Improved AGN light curve analysis with the z-transformed discrete correlation function

Tal Alexander

Department of Particle Physics & Astrophysics

Faculty of Physics

Weizmann Institute of Science

POB 26, Rehovot 76100, Israel

Email: tal.alexander@weizmann.ac.il

February 7, 2013

Abstract

The cross-correlation function (CCF) is commonly employed in the study of AGN, where it is used to probe the structure of the broad line region by line reverberation, to study the continuum emission mechanism by correlating multi-waveband light curves and to seek correlations between the variability and other AGN properties. The z-transformed discrete correlation function (ZDCF) is a new method for estimating the CCF of sparse, unevenly sampled light curves. Unlike the commonly used interpolation method, it does not assume that the light curves are smooth and it does provide errors on its estimates. The ZDCF corrects several biases of the discrete correlation function method of Edelson & Krolik (1988) by using equal population binning and Fisher's z-transform. These lead to a more robust and powerful method of estimating the CCF of sparse light curves of as few as 12 points. Two examples of light curve analysis with the ZDCF are presented. 1) The ZDCF estimate of the auto-correlation function is used to uncover a correlation between AGN magnitude and variability time scale in a small simulated sample of very sparse and irregularly sampled light curves. 2) A maximum likelihood function for the ZDCF peak location is used to estimate the time-lag between two light curves. FORTRAN 77 and FORTRAN 95 code implementations of the ZDCF and the maximum likelihood peak location algorithms are freely available (see `http://www.weizmann.ac.il/weizsites/tal/research/software/`).

1 Introduction

The problem of analysing sparse, unevenly sampled light curves is frequently encountered in the study of AGN, notably in line reverberation mapping and in multi-wavelength variability studies. In many cases, despite great observational efforts, the light curves can not be reliably analysed by Fourier inversion or other inversion methods (Maoz, 1994). There is, to date, no satisfactory alternative but to carry the analysis

in the time domain. A widely used tool for extracting information from such light curves is the cross-correlation function (CCF) Blandford & McKee (1982)

$$\mathrm{CCF}(\tau) = \frac{E[(a(t) - E_a)(b(t+\tau) - E_b)]}{\sqrt{V_a V_b}}\,, \tag{1}$$

where a and b are the two light curves, τ the time-lag, E the expectation value and V the variance over the light curve.

The CCF of AGN continuum and emission line light curves is used in line reverberation mapping to estimate the size and geometry of the broad line region Peterson (1994, and references therein). The CCF between different AGN continuum wave bands is used for studying the continuum emission mechanism by looking for causal connection between the various wavebands (e.g. Korista et al. 1995). Other uses of the CCF include the determination of the Hubble constant from the time lag between variations of the multiple images of a macro-lensed QSO (e.g. Vanderriest et al. 1989) and the linear reconstruction of discretely sampled light curves, whose input is the time-lag dependence of the auto-correlation function (ACF) Rybicki & Press (1992).

In practice, finite segments of the light curves are sampled discretely with measurement errors, at unevenly spaced intervals. The expectation values and variances of the light curves are unknown and must also be estimated from the observations. Three conditions are implicitly assumed in the process of estimating the correlation function: 1) statistical stationarity, 2) the ergodic assumption (i.e. that the ensemble average of all the light curves, which are statistically similar to the observed one, is equivalent to the time average over a single infinite light curve) and 3) random sampling of the light curves. If these are satisfied, then the data at hand can indeed yield an estimate for the population correlation between signals that are statistically similar to the observed light curves. The observed data are the result of a two level sampling process. First, Nature 'draws' a light curve from the continuum of possible light curves. Second, the observations sample the continuum of points in a finite segment of this given light curve (See Fig. 1). Conditions (1) and (2) relate to first sampling level. They must usually be *assumed*, since they cannot be inferred from the observed data. The ZDCF (like the two other CCF estimators that are described below) deals only with the second level of sampling, where the parent distribution is the continuous distribution of the points in the given finite segments.

There are in general two approaches for dealing with the uneven sampling: interpolation and the discrete correlation function (DCF). The interpolation method of Gaskell & Peterson (1987) calculates the CCF by averaging the cross-correlation obtained by pairing the observed $a(t_i)$ with the interpolated value $b(t_i - \tau)$, and that obtained by pairing the observed $b(t_j)$ with the interpolated $a(t_j + \tau)$. The DCF method Edelson & Krolik (1988, EK) initially uses the observed points to calculate

$$\mathrm{DCF}_{ij} = \frac{(a_i - \bar{a}')(b_j - \bar{b}')}{s'_a s'_b} \tag{2}$$

where a_i, b_j are the observed fluxes at times t_i and t_j, respectively, and $\bar{a}'$, $\bar{b}'$, s'^2_a and s'^2_b are the sample means and variances over the entire light curves. Subsequently, $\{\mathrm{DCF}_{ij}\}$ are binned by their associated time-lag, $\tau_{ij} = t_i - t_j$, into equal width bins,

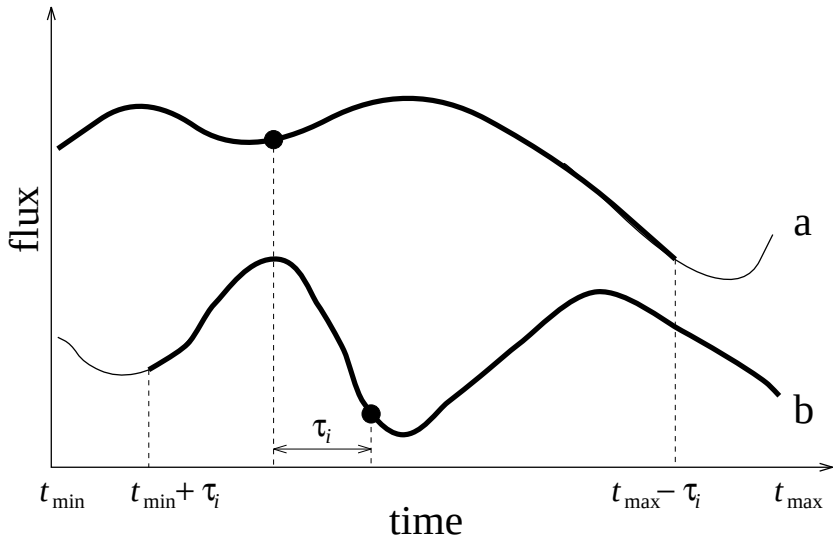

Figure 1: The two observed points, with time lag τ_i, sample the correlation at this lag between the bold intervals of the two continuous segments of the light curves.

$\tau \pm \delta\tau$. The bin average is used to estimate CCF(τ) and the bin's standard deviation to estimate the error.

The two methods were compared by several authors Rodriguez-Pascual, Santo-Lleó & Clavel (1989); White & Peterson (1994) in the context of AGN variability studies, and the general conclusion seems to be that the interpolation method is equal or superior to the DCF in most cases. Nevertheless, interpolation has several obvious drawbacks. Adding interpolated points between those actually observed amounts to inventing data or assuming that the light curve varies smoothly. The interpolation method does not give error estimates on the reconstructed CCF, which are necessary for model fitting. The DCF concept is a more cautious approach to reconstructing the CCF. However, the original DCF method has several problems, which are discussed in detail below. These problems can have adverse effects on the DCF in the realistic case of poorly and unevenly sampled light curves.

This work does not attempt to deal with the many potential pitfalls in the statistical interpretation of the CCF and its significance, which may arise from incorrect assumptions of stationarity or ergodicity or from an incorrect measurement error model (e.g. Box & Newbold 1971). The use of the CCF in AGN light curve analysis is dictated by the present-day quality of the data and the nature of current AGN research. In this given situation, it is important to try and develop more reliable methods for estimating the CCF. The validity of the underlying assumptions in the case of AGN will be ultimately justified if the accumulated results from many observation campaigns converge into a coherent physical picture.

The approach taken here is empirical, in that analytical arguments and approximate assumptions are used only as motivation for the proposed improvements, whose final

test is by simulations with AGN-like light curves. Care is taken to verify that in cases where the assumptions do not hold, the CCF estimate is conservative rather than deceptively significant. Several changes in the original DCF method are suggested, the major ones being the use of the z-transform and equal population binning. These result in the z-transformed discrete correlation function (ZDCF), which is an improved, more robust method for reconstructing the CCF under realistically unfavourable conditions.

The ZDCF method is described and contrasted with the DCF in section 2. Their performance is compared by simulations in section 3. Section 4 presents two sample applications of the ZDCF. The usefulness of the ZDCF in extracting information from a small number of observations is demonstrated in section 4.1, where it is used to find a correlation between the variability time scale and magnitude in a small sample of simulated AGN light curves. Section 4.2 presents a maximum likelihood method for estimating the time-lag between two light curves. The results are discussed and summarised in section 5.

2 The z-transformed Discrete Correlation Function

2.1 Estimating the correlation with the z-transform

Let n be the number of $\{a_i, b_i\}$ pairs in a given time-lag bin. CCF(τ) is estimated by the correlation coefficient

$$r = \frac{\sum_i^n (a_i - \bar{a})(b_i - \bar{b})/(n-1)}{s_a s_b} \,. \tag{3}$$

where $\bar{a}$, $\bar{b}$ are the estimators of the bin averages, and s_a, s_b the estimators of the standard deviations

$$s_a^2 = \frac{1}{n-1} \sum_i^n (a_i - \bar{a})^2 \,, \tag{4}$$

with s_b^2 is similarly defined. Note, in contrast, that the DCF is not normalized correctly since the moments of the full light curves, which appear in DCF_{ij} (equation 2), are used to normalize the individual bins. The correct normalization, as well as ensuring that $|r| \leq 1$, reduces the errors when the light curves are non-stationary by limiting the summation to those points that actually contribute to the bin White & Peterson (1994). The sampling distribution of r is known to be highly skewed and far from normal and therefore estimating its sampling error by the sample variance s_r can be very inaccurate. When a and b are drawn from the bivariate normal distribution it is possible to transform r into an approximately normally distributed random variable, Fisher's z (e.g. Kendall & Stuart 1969, p. 390, Kendall & Stuart 1973, p. 486 and references therein). Defining

$$z = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right) \,, \quad \zeta = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right) \,, \quad r = \tanh z \,, \tag{5}$$

the mean and variance of z are approximately equal to

$$\bar{z} = \zeta + \frac{\rho}{2(n-1)} \times \left[1 + \frac{5+\rho^2}{4(n-1)} + \frac{11+2\rho^2+3\rho^4}{8(n-1)^2} + \cdots\right], \quad (6)$$

and

$$s_z^2 = \frac{1}{n-1}\left[1 + \frac{4-\rho^2}{2(n-1)} + \frac{22-6\rho^2-3\rho^4}{6(n-1)^2} + \cdots\right], \quad (7)$$

where ρ is the unknown population correlation coefficient of the bin. In order to estimate $\bar{z}$ and s_z, the *ansatz* $\rho = r$ is assumed. Transforming back to r, the interval corresponding to the normal $\pm 1\sigma$ error interval can be estimated by

$$\delta r_{\pm} = |\tanh(\bar{z}(r) \pm s_z(r)) - r|. \quad (8)$$

The validity of this *ansatz* was verified by simulations with normal bivariate distributions of different correlation coefficients and n = 11. The results show that the probability of ρ lying outside the empirical interval $r \pm \delta r_{\pm}$ is 1.03 to 1.08 times higher than the normal value of 0.3174, implying a slight under-estimation of the errors. Unlike the $r \pm s_r$ error interval of the DCF, the z-transformed error interval satisfies $|r \pm \delta r_{\pm}| \leq 1$. As $|r|$ tends to 1 it becomes both asymmetric around r and smaller than $r \pm s_r$, thus improving the error estimates in the physically interesting extrema of the CCF. The normality of z is known to hold down to $n = 11$ for an exact normal bivariate distribution or for $\rho = 0$ and otherwise tends asymptotically with n towards normality. As will be discussed below, the dependence of the z-transform on the shape of the parental distribution plays an important role in determining the properties and limitations of the ZDCF. The effects of deviations from binormality, and especially from mesokurtosis, can be significant Gayen (1951). While the resulting bias in $\bar{z}$ is moderate and tends to zero as n increases, the bias in s_z may be non-negligible and its behaviour as n increases varies depending on the underlying bivariate distribution. Nevertheless, the $n = 11$ lower limit still holds as long as the deviations from binormality are moderate. There exist also further refinements of the z transform, z^* and z^{**} Hotelling (1953). Their effect on the ZDCF was checked by simulations, similar to those described in section 3 below and the improvement in the results was found to be negligible and not worth the added calculational complexity.

2.2 The binning method

The ZDCF binning method differs from that of EK in both the binning criterion and the treatment of interdependent pairs in the bin. The ZDCF bins by equal population and as a result the bins are not equal in time-lag width. Each bin contains at least $n_{\min} = 11$ points, (the minimum for a meaningful statistical interpretation) and does not contain interdependent pairs, which are discarded. Consequently, light curves of less then 12 observations cannot be analysed by this method.

2.2.1 Statistical properties

The main source of bias in estimating the bin's correlation coefficient is the existence of inner correlations within the sets $\{a_i\}$ and $\{b_i\}$. The mean time between the observations in a bin associated with time-lag τ is

$$\overline{\Delta t}_\tau = \frac{T-\tau}{n_{\min}} \tag{9}$$

where T is the duration of the two observed light curves (for simplicity, it will be assumed from this point on that the two light curves were observed over the same period). If the light curve is significantly auto-correlated over a coherence time scale τ_0, then when $\overline{\Delta t}_\tau < \tau_0$, the sampling is no longer random since consecutive points in the bin are not independent[1]. $\overline{\Delta t}_\tau$ should not be confused with the mean time between observations, $\overline{\Delta t} = T/(n_{\rm obs}-1)$. $\overline{\Delta t}_\tau$ is not a function of the number of observations, $n_{\rm obs}$, and therefore equal population binning automatically stabilises the results to changes in $n_{\rm obs}$.

It is instructive to discuss the effect of auto-correlation on the z transform in terms of its marginal parental distributions. The irregularly spaced observation times, $\{t_i\}$, can be viewed as sampling the distribution of a random variable t. The observed points are therefore distributed as functions of the random variable t, namely $a(t)$ and $b(t)$. Since these distributions are not necessarily normal, the z transform may be biased, even when the signals are not auto-correlated. The bias caused by auto-correlation has a distinctive signature which can be demonstrated by studying smooth light curves in the limit of short T. In this case the coherence time scales of both light curves are of the order of T itself, the observed segments are almost linear and the distributions of $\{a_i\}$ and $\{b_i\}$ are approximately linear functions of the distribution of t. Suppose further that t is uniformly distributed, so that $kurt(t) = -1.2$. Since the kurtosis is invariant under linear transformations of the random variable, the kurtosis of a and b will also be -1.2. Negative kurtosis in a or b can contribute to significant over-estimation of s_z Gayen (1951, eqs. 81–84), since

$$\begin{aligned} s^2_{z,\,\rm bias} &= s^2_{z,\,\rm true} - \\ &\quad \frac{(n-3-\rho^2)}{(n-1)^2}\,\frac{\rho^2(kurt(a)+kurt(b))}{4(1-\rho^2)^2} + \cdots \end{aligned} \tag{10}$$

This qualitative explanation of the tendency of the ZDCF to over-estimate the errors of auto-correlated signals is supported by the simulations in section 3. The DCF is also known to suffer from this bias.

Astronomical observations are often clustered on daily or seasonal timescales. One simple model for this distribution is that of N periods, each divided into two sub-periods. At the first sub-period, of length T_1, the object can be observed, and the observing times are uniformly distributed. At the second sub-period, of length T_2, the

[1]The coherence time scale which is relevant here is that of finite segments of length T and not that of infinite light curves, which may be larger. This distinction is relevant, for example, for light curves with power-law power spectra, where τ_0 increases with T as the lower frequency, higher amplitude variations become more dominant.

object cannot be shown that the kurtosis of this distribution is also negative, $kurt = -1.2F(N, T_2/T_1)$, where the function F is bracketed between 1 and 5/3 and approaches 1 very rapidly with increasing N. Such sampling patterns will therefore also lead to over-estimation of the sampling errors.

The robustness of $\bar{z}$ and s_z to deviations from the assumed binormality of the parental distributions determines their usefulness in analysing auto-correlated light curves. The nature and magnitude of these deviations depend on the ratio $\overline{\Delta t}_\tau/\tau_0$ and on the specific statistical properties of the light curve. The simulations below show that in the case of AGN-like data sets, these deviations are still small enough for the z-transform to be of practical use.

Another source of bias is the occurrence of interdependent pairs, such as $\{a_i, b_j\}$ and $\{a_i, b_k\}$, in the same bin. If, for example, a and b are fully correlated at time-lag τ, $a(t) \propto b(t+\tau)$, then $r(\tau)$ will be biased towards zero because the repeated occurrences of the point a_i are equivalent to substituting the true a by a constant light curve. Although the frequency of these multiple occurrences can be lowered by decreasing the bin width, this is inconsistent with the requirement that the bin contain at least 11 points. The ZDCF addresses this problem directly by discarding the interdependent pairs. This is to be contrasted with the suggestion of EK that the effect of these pairs can be taken into account by heuristically increasing the DCF error estimates beyond the actual scatter in the bin.

The binned correlation coefficient associates with the mean bin lag, $\bar{\tau}$, an estimate of the bin average of the correlation coefficient. This is not the same as estimating $\rho(\bar{\tau})$. The approximation involved in using the binned average and its interpretation are discussed in appendix A.

2.2.2 The binning algorithm

The binning is implemented by ordering all the possible pairs $\{a_i, b_j\}$ by their associated time-lag $\tau = t_i - t_j$. The ordered list is then divided, bin by bin, into bins of $n_{\min}$ pairs. In the process of adding pairs to a bin, a new pair whose a or b points have previously appeared in that bin, is discarded. The artificial separation of pairs of very close time-lags into adjacent bins is prevented by defining a small parameter ϵ so that a new pair with time-lag τ_{i+1} will be added to the bin as long as $\tau_{i+1} - \tau_i < \epsilon$, even if $n_{\min}$ is exceeded. Most of the information in the ZDCF is at the time lags where the overlap between the two light curves is large (cf. Fig. 6). It is therefore desirable that the binning algorithm minimize the bin widths at these lags. This can be achieved by starting the allocation of pairs to bins at the lag where the pairs are densest. For the auto-correlation function, this means that the binning proceeds from $\tau = 0$ up to $\tau_{\max}$. For the CCF, the binning proceeds from the median τ up to $\tau_{\max}$ and then from the median τ down to $\tau_{\min}$.

The allocation of pairs to bins depends on the order in which the binning is carried out and on the choice of the interdependent pairs which are discarded. While this does not affect the average statistical properties of this method, different choices may lead to different ZDCF results for a given poorly sampled light curve. This ambiguity can be resolved when the ZDCF points are used for fitting a model CCF. In this case the fit score can be averaged over the possible binning choices.

2.3 Measurement errors

The measured light curves a, b are the sum of the true light curves x, y and random noises n_a, n_b

$$a(t_i) = x(t_i) + n_a(t_i) \quad , \quad b(t_j) = y(t_j) + n_b(t_j)\,. \tag{11}$$

As the measurement errors propagate into the CCF in a complicated, non-linear manner, it is suggested that their effect be estimated by Monte Carlo simulations. This is performed by assuming a distribution for the measurement errors (usually the normal distribution), adding a randomly drawn error value to each point and recalculating the ZDCF. The Monte Carlo average of z is then used to obtain r and $\delta r_{\pm}$. The resulting ZDCF is conservative in that it estimates the mean CCF that is *consistent* with the observations and $\sqrt{2}$ times the quoted measurement errors. The over-estimated errors are an unavoidable consequence of having no model for the light curve but the observed points themselves. The errors on the ZDCF points are the sampling errors and should not be interpreted as estimating the deviation from the CCF of the true signals due to the measurement errors. For example, the ZDCF of very noisy light curves tends to zero irrespective of the correlation between the true signals. It should be emphasized that the data points are not weighted by their errors. It is therefore recommended that points with atypically high measurement errors (relative to the mean error) be judiciously discarded from the light curve before the ZDCF is applied.

This approach to estimating the effects of the measurement errors is safer than that suggested by EK, which diverges in the limit of a small number of observations. This is discussed in more detail in appendix B

3 Results

The properties of AGN power spectra are currently well known only in the X-ray range, where the power spectrum can be approximated by a power law, $P(\nu) \propto \nu^{-\alpha}$, with $1 \lesssim \alpha \lesssim 2$ Green, M$^{\rm c}$Hardy & Lehto (1993). In order to compare the performance of the DCF and ZDCF on AGN-like light curves, the two methods were applied to two types of simulated light curves (Fig 2): Noisy light curves with a $P(\nu) \propto \nu^{-1}$ power spectrum (flicker noise) and softer light curves with a $P(\nu) \propto \nu^{-2}$ power spectrum. Both were generated by choosing random phases for frequencies in the range $\nu_{\rm min} = 1/4$, $\nu_{\rm max} = 100$ at increments of $\Delta\nu = 1/4$ (in arbitrary inverse time units). A light curve was then calculated from $t = 0$ to 1.

The simulations consisted of drawing eight sets of 100 light curves, one for each of the two power spectra and $n_{\rm obs} = 15, 20, 25$ and 30. For each of these light curves, the times of the $n_{\rm obs}$ simulated observations were randomly drawn so that the time difference between two successive observations was uniformly distributed. The light curve was then sampled at these random times and the discrete auto-correlation function was calculated by both the DCF and ZDCF methods. In order to have a common basis for comparing the two methods, the number of bins used by the DCF was adjusted to equal the mean one used by the ZDCF. The ZDCF was calculated with $n_{\rm min} = 11$ which resulted in $\langle n_{\rm bin} \rangle = 2, 6, 14$ and 22 for $n_{\rm obs} = 15, 20, 25$ and 30, respectively. In order

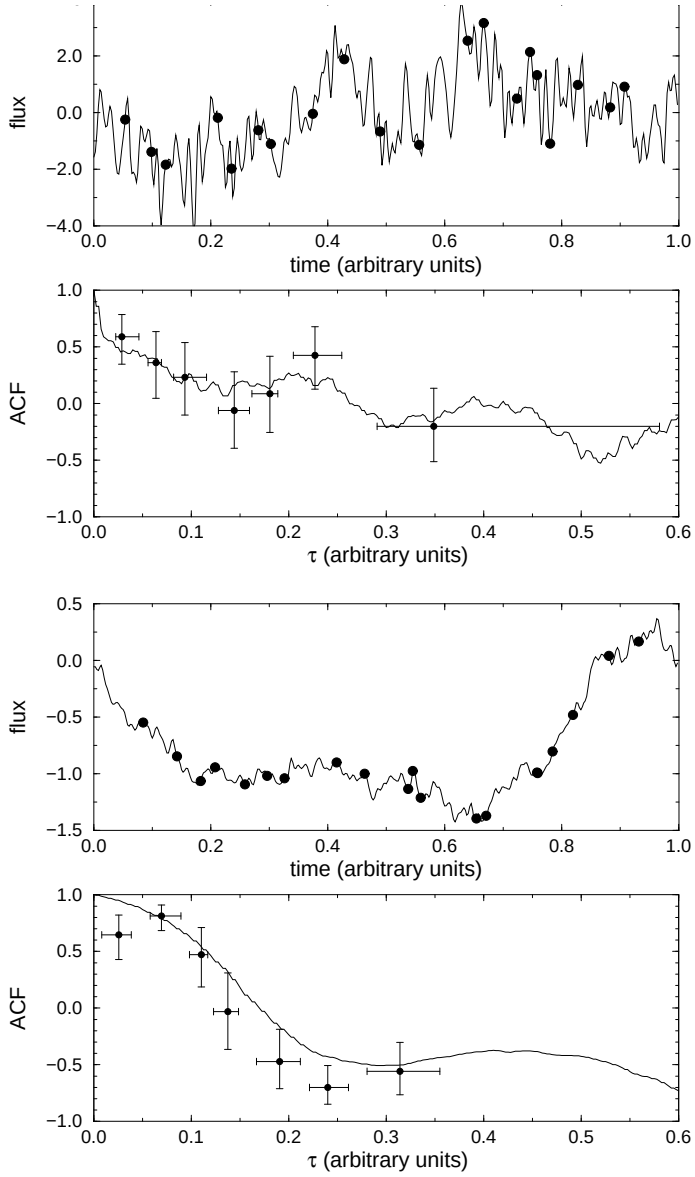

Figure 2: Two examples of simulated light curves and their ZDCF. The time differences between each successive pair of the 20 observations (black dots) are drawn from the uniform distribution. No measurement errors added. Top 2 panels: flicker-noise light curve ($P(\nu) \propto 1/\nu$) and its ACF. Bottom 2 panels: soft spectrum light curve ($P(\nu) \propto 1/\nu^2$) and its ACF. The simulated light curve is shown in solid line and the discrete observations in dots. The true ACF is shown in solid line and the auto-correlation function calculated by the ZDCF method is shown as points with error bars.

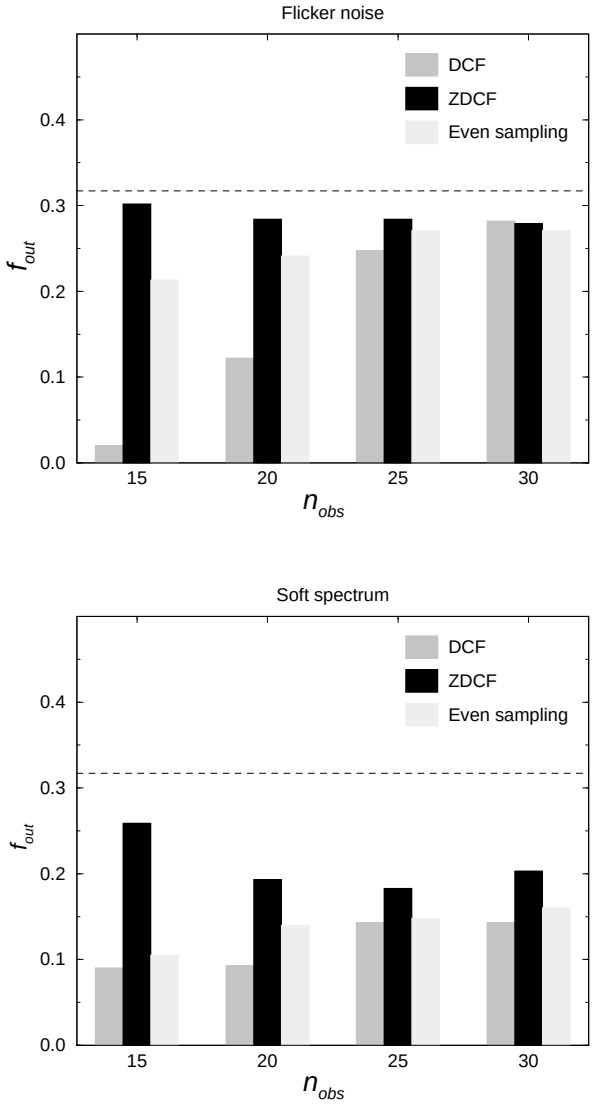

Figure 3: The fraction of outlying points, $f_{\rm out}$ as function of the number of simulated observations, $n_{\rm obs}$. The results are the average over 100 simulations. Top: flicker-noise light curves. Bottom: soft, $P(\nu) \propto 1/\nu^2$, light curves. The even sampled bins were thinned out so as to have only 11 points per bin. No measurement errors added.

to isolate the possible effects of the binning procedure on the results, eight additional sets of simulated light curves were similarly generated and sampled at evenly spaced time intervals. The ZDCF was then calculated after the resulting zero width bins were thinned down to 11 randomly chosen points per bin to avoid possible biases due to the z transform's n dependence. The fit of the ZDCF to the true ACF was evaluated by the fraction of outlying bins, $f_{\rm out}$, whose error interval does not include $\rho_{\rm bin}$, which was calculated from the full light curve for each bin. Similarly, for the DCF the comparison was to $\bar{\rho}$, averaged over the interval $\tau_i \pm \delta\tau$. For each of the two light curve types the average kurtosis of a and b and a χ^2 fit of the marginal distributions to the normal distribution were also calculated to asses the deviations from the idealized assumptions of the z transform.

Figure 3 shows the fraction of outliers, $f_{\rm out}$, as a function of $n_{\rm obs}$ for the two light curve types and for the ZDCF, DCF and evenly sampled ZDCF cases. In all the simulations checked here, both methods *over-estimate* the errors, but the ZDCF results are always closer than the DCF's to the normal value $f_{\rm out} = 0.3173$. There is a marked trend of an increase in the over-estimation for softer power spectra. This is correlated with the deviations from normality. The flicker noise average kurtosis is $kurt = -0.27$ and the χ^2 fit probability for the normality of the marginal distributions is $P(\chi^2) \sim 0.3$ whereas for the soft power spectrum the values are $kurt = -0.80$ and $P(\chi^2) \sim 0.03$, respectively. The evenly sampled ZDCF dependence on the underlying power spectrum is similar to that of the binned ZDCF, confirming that this effect is not due to the binning procedure.

As anticipated by equation 9, the behaviour of the ZDCF, unlike that of the DCF, is not affected by the number of observations. The difference between the two methods is most marked in the case of the flicker noise power spectrum, where the DCF becomes highly biased as the number of observations decreases.

Qualitatively similar results were obtained for simulations performed with Gauss-Markov random signals Brown & Hwang (1992) and signals described by Gaussian shaped peaks of random height and width, randomly superimposed on a high order polynomial.

4 examples

4.1 Correlations involving variability time scales

The physical origin of AGN variability is poorly understood. One line of approach to this problem is to look for correlations between the variability and other AGN properties, such as the luminosity (e.g. Hook et al. 1994). The following example is a simplified, simulated version of such an analysis, which was performed on observed data by Netzer et al. (1996). The ZDCF bias depends on the statistical properties of the light curves, which are generally not known in advance. The following example demonstrates that the reduced bias of the ZDCF, even when it cannot be quantified, makes it more efficient in extracting information from the data than the DCF and thus useful even when applied to sparsely sampled light curves.

The observational situation was simulated as follows. A sample of 30 light curves

with power law power spectra, $P(\nu) \propto \nu^{-\alpha}$ were generated as described in section 3. The 30 light curves consisted of three light curves for each of the 10 values $\alpha = 1.1$, $1.2, \ldots, 2.0$. All the light curves were then sampled at the same 15 times (this simulates the situation where the AGN sample is recorded simultaneously on one photographic plate). The sampling times (Fig. 4) have a clustered pattern typical of astronomical observations and were taken from the observed light curve shown in Fig. 6 (The original light curve was diluted down to 15 points by throwing very close observing times). Normally distributed 3% measurement errors were added to the simulated observations. A toy model for the magnitude–power spectrum relation was used to associate each light curve with a magnitude

$$M(\alpha) = 26 + \alpha + \epsilon\,, \tag{12}$$

where ϵ is a gaussian variate with zero mean and a standard deviation of 0.1. ϵ represents an additional scatter in the AGN luminosity that is independent of the power spectrum. This may be due to measurement errors in M or to a hidden dependence of M on some physical parameters other than α.

The ACF of each of the light curves was estimated by both the DCF and ZDCF. The 15 observations yielded only two ZDCF bins, and accordingly, the DCF was also calculated with only two bins. The coherence timescale, τ_0, is a measure of the typical variability timescale. Following Netzer et al. (1996), τ_0 was formally defined as the shortest lag where $\mathrm{ACF}(\tau_0) = 0$ and was estimated by minimal squares fitting of the straight line $r = 1 - \tau/\tau_0$ through the DCF and ZDCF points. The fit took into account both the δr and $\delta\tau$ errors. Finally, Spearman's rank order correlation coefficient between τ_0 and M was calculated for the 30 AGN. Note that the rank order correlation is insensitive to the exact functional form of $M(\alpha)$, and is only affected by the ratio between the scatter in M due to the distribution of α in the sample and that due to ϵ.

This entire procedure was repeated with 1000 random samples of 30 light curves each. Fig. 5 shows the distribution of the calculated correlation coefficient and of the proability that the correlation is not random, for both the ZDCF and DCF. For power law power spectra, τ_0 increases with α and therefore correct estimates of τ_0 should yield significant positive correlation coefficients. Fig. 5 shows that the ZDCF is much more efficient than the DCF and has a significantly better chance of uncovering the underlying correlation even in a small sample of sparsely sampled light curves. For example, in 40% of the ZDCF simulations, a positive correlation was detected at the 0.99 significance, as compared to only 6% for the DCF. In 76% of the ZDCF simulations, a positive correlation was detected at the 0.90 significance, as compared to only 23% for the DCF.

4.2 Time lag estimation by maximum likelihood

Figure 6 shows the R and B light curves of the optically violent variable quasar 3C 454 (Netzer et al., 1996) and their ZDCF. The time lag between the two bands, and the question whether it is consistent with zero, are important for testing models of AGN continuum emission. In the following example, the ZDCF is used to estimate the uncertainty in the peak position by using, for the first time in this context, the fiducial interpretation of the likelihood function.

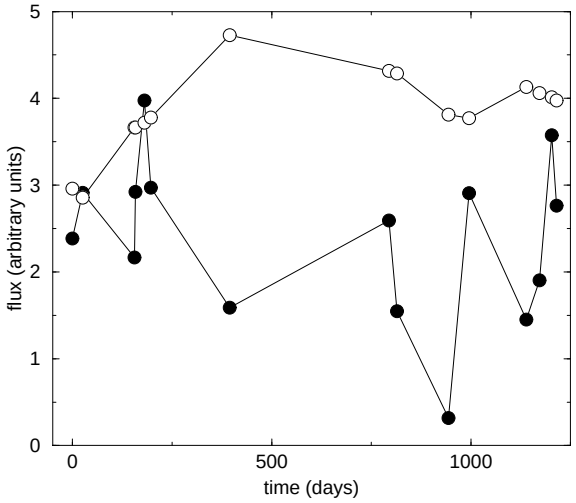

Figure 4: Two of the simulated light curves used for studying the correlation between variability timescale and magnitude. Both light curves are sampled simultaneously 15 times with a clustered sampling pattern and have a power law index of $\alpha = 2$ (open circles) and $\alpha = 1$ (filled circles).

The likelihood that point i is the maximum, L_i, is approximately the product of the probabilities for point i being larger than point j

$$L_i = \int_{-\infty}^{\infty} d\zeta_i \frac{1}{\sqrt{2\pi} s_{z,i}} \exp\left[-\frac{1}{2}\left(\frac{z_i - \bar{z}_i}{s_{z,i}}\right)^2\right] \times \prod_{j \neq i} \int_{-\infty}^{\zeta_i} d\zeta_j \frac{1}{\sqrt{2\pi} s_{z,j}} \exp\left[-\frac{1}{2}\left(\frac{z_j - \bar{z}_j}{s_{z,j}}\right)^2\right], \quad (13)$$

where $\bar{z}$ and s_z are functions of ζ through ρ and z is a function of r, the empirical correlation. The approximation results from neglecting the possible correlations between the ZDCF points (e.g. Box & Jenkins 1970) and from the fact that the ZDCF points are only approximately normally distributed. It is more convenient to express L_i as an integral over ρ

$$L_i = \int_{-1}^{+1} \frac{1}{\sqrt{2\pi} s_{z,i}} \exp\left[-\frac{1}{2}\left(\frac{z_i - \bar{z}_i}{s_{z,i}}\right)^2\right] \times \prod_{j \neq i} \Phi(z_j, \rho_i) \frac{d\rho_i}{(1 - \rho_i^2)}, \quad (14)$$

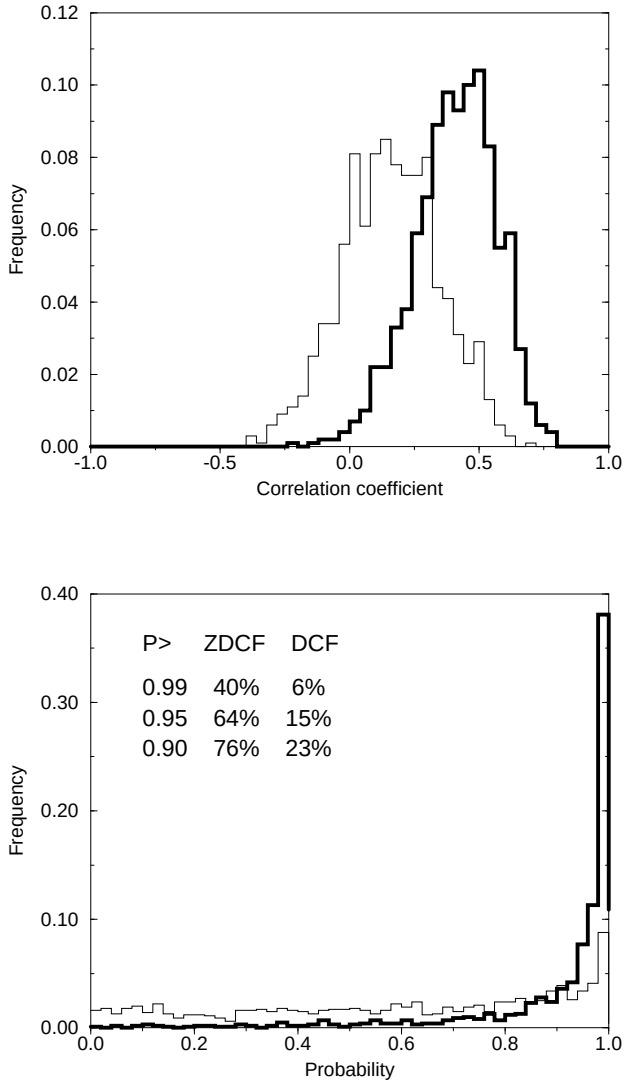

Figure 5: The sample distribution of Spearman's rank order correlation coefficient between τ_0 and M (top) and its probability (bottom) calculated with the DCF (thin line) and ZDCF (bold line). Also listed are the fractions of results with positive correlation and probabilities greater than 0.90, 0.95 and 0.99.

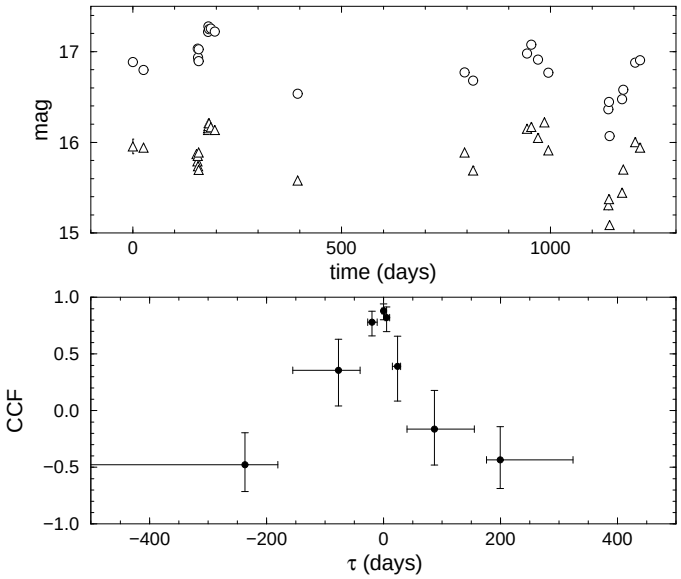

Figure 6: Top: The R band (triangles) and B band (circles) light curves of the optically violent variable quasar 3C 454. The measurement errors are smaller than the point size. Bottom: The central peak of the ZDCF, calculated with $n_{\min} = 11$ and 100 Monte Carlo draws for estimating the effects of the measurement errors.

where

$$\Phi(z,\rho) = \int_{-1}^{\rho} \frac{1}{\sqrt{2\pi}s_z} \exp\left[-\frac{1}{2}\left(\frac{z-\bar{z}}{s_z}\right)^2\right] \frac{d\rho'}{(1-\rho'^2)}\,. \tag{15}$$

L_i can be easily calculated numerically. L_i reaches its maximum at the highest ZDCF point, so that the maximum likelihood (ML) estimate coincides with the ZDCF peak. The sampling distribution of ML estimators is generally known only in the large sample limit, where it is Gaussian. Here, having a large sample means having many pairs of R and B observations of the same continuous light curve, observed simultaneously by different observers. The actual data set consists of only one such pair, and it is therefore impossible to assign a a confidence interval for the peak.

There is an alternative approach for obtaining interval estimates from the likelihood function, which is related, but not identical, to Bayesian statistics. The normalized likelihood function, defined as the *fiducial distribution*, is interpreted as expressing the "degree of belief" in the possible value of the estimated parameter, and the 68% interval around the likelihood function's maximum is defined as the 68% *fiducial interval* (e.g. Frodesen, Skjeggestad & Tøfte 1979). The fiducial function for the peak is calculated by interpolating between the points of the likelihood function. The position of the fiducial distribution's maximum is the maximum likelihood estimate of the time-lag[2]

[2]When the fiducial function is interpolated linearly, the most likely peak position is that of the highest

and the fiducial interval is that which includes 0.3414 (normal 1σ) of the area left and right of the maximum, respectively. In practice, the fiducial interval cannot be narrower than the bin width $\delta\tau_{\pm}(r_{\max})$. The fiducial interval can be interpreted as the interval where 68% of the likelihood-weighted ensemble of all possible CCFs reach their peaks. It should be emphasized that a fiducial interval is conceptually different from a confidence interval, and one should not misinterpret it as meaning that if the light curves are resampled and their ZDCF peak calculated, then on the long run, the true peak will lie inside the fiducial intervals 68% of the time. A detailed discussion of the fiducial interval and its relation to Bayesian statistics can be found in Kendall & Stuart (1973).

The fiducial estimate of the peak has several advantages over other approaches that were used in AGN variability studies. Unlike methods that use simulations to estimate the uncertainty on the time-lag (e.g. Maoz & Netzer, 1989), the fiducial estimator does not assume anything about the mechanism that generates the light curves. Another commonly used method for estimating the peak location is to fit the CCF peak to a peaked function, such as a parabola or a Gaussian. Some parameterization of the function's width is then used as a measure of the uncertainty in the peak position. The arbitrary choice of the function introduces an unwanted degree of freedom to the final result. It is also often the case that the peak, unlike the fitted functions, is asymmetric and this leads to skewed estimates of the peak location. In contrast, The fiducial estimator does not assume anything about the shape of the peak and thus avoids these problems.

These advantages come at the price of having to make the approximations mentioned above in calculating L_i, as well as the approximation involved in interpolating the likelihood function[3]. Another unavoidable consequence of this approach is the assumption of a Bayesian prior, namely the uniform distribution of the CCF points in z-space.

The CCF peak of 3C 454 lies between time-lags $\tau = -200$ to 200 days. The maximum likelihood estimate that is calculated for the points in this interval is $0.1^{+10.0}_{-23.7}$ days (R lags after B). This is consistent with the hypothesis that the R and B bands of 3C 454 vary simultaneously.

5 Discussion and summary

Monitoring light curves of astronomical objects over long periods of time requires great observational efforts. Inevitable limitations and difficulties stand in the way of obtaining regular, well sampled light curves and severely restrict the reliability of spectral analysis. The result is that in many cases the light curves are analysed directly in the time domain by the CCF. Such a situation calls for an effort, on par with the observational one, to develop improved methods for estimating the CCF, even at the cost of added computational complexity. In many cases the only way to extract information from meager data is by modeling. It is therefore important to have error estimates that

ZDCF point.

[3]Note that these approximations are also shared by the function-fitting method, and that methods that use the centroid to parameterize the time-lag also ignore the biases that may be introduced by the correlations

can be used for fitting theoretical CCF models to the data and estimate the location and significance of the CCF extrema.

The ZDCF method attempts to correct the biases that affect the original DCF. The simulations show that the ZDCF performs better or at least as good as the DCF under a variety of conditions. The performance of the ZDCF depends on the ratio between $\overline{\Delta t}_\tau$, the typical time between observations in a bin, and τ_0, the coherence timescale. As long as $\overline{\Delta t}_\tau > \tau_0$, the z-transform biases are small and the error estimates are realistic. Since $\overline{\Delta t}_\tau$ does not depend on $n_{\rm obs}$, it is possible to increase the time lag resolution of the ZDCF without increasing its bias by increasing sampling rate (i.e increasing $n_{\rm obs}$ at a constant total time T). When $\overline{\Delta t}_\tau < \tau_0$, the error estimates of the z-transform may be over-estimated. However, in this case interpolation is justified. This is seen by noting that $\overline{\Delta t}_\tau$ is a decreasing function of τ, which is approximately bounded from below by

$$\overline{\Delta t}_\tau \gtrsim \left(\frac{n_{\rm min}-1}{n_{\rm min}}\right)\overline{\Delta t} \sim \overline{\Delta t}\,. \tag{16}$$

It then follows that if $\overline{\Delta t}_\tau < \tau_0$, then $\overline{\Delta t} < \tau_0$ and therefore that the light curves are well sampled. Even in this case the ZDCF offers a more conservative alternative than interpolation. $\bar{z}$, unlike s_z, is relatively unbiased and therefore $r_{\rm bin}$ estimates the CCF much better than implied by the formal error estimates.

The simulations were limited to the uniform random sampling pattern. Sampling patterns which are far from uniform may introduce complicated biases to the ZDCF, which can be traced to the fact that the Fourier transform of the underlying signal is convolved with that of the sampling pattern itself. This is a fundamental problem that lies beyond the scope of the ZDCF. A partial solution is to divide the light curve into segments which are roughly uniformly sampled and perform the ZDCF on the union of these segments, at the price of estimating the correlation function only for small values of τ.

To summarize, the calculation of the ZDCF involves the following steps:

1. Atypically noisy measurements should preferably be discarded from the light curve .

2. All possible pairs of observations, $\{a_i, b_j\}$, are sorted according to their time-lag $t_i - t_j$ and binned into equal population bins of at least 11 pairs. Multiple occurrences of the same point in a bin are discarded so that each point appears only once per bin (section 2.2).

3. Each bin is assigned its mean time-lag and the intervals above and below the mean that contain 1σ (normal) of the points each (section 2.2).

4. The correlation coefficients of the bins are calculated (equation 3) and z-transformed (equation 5). The error is calculated in z-space (equation 7) and transformed back to r-space (equation 8).

5. The effect of measurement errors is estimated by Monte Carlo runs, where at each step a random error is added to each point according to its quoted error.

The ZDCF is averaged in z-space and the average is transformed back to r-space (section 2.3).

The resulting CCF error bars are roughly equivalent to normal 1σ errors. The simulations performed here show a consistent trend towards over-estimation of the errors and suggest that the error interval may be as large as 1.4σ for strongly auto-correlated light curves. A FORTRAN 77 code of the ZDCF method is available from the author on request.

A Properties of the binned average

The binned correlation coefficient is an average of the correlation coefficient over the bin's time lag interval. The exact nature of this average can be expressed in terms of the mixture of time-lags $\{\tau_i\}$, associated with the observed pairs $\{a_i, b_i\}$ in the bin. Let the weight ω_i be the fraction of the pairs with time-lag τ_i (usually $\omega_i = 1/n_{\min}$). Define $E_a^-(\tau_i)$, $V_a^-(\tau_i)$ to be the mean and variance over the continuous light curve a from $t_{\min}$ to $t_{\max} - \tau_i$, $E_b^+(\tau_i)$, $V_b^+(\tau_i)$ to be the mean and variance over the continuous light curve b from $t_{\min} + \tau_i$ to $t_{\max}$ and $C_{a,b}^{\pm}(\tau_i)$ to be the covariance of the two light curves with the intervals similarly defined (See Fig. 1). It is straightforward to show that the binned variance and covariance over the continuous light curves are

$$\begin{aligned} V_a &= \sum_{i=1}^{n} \omega_i V_a^-(\tau_i) + \sum_{i<j} \omega_i \omega_j (E_a^-(\tau_i) - E_a^-(\tau_j))^2 \,, \\ V_b &= \sum_{i=1}^{n} \omega_i V_b^+(\tau_i) + \sum_{i<j} \omega_i \omega_j (E_b^+(\tau_i) - E_b^+(\tau_j))^2 \,, \end{aligned} \tag{17}$$

and

$$\begin{aligned} C_{a,b} = &\sum_{i=1}^{n} \omega_i C_{a,b}^{\pm}(\tau_i) + \\ &\sum_{i<j} \omega_i \omega_j (E_a^-(\tau_i) - E_a^-(\tau_j))(E_b^+(\tau_i) - E_b^+(\tau_j)) \,. \end{aligned} \tag{18}$$

It is clear that the sample binned correlation coefficient estimates the quantity

$$\rho_{\rm bin} = \frac{C_{a,b}}{\sqrt{V_a V_b}} \,, \tag{19}$$

which is not identical to the simple weighted arithmetic mean

$$\bar{\rho} = \sum_{i=1}^{n} \omega_i \frac{C_{a,b}^{\pm}(\tau_i)}{\sqrt{V_a^-(\tau_i) V_b^+(\tau_i)}} \,, \tag{20}$$

although both values approach each other as the bin width is decreased. r estimates an averaged value of ρ, weighted by the density distribution of the time-lag points in

the bin. In order to reflect this, the time-lag associated by the ZDCF with the bin is estimated by the mean lag $\bar{\tau}$. The uncertainty in τ is estimated by the intervals $\delta\tau_{\pm}$ that contain 0.3414 (normal 1σ) of the points in the bin above and below $\bar{\tau}$, respectively. The resulting error bars are asymmetric and, unlike usual error bars, do not give the $\pm 1\sigma$ uncertainty on the 'true' value $\bar{\tau}$ but rather describe the interval over which $\sim 2/3$ of the averaging was performed. Simulations of sparsely sampled light curves, where $\tau_{\max}$ is small and $E(\tau_i)$, $V(\tau_i)$ and $C(\tau_i)$ are therefore calculated on largely overlapping stretches of the light curves, confirm that $\rho_{\rm bin} \simeq \bar{\rho}$ to a very good approximation.

B The effect of measurement errors on the DCF

EK suggest that the sample CCF of the true signals, $r_{x,y}$, can be recovered from $r_{a,b}$ by a simple analytic correction of replacing equation 3 with

$$r'_{x,y} = \frac{\sum_i^n (a_i - \bar{a})(b_i - \bar{b})/(n-1)}{\sqrt{s_a^2 - s^2(n_a)}\sqrt{s_b^2 - s^2(n_b)}}\,, \tag{21}$$

where $s(n_a)$ and $s(n_b)$ are the typical measurement errors. $|r'_{x,y}| > |r_{a,b}|$ by definition, and in the limit of low S/N, $|r'_{x,y}|$ may become arbitrarily large (even larger than 1), despite the fact that the observed light curves are almost pure noise. The origin of this problem is that this correction applies only in the limit of an infinitely sampled, infinite light curve. This is demonstrated by writing $\rho_{x,x}$ explicitly in the case of auto-correlation. Assuming for simplicity that the bin contains only a single time-lag τ, the correlation coefficient over a continuous finite segment of the light curve is

$$\rho_{x,x}(\tau) = \frac{C^{\pm}_{a,a} - C^{\pm}_{x,n} - C^{\pm}_{n,x} - C^{\pm}_{n,n}}{\sqrt{(V_a^- - V_n^- - 2C^-_{x,n})(V_a^+ - V_n^+ - 2C^+_{x,n})}}\,. \tag{22}$$

The errors are assumed to be uncorrelated with the signal and with themselves, and therefore in the limit of an infinite light curve,

$$C^{\pm}_{x,n}(\tau) = C^{\pm}_{n,x}(\tau) = C^{+}_{x,n}(\tau) = C^{-}_{x,n}(\tau) = 0\,, \tag{23}$$

and for $\tau \neq 0$,

$$C^{\pm}_{n,n}(\tau) = 0 \tag{24}$$

also holds. $r'_{x,x}$ approaches $\rho_{x,x}$ only in this limit. Moreover, even in cases where this is an adequate approximation, it is still unclear whether this estimator has the required statistical properties (i.e. consistency, efficiency and lack of bias).

Acknowledgements

I am grateful to Rick Edelson and Julian Krolik for comprehensive discussions of issues relating to the DCF and ZDCF. Stimulating conversations with Shai Kaspi, Shai Zucker and Eyal Neistein are much appreciated. This work was partially supported by the US-Israel Binational Science Foundation grant no. 89-00179.

References

Blandford R. D., McKee C. F., 1982, ApJ, 255, 419

Box G. E. P., Jenkins G. M., 1970, Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco, p. 376

Box G. E. P., Newbold P., 1971, J. Roy. Stat. Soc., A134, 229

Brown R. G., Hwang P. Y. C., 1992, Introduction to Random Signals and Applied Kalman Filtering, 2nd ed., John Wiley & Sons, New York, p. 134

Edelson R. A., Krolik J. H., 1988, ApJ, 333, 646 (EK)

Frodesen A. G., Skjeggestad O., Tøfte H., 1979, Probability and Statistics in Particle Physics, Universitetsforlaget, Bergen, p. 229

Gaskell C. M., Peterson B. M., 1987, ApJ Suppl., 65, 1

Gayen A. K., 1951, Biometrika, 38, 219

Green, A. R., M^{c}Hardy I. M., Lehto H. J., 1993, MNRAS, 265, 664

Hook, I. M., M^{c}Mahon R. G., Boyle, B. J., Irwin, M. J., 1994, MNRAS, 268, 305

Hotelling H., 1953, J. Roy. Statist. Soc., B15, 193

Kendall M. G., Stuart A., 1969, The Advanced Theory of Statistics, 3rd ed., vol. I, Griffin, London

Kendall M. G., Stuart A., 1973, The Advanced Theory of Statistics, 3rd ed., vol. II, Griffin, London

Korista K. T. et al., 1995, ApJ Suppl., 97, 285

Maoz D., 1994, in Gondhalekar P. M., Horne K., Peterson B. M., eds, Reverberation Mapping of the Broad-Line Region in Active Galactic Nuclei. A.S.P. Conf. Ser., Vol. 69, A.S.P., San Francisco, p. 95

Maoz D., Netzer H., 1989, MNRAS, 236, 21

Netzer H. et al., 1996, MNRAS, 279, 429

Peterson B. M., 1994, in Gondhalekar P. M., Horne K., Peterson B. M., eds, Reverberation Mapping of the Broad-Line Region in Active Galactic Nuclei. A.S.P. Conf. Ser., Vol. 69, A.S.P., San Fransisco, p. 1

Rodriguez-Pascual P. M., Santo-Lleó M., Clavel J., 1989, AA, 219, 101

Rybicki G. B., Press W. H., 1992, ApJ, 398, 169

Vanderriest C., Schneider J., Herpe G., Chevreton M., Moles M., Wlérick G., 1989, A&A ,215, 1

White R. J., Peterson B. M., 1994, PASP, 106, 879