way of measuring the time delay between light curves of two images of a gravitationally lensed system, based on generalised linear regression with fixed- and variable-width Gaussian basis functions (Kernels) (see §2.2.1 and §2.2.2). On a large set of controlled experiments using artificially generated data, we compare the accuracy of our methods with that of other methods used in the literature for time delay estimation, notably the DCF, LNDCF, PRH and Dispersion spectra methods (see Table 4).

Running a controlled set of experiments is essential for a well-grounded comparison of competing models. For the artificial data, unlike in the case of observed fluxes, we have the luxury of knowing exactly the magnification ratio M and the time delay ; the noise process is also known. Therefore, we can reliably measure the bias ($_\mu$, top of Figs. 3–10) and variance ($_\sigma$, bottom of Figs. 3–10) of the time delay estimates given by the studied methods. Obviously, one can never fully measure the bias when estimating the time delay from real observations. On the artificial data, our kernel-based methods presented in this paper came across as the most accurate and stable methodologies for estimating the time delays between multiple images of a gravitationally lensed quasar.

Previous attempts at generating similar artificial data have tried to simulate specific data sets (see Pijpers 1997; Burud et al. 2001). Our artificial data sets contain simulated light curves of widely varying (but still realistic) shapes, observational gaps and noise levels (these can be made available on request– see more plots at http://www.cs.bham.ac.uk/~jcc/artificial). At the bottom of Figs. 3–10, we can observe a general trend of increased uncertainty as the gap size increases. The uncertainty is also proportional to the noise level.

Our methods for estimating the time delay introduced in sections §2.2.1 and §2.2.2, give similar results (see Figs. 3 & 4), although the variable kernel width method tends to require less computational time.

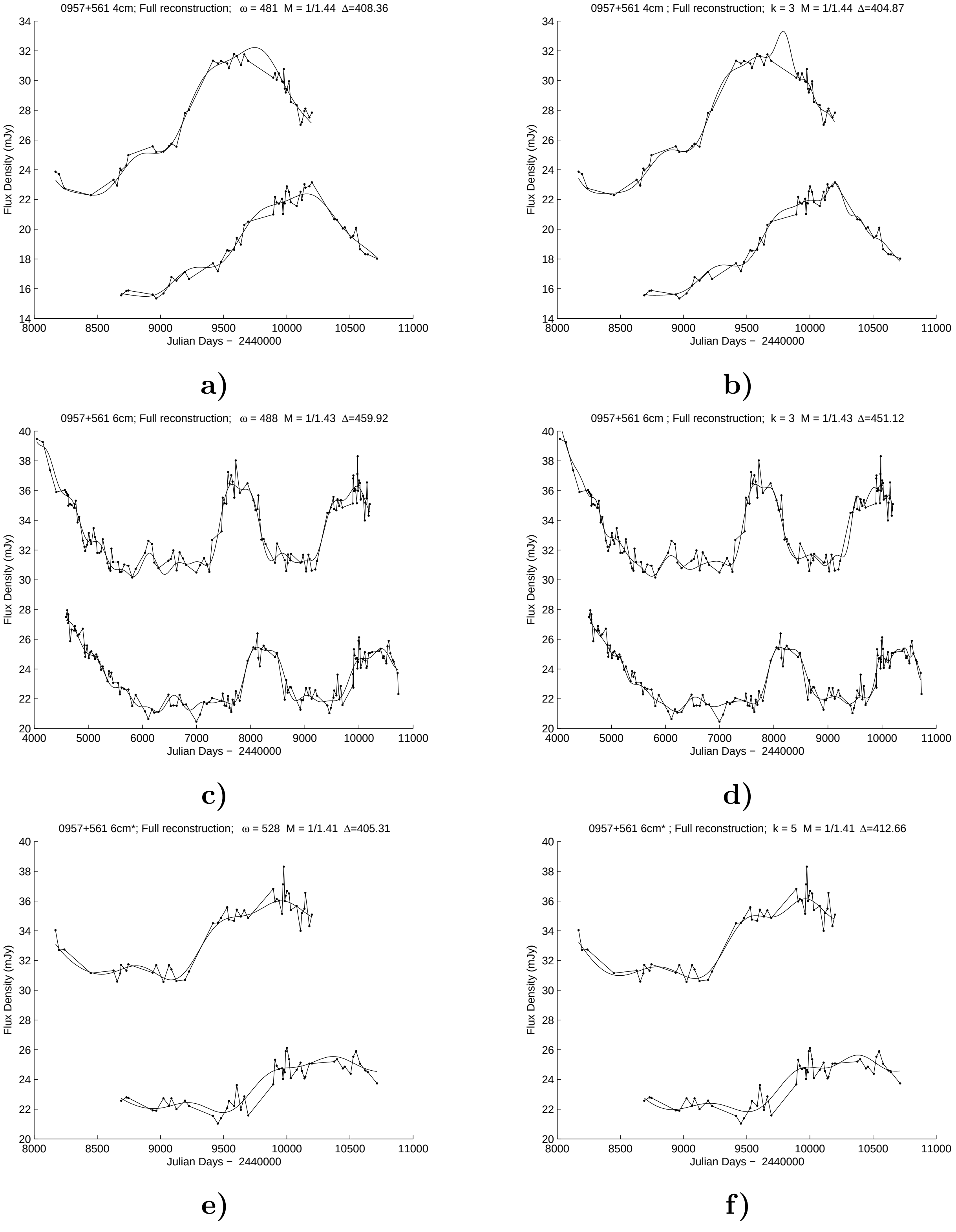
Finally, we have estimated the time delay between of two images of the quasar Q0957+561 from radio observations at 4 cm and 6 cm. The time delay estimates given by our methods are in the range of 405–412 days (see Table 3), which are lower than, but consistent with, most of the other estimates from the same or similar radio data sets (see §6). However, the 6 cm* data set, which by construction includes only observations that are performed at the same epoch as those in the 4 cm data set, yields essentially the same value for the time delay at that obtained from the 4 cm data set (these can be combined to yield 408 ± 12 days, the errors being lower for just the 4 cm data), as opposed to a value of ~450 days as obtained from the full 6 cm data set, which covers a longer monitoring period.

We conclude that such systematic differences between results obtained from observations at various wavelengths are due to the irregular sampling, and in particular, due to the presence of large gaps in the monitoring data. Experiments with simulated data sets lie ours help in the understanding of how the results depends on the sampling, and in assessing the reliability of the time delays obtained by various methods.

Acknowledgements. Juan C. Cuevas-Tello thanks PROMEP and the UASLP, his sponsors.

References

Burud, I., Magain, P., Sohy, S., & Hjorth, J. 2001, A&A, 380, 805
Colley, W., Schild, R., Abajas, C., et al. 2003, A&A, 587, 71
Edelson, R. & Krolik, J. 1988, ApJ, 333, 646
Haarsma, D., Hewitt, J., Lehar, J., & Burke, B. 1997, ApJ, 479, 102
Haarsma, D., Hewitt, J., Lehar, J., & Burke, B. 1999, ApJ, 510, 64
Hastie, T., Tibshirani, R., & Friedman, J. 2001, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer)
Haykin, S. 1999, Neural Networks: a Comprehensive Foundation (Prentice Hall)
Kochanek, C. & Schechter, P. 2004, Carnegie Observatories Astrophysics Series, 2
Kundic, T., Turner, E., Colley, W., et al. 1997, ApJ, 482, 75
Larsen, R. & Hansen, P. 1997, A&A, 121, 587
Lehar, J., Hewitt, J., Roberts, D., & Burke, B. 1992, ApJ, 384, 453
Oscoz, A., Alcalde, D., Serra-Ricart, M., et al. 2001, ApJ, 552, 81
Oscoz, A., Mediavilla, E., Goicoechea, L., Serra-Ricart, M., & Buitrago, J. 1997, ApJ, 479, L89
Ovaldsen, J., Teuber, J., Schild, R., & Stabell, R. 2003, A&A, 402, 891
Pelt, J., Hjorth, J., Refsdal, S., Schild, R., & Stabell, R. 1998a, A&A, 337, 681
Pelt, J., Kayser, R., Refsdal, S., & Schramm, T. 1996, A&A, 305, 97
Pelt, J., Refsdal, S., & Stabell, R. 2002, A&A, 389, L57
Pelt, J., Schild, R., Refsdal, S., & Stabell, R. 1998b, A&A, 336, 829
Pijpers, F. 1997, MNRAS, 289, 933
Press, H., Flannery, B., Teukolsky, S., & Vetterling, W. 1986, Numerical Recipes (Cambridge University Press)
Press, W., Rybicki, G., & Hewitt, J. 1992, ApJ, 385, 404
Rabello-Soares, M., Basu, S., & Christensen-Dalsgaard, J. 1999, MNRAS, 309, 35
Refsdal, S. 1964, MNRAS, 128, 307
Refsdal, S. 1966, MNRAS, 134, 315
Rybicki, G. & Press, W. 1992, ApJ, 398, 169
Saha, P. 2004, A&A, 414, 425
Shawe-Taylor, J. & Cristianini, N. 2004, Kernel Methods for Pattern Analysis (Cambridge University Press)



a) b) c) d) e) f)

Fig. 11. Radio observations of gravitationally lensed images A & B of Q0957+561. We show reconstructions with fixed parameters: **a)** 4 cm with fixed width (ω = 481, = 408.3), **b)** 4 cm with variable width (k = 3, = 404.8), **c)** 6 cm with fixed width (ω = 488, = 459.9), **d)** 6 cm with variable width (k = 3, = 451.1), **e)** 6 cm* with fixed width (ω = 528, = 405.3), and **f)** 6 cm* with variable width (k = 5, = 412.6). Within each plot, at the top is the image A and at the bottom is image B. The continuous lines are our reconstructed underlying light curves, $h_A(t_u)$ and $h_B(t_v)$ in Eq. 9.