IS AGN VARIABILITY CORRELATED WITH OTHER AGN PROPERTIES?—ZDCF ANALYSIS OF SMALL SAMPLES OF SPARSE LIGHT CURVES

To appear in Astronomical Time Series, 1997, D. Maoz, A. Sternberg, E. Leibowitz, eds, Kluwer, Dordrect

TAL ALEXANDER
School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel

Abstract. A new method for estimating the cross-correlation function (CCF) of light curves, the z-transformed discrete correlation function (ZDCF), is demonstrated by uncovering a noisy correlation between AGN variability timescale and luminosity in a small sample of sparsely sampled light curves.

1. Introduction

The analysis of variability in the frequency domain requires well sampled light curves (LCs). Unfortunately, this is not always the case for astronomical time series, and the analysis has to be carried out in the time domain with the CCF (e.g. Netzer & Peterson, this volume). Two commonly used methods for estimating the CCF are the interpolated CCF (ICCF) (Gaskell & Peterson, 1987) and the discrete correlation function (DCF) (Edelson & Krolik, 1988). The ICCF approximates the CCF integrals by interpolating between the observed points. However, this may be unreliable if the LCs are under-sampled, and the ICCF offers no error estimates on the reconstructed CCF. In the DCF, all pairs of points from the two LCs are ordered according to their time difference, τ_{ij}, and binned according to the user's discretion. The CCF of each bin is then estimated by

$$r_{\mathrm{dcf}}(\tau) = \frac{1}{n} \sum_{\tau_{ij} \in \mathrm{bin}} \frac{(x_i - \bar{x})(y_j - \bar{y})}{s_x s_y} = \frac{1}{n} \sum_{\tau_{ij} \in \mathrm{bin}} u_{ij} \, , \tag{1}$$

and the error on the estimate by the scatter of the u_{ij} terms ($\bar{x}$, $\bar{y}$, s_x^2, s_y^2 are the sample means and variances). The problem is that the statistical properties of r_{dcf} are not well understood. Its similarity to Pearson's estimator of the linear correlation coefficient, which differs only by having a $1/(n-1)$ normalization factor, suggests that its distribution is likewise very skewed. Comparative studies of the ICCF and DCF show that the DCF error estimates are strongly over-estimated, and that it is not better than the ICCF (Rodriguez-Pascual, Santo-Lleó & Clavel, 1989; White & Peterson, 1994). Nevertheless, the idea of avoiding interpolation is attractive, and the question is whether the DCF can be improved, in spite of the lack of a simple, general statistical description of the bin's parent distribution.

2. The z-transformed discrete correlation function

The goals of the ZDCF method are exploratory (i.e. no prior models assumed), conservative analysis of sparse data, and a *qualitative* improvement over the DCF and ICCF. These yield a *quantitative* improvement when the ZDCF is used together with non-parametric estimators (Section 3).

The ZDCF approximates the bin's distribution by the bi-normal distribution, for which Fisher's z-transform (Fisher, 1921)

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

is roughly normally distributed with a known mean, $\bar{z}(\rho)$ and variance, $s_z^2(\rho)$. These are estimated by $\bar{z}(r)$ and $s_z^2(r)$. The bin's ZDCF estimate is then

$$r_{\mathrm{zdcf}}(\tau) = r^{+(\tanh(\bar{z}+s_z)-r)}_{-(r-\tanh(\bar{z}-s_z))}\,. \tag{3}$$

The binning is by equal population, rather than equal $\Delta\tau$. The z-transform's convergence requires a minimum of $n_{\mathrm{min}} = 11$ points per bin. Dependent pairs (e.g. τ_{ij}, τ_{ik}) are discarded to avoid bias. The effect of measurement errors is conservatively over-estimated by the Monte Carlo averaged ZDCF of LCs with simulated random errors. This over-estimate is an unavoidable result of not assuming prior models for the LCs.

This procedure has several consequences: the ZDCF has unequal $\Delta\tau$ bins; only one free parameter, n_{min}, governs the binning, as opposed to the DCF's two ($\Delta\tau$ and τ_{max}); the LCs must have at least 12 points. When the LCs are auto-correlated, the bin's parent distribution reflects that of the observing times, which is usually broader than normal (negative kurtosis). A known bias of the z-transform is that this leads to conservative, over-estimated errors. Finally, as a result of equal population binning, this bias depends on n_{min} only and not of the number of observations. This is an advantage for analyzing heterogeneous samples of unequally sampled LCs.

The z-tranform can be tested separately of the binning by drawing n pairs from distributions of a known correlation coefficient ρ, calculating r_{dcf}, r_{zdcf} and their errors, and noting the fraction of times where ρ lies outside the error interval. Figure 1 shows the results for the bi-normal and

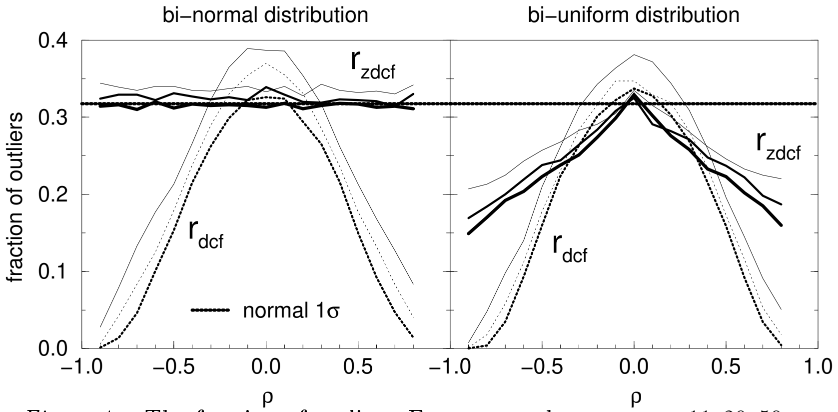

Figure 1. The fraction of outliers. From top to bottom: $n = 11, 20, 50$.

$n_{\rm obs}$	$\alpha = 1$			$\alpha = 2$		
	$\frac{\Delta(ZDCF)}{\Delta(DCF)}$	$\frac{\Delta(ZDCF)}{\Delta(ICCF)}$	$\frac{\Delta(DCF)}{\Delta(ICCF)}$	$\frac{\Delta(ZDCF)}{\Delta(DCF)}$	$\frac{\Delta(ZDCF)}{\Delta(ICCF)}$	$\frac{\Delta(DCF)}{\Delta(ICCF)}$
15	0.75	0.74	0.99	0.97	2.48	2.96
20	0.74	0.94	1.27	0.54	2.50	2.52
25	0.85	1.00	1.18	0.46	2.02	1.91

TABLE 1. The ratios of the mean $\Delta = |\tau_{\max r} - \tau_{\rm peak}|$ over 2500 simulations.

the bi-uniform distributions. In both cases $r_{\rm zdcf}$ performs much better than $r_{\rm dcf}$, whose errors are highly exaggerated for $\rho \neq 0$ and strongly *increase* with n.

The binning algorithm is tested separately of the z-transform by comparing the success of the different methods in estimating the time lag of the CCF peak. This is done by generating a random LC with a power law power spectrum $P(\nu) \propto \nu^{-\alpha}$, shifting a copy of it by a random timelag, sampling both at $n_{\rm obs}$ random times, calculating the DCF, ZDCF and ICCF and noting Δ, the absolute difference between the true time lag of the peak and the estimate's maximum. Table 1 shows that the ZDCF out-performs the DCF for both hard ($\alpha = 1$) and soft ($\alpha = 2$) LCs, and unlike the DCF, is in fact better than the ICCF for under-sampled LCs ($\alpha = 1$ and $n_{\rm obs} < 25$).

3. Uncovering correlations involving the variability timescale

I demonstrate how the ZDCF can be used to uncover correlations between the variability timescale and other AGN properties by using a toy model for the luminosity–variability relation. I assume that all AGN have power-law power spectra, where the index α is drawn from the uniform distribution $U(1, 2)$. The variability properties are linked to the absolute magnitude through the relation $M(\alpha) = 26 + \alpha + \mathrm{d}M$, where $\mathrm{d}M \sim N(0, \Delta M)$ is a normal random scatter, which simulates measurement errors in M and

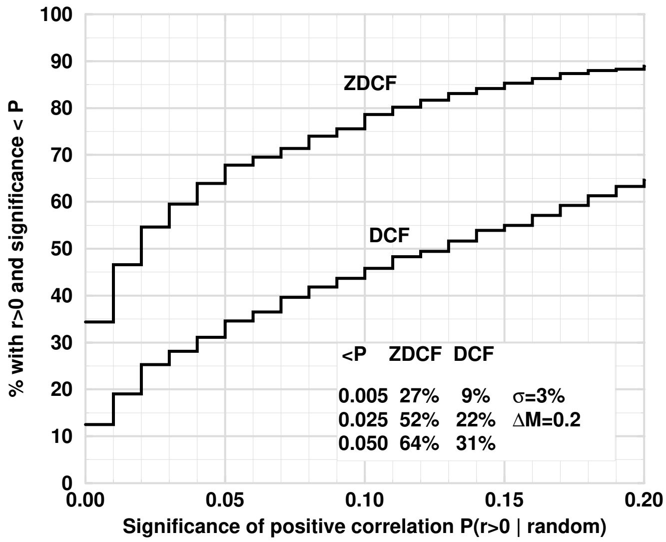

Figure 2. The fraction of simulated samples where a positive correlation between τ_0 and M was detected at a significance below a given threshold.

/ or the effects of additional physical parameters, other than α, on the luminosity.

A simulated LC sample was generated by drawing 30 α values, and using each to generate a random LC. Each LC was then associated with an absolute magnitude M with $\Delta M = 0.2$. The 30 LCs were then sampled with 3% measurement errors at the same 15 times, where the observation pattern, which displays the typical clustered pattern of astronomical time series, was taken from one of the LCs observed at the Wise Observatory. For each LC, the ACF was estimated by both the DCF and ZDCF, and the typical variability timescale was quantified by the zero-crossing time, τ_0, which is the shortest time it takes the ACF to fall to zero. τ_0 was estimated by an error-weighted least squares fit of the straight line $\mathrm{ACF}(\tau) = 1 - \tau/\tau_0$ to the ACF (Netzer et al., 1996). Finally, the correlation between M and τ_0 was estimated by Spearman's rank order correlation coefficient. This procedure was repeated 10^3 times. Figure 2 shows the fraction of results where a positive correlation was uncovered at a significance less than a given threshold. The ZDCF is 2–3 times more efficient in detecting the correlation than the DCF.

To conclude, the ZDCF is much more efficient than the DCF in uncovering correlations involving the variability timescale, and deals with undersampled LCs better than both the DCF and ICCF.

References

Edelson R. A., Krolik J. H., 1988, ApJ, 333, 646
Fisher R. A., 1921 Metron, 1, No. 4, 1

Gaskell C. M., Peterson B. M., 1987, ApJ Suppl., 65, 1
Gayen A. K., 1951, Biometrika, 38, 219
Netzer H. et al., 1996, MNRAS, 279, 429
Rodriguez-Pascual P. M., Santo-Lleó M., Clavel J., 1989, AA, 219, 101
White R. J., Peterson B. M., 1994, PASP, 106, 879