Time Series Analysis in Studies of AGN Variability

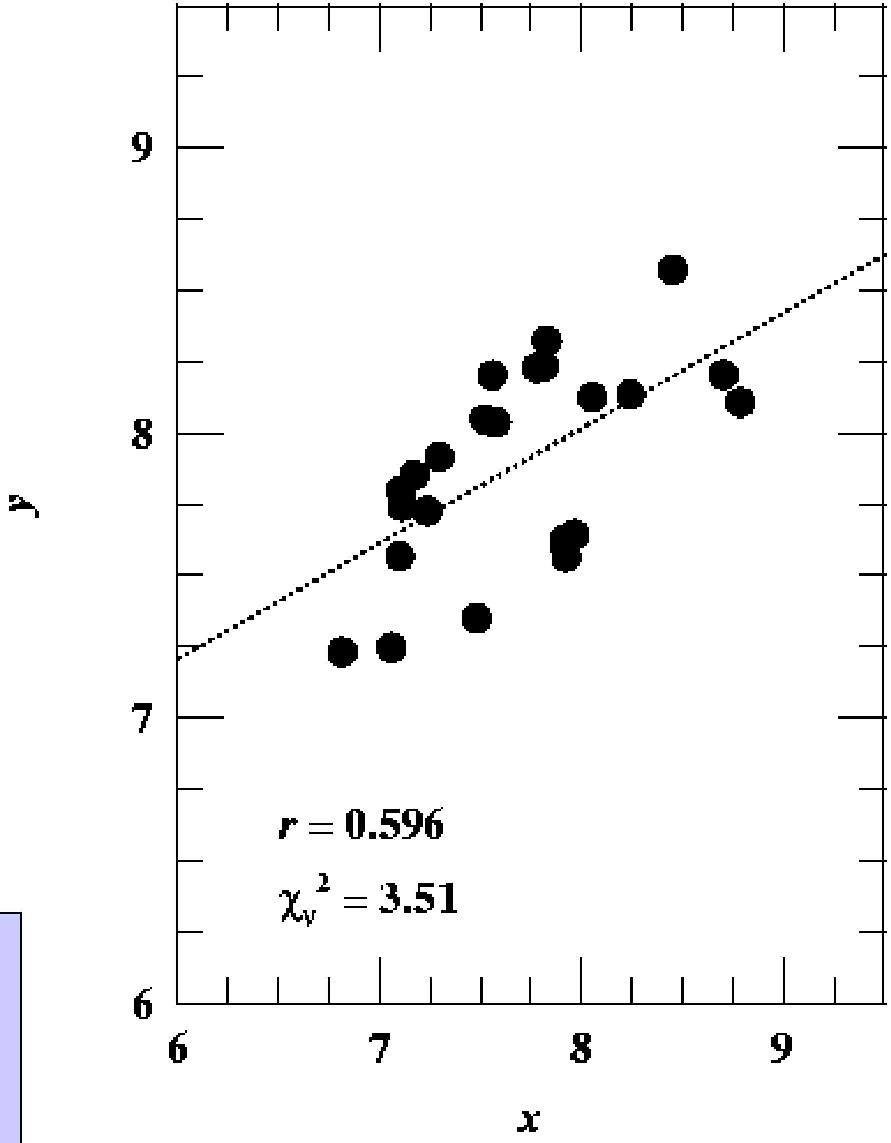
Bradley M. Peterson

The Ohio State University

Linear Correlation

- Degree to which two parameters are ***linearly*** correlated can be expressed in terms of the linear correlation coefficient:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\left(\sqrt{\sum_i (x_i - \bar{x})^2}\right)\left(\sqrt{\sum_i (y_i - \bar{y})^2}\right)}$$



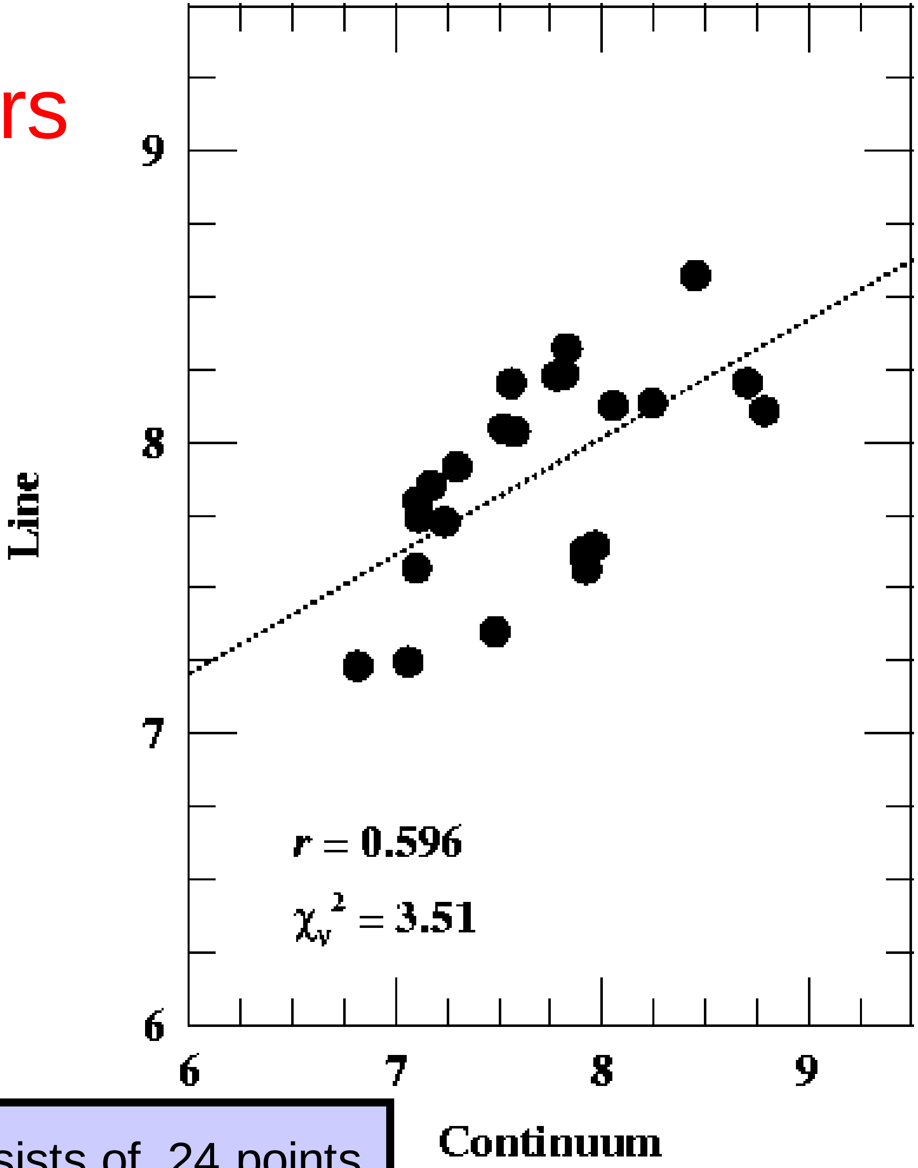
r = 1: perfect correlation

r = 0: no correlation

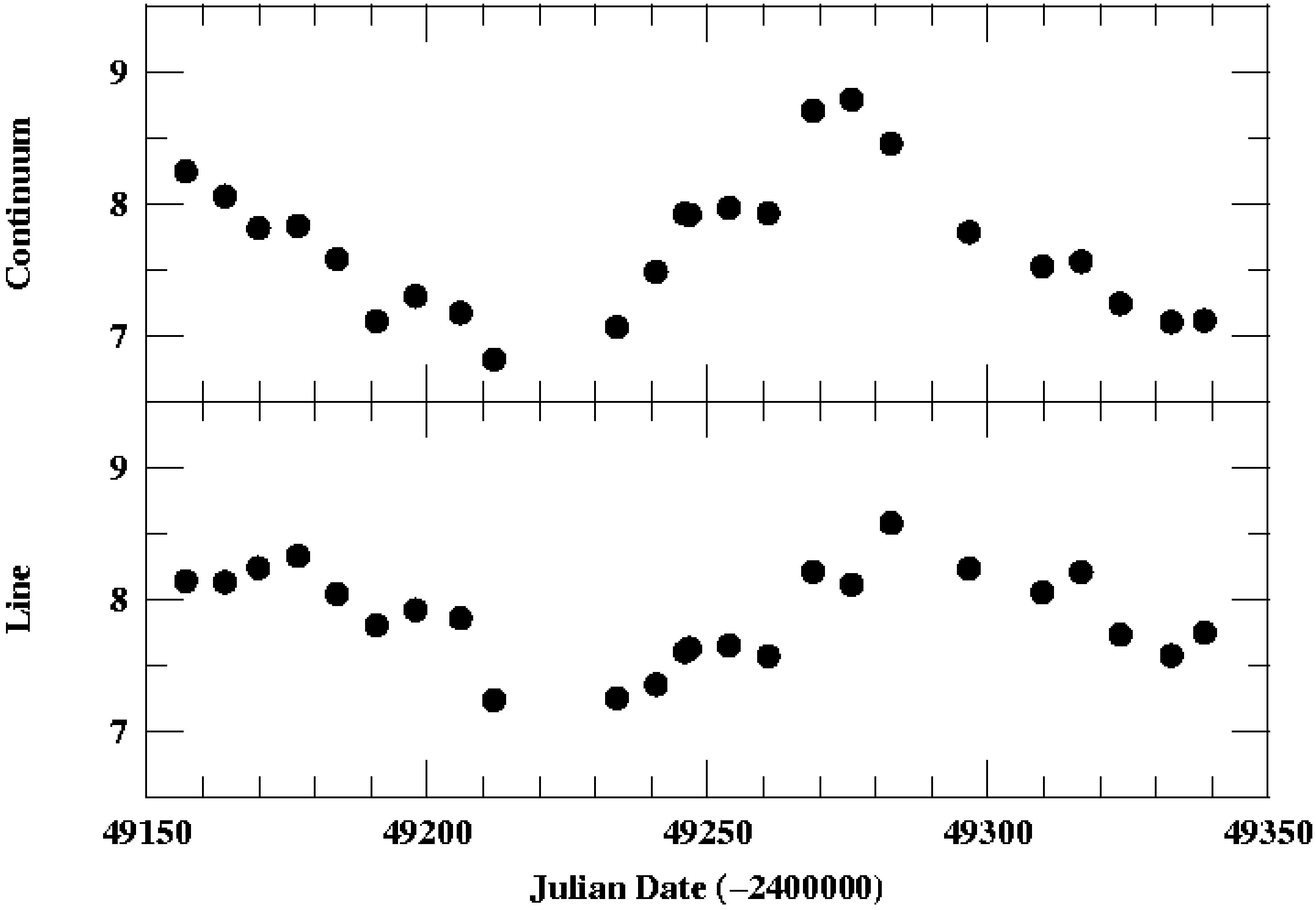
r = –1: perfect anticorrelation

Correlation Between Time-Varying Parameters

- In fact, the data shown in the example are continuum and Hβ fluxes in a variable Seyfert 1 galaxy, Mrk 335.
  - $x = C(t)$
  - $y = L(t)$
- The continuum and emission-line fluxes are highly correlated.

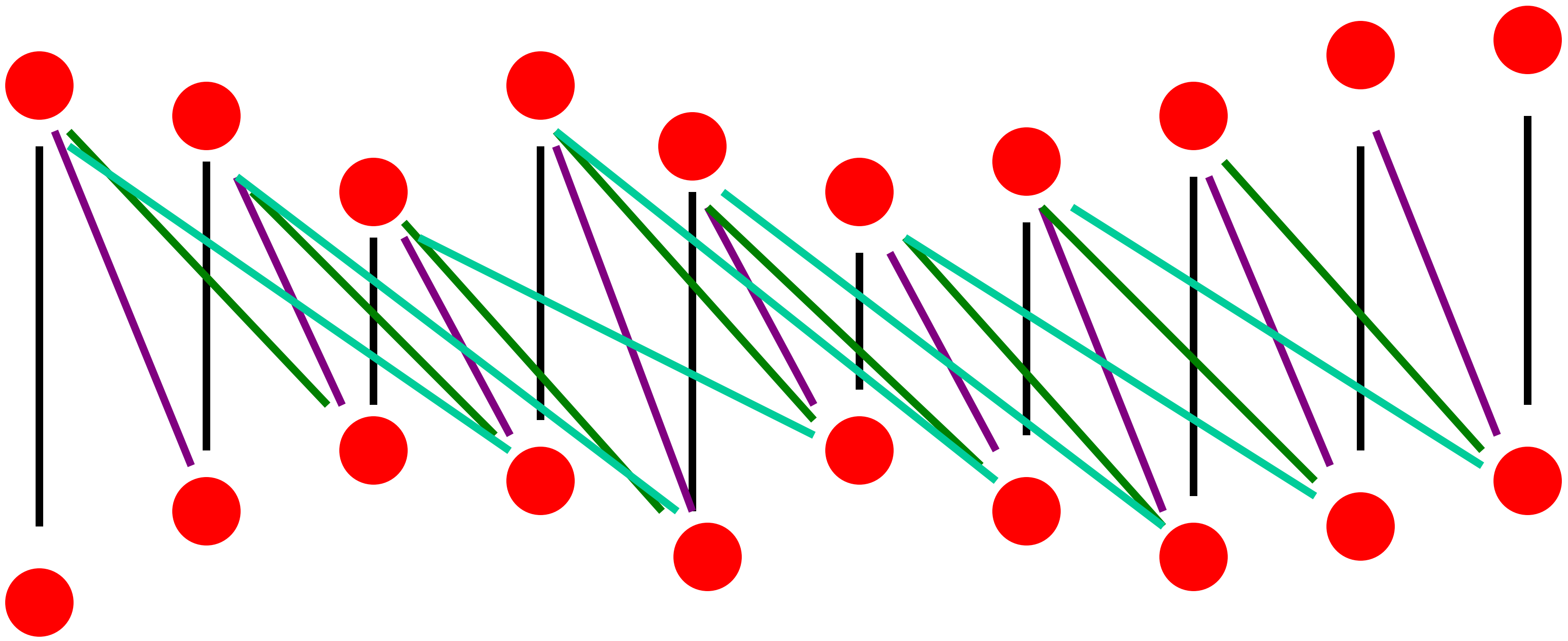


Mrk 335 data consists of 24 points average spacing of 7.9 days.



Instead of letting x = $C(t)$ and y = $L(t)$, improve the correlation by letting x = $C(t)$ and y = $L(t + \tau)$, where τ is the time-shift or “lag”

Cross-correlating evenly spaced data is trivial



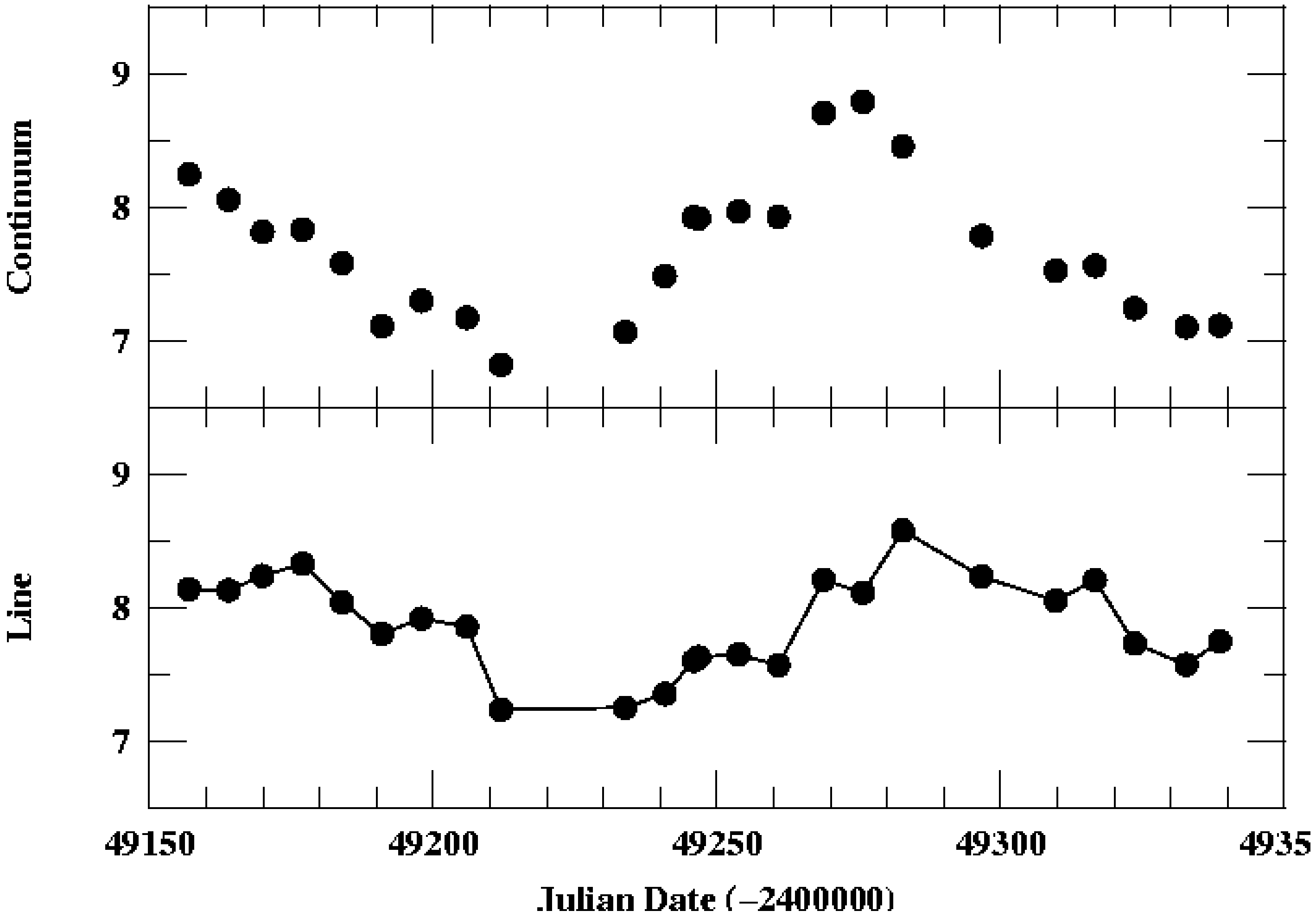
Shift = 3 units

Shift = 2 units

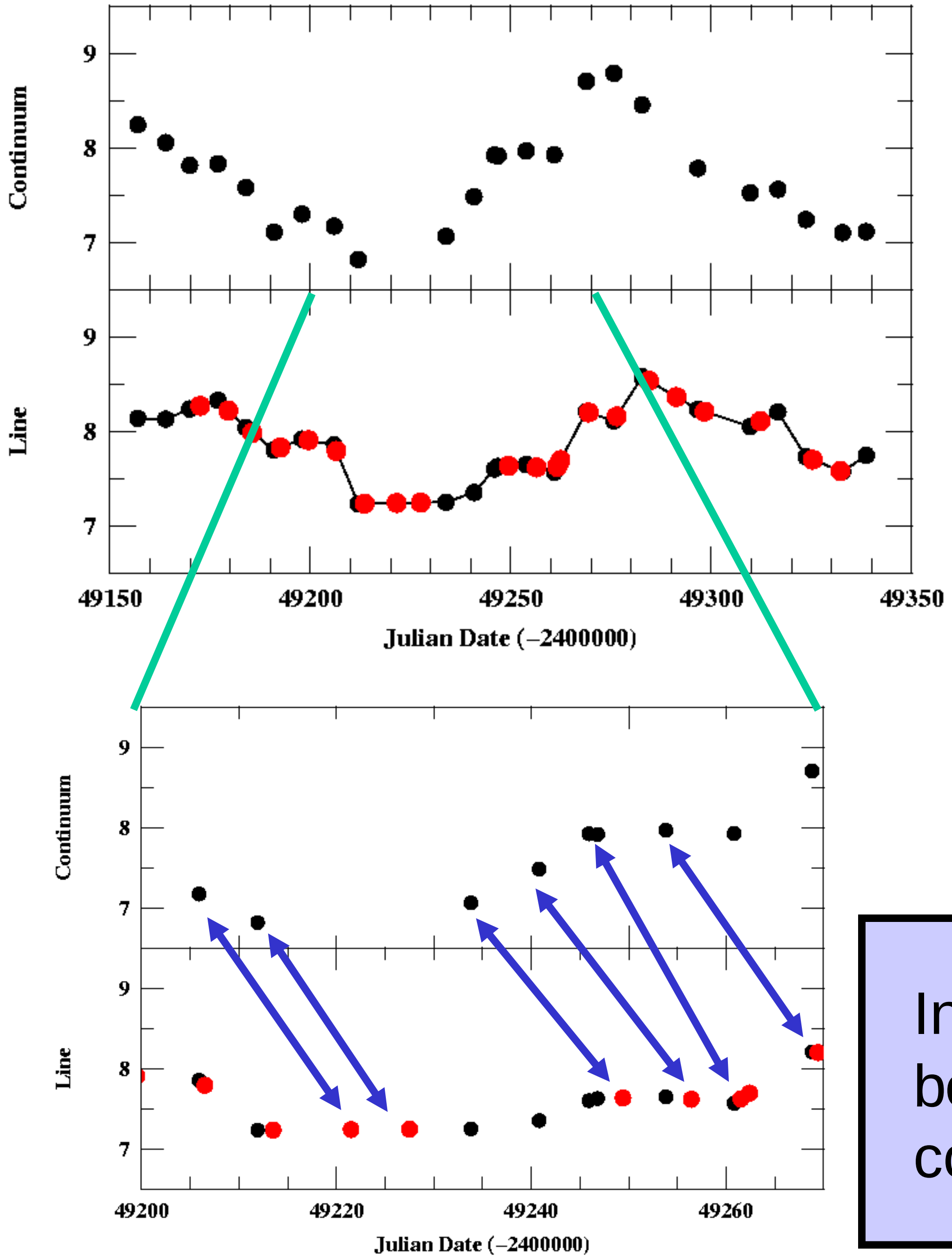
Shift = 0 units

Shift = 1 unit

Goal: find the value of the shift that maximizes the correlation coefficient.



First practical problem: in general, data are not evenly spaced. One solution is to interpolate between real data points.

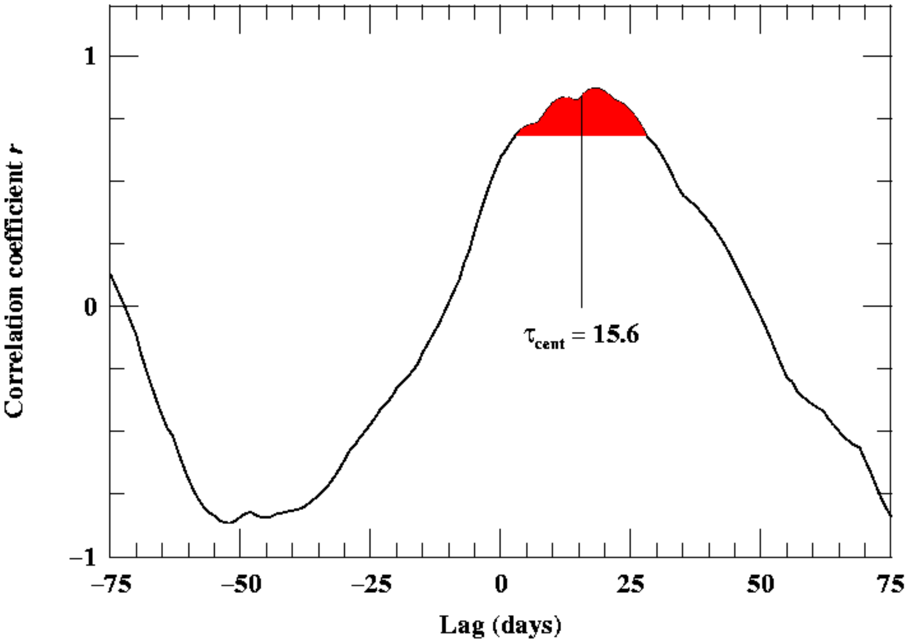


Each real datum $C(t)$ in one time series is matched with an interpolated value $L(t + \tau)$ in the other time series and the linear correlation coefficient is computed for all possible values of the lag τ.

Interpolated line points lag behind corresponding continuum points by 16 days.

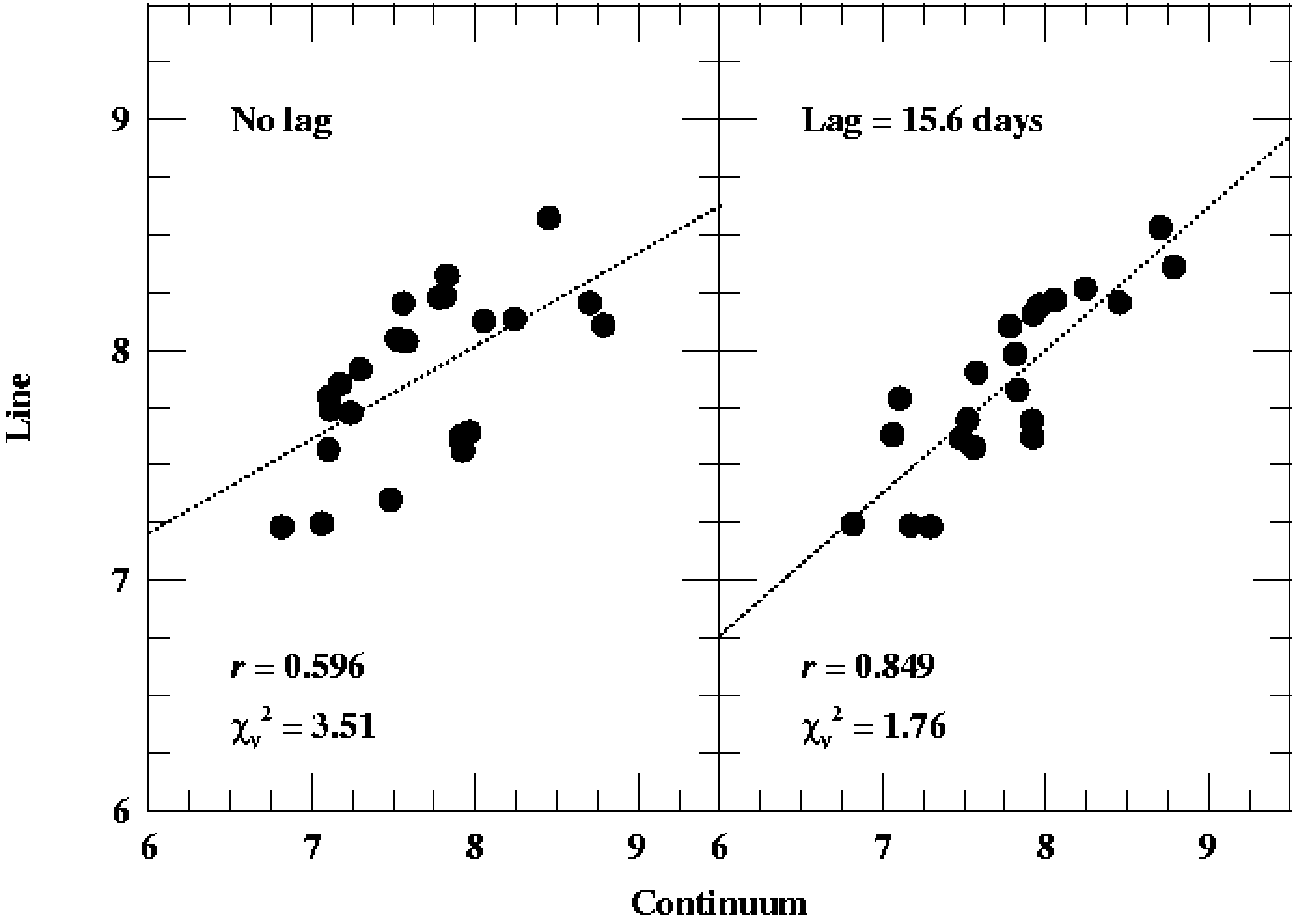
Cross-Correlation Function

- Linear correlation coefficient as a function of time lag is the “cross-correlation function” (CCF).
- The formal definition of the CCF as a continuous function is the convolution integral:



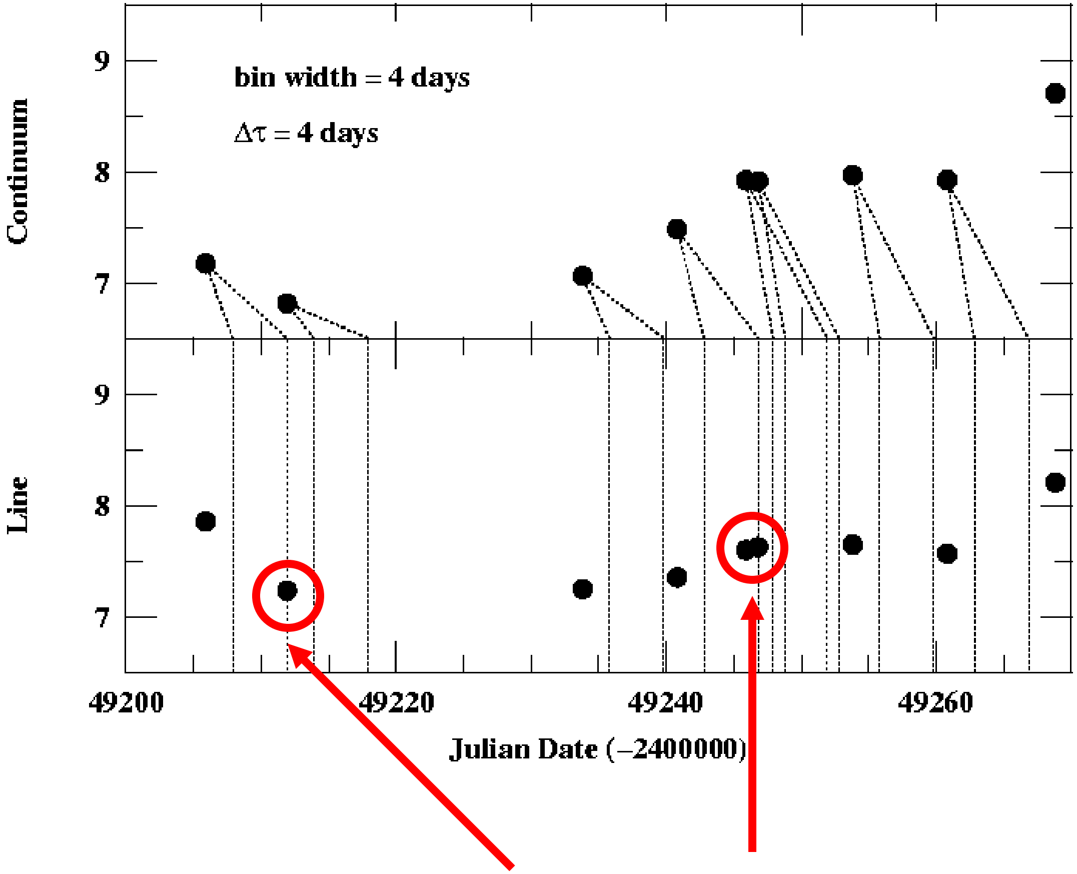
$$\mathrm{CCF}(\tau) = \int_{-\infty}^{+\infty} L(t)\, C(t-\tau)\, dt$$

The Time-Shift Improves the Linear Correlation



Discrete Correlation Function (DCF)

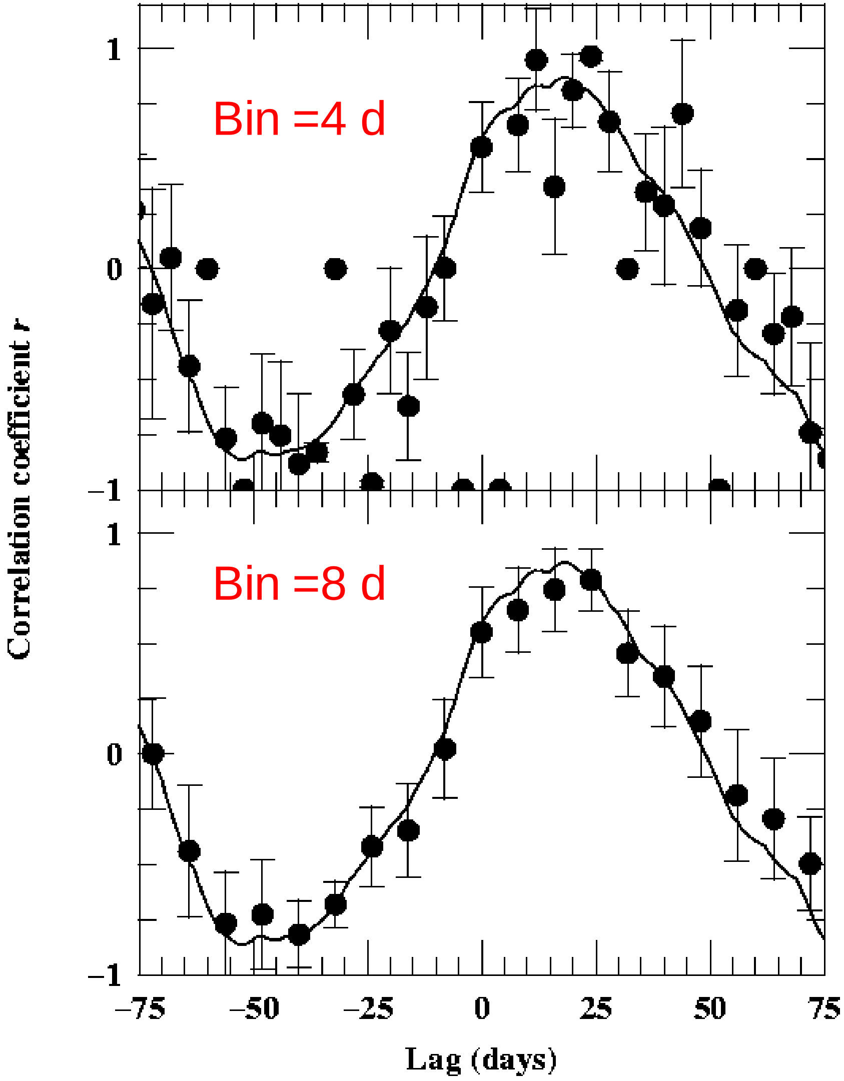
- Potential problem: what if you cannot reasonably interpolate between actual observations?
- Alternative method is the DCF. For each continuum point, search for line points within a certain bin width of the value of τ being computed.



Only contributing points in the current computation

Discrete Correlation Function (DCF)

- Same example shown.
- Bin width is somewhat arbitrary, but median sampling (lower panel) seems to give good results.
- DCF is a very conservative approach.
  - Valuable check when gaps in data.
  - In general performs much worse than interpolated CCF.

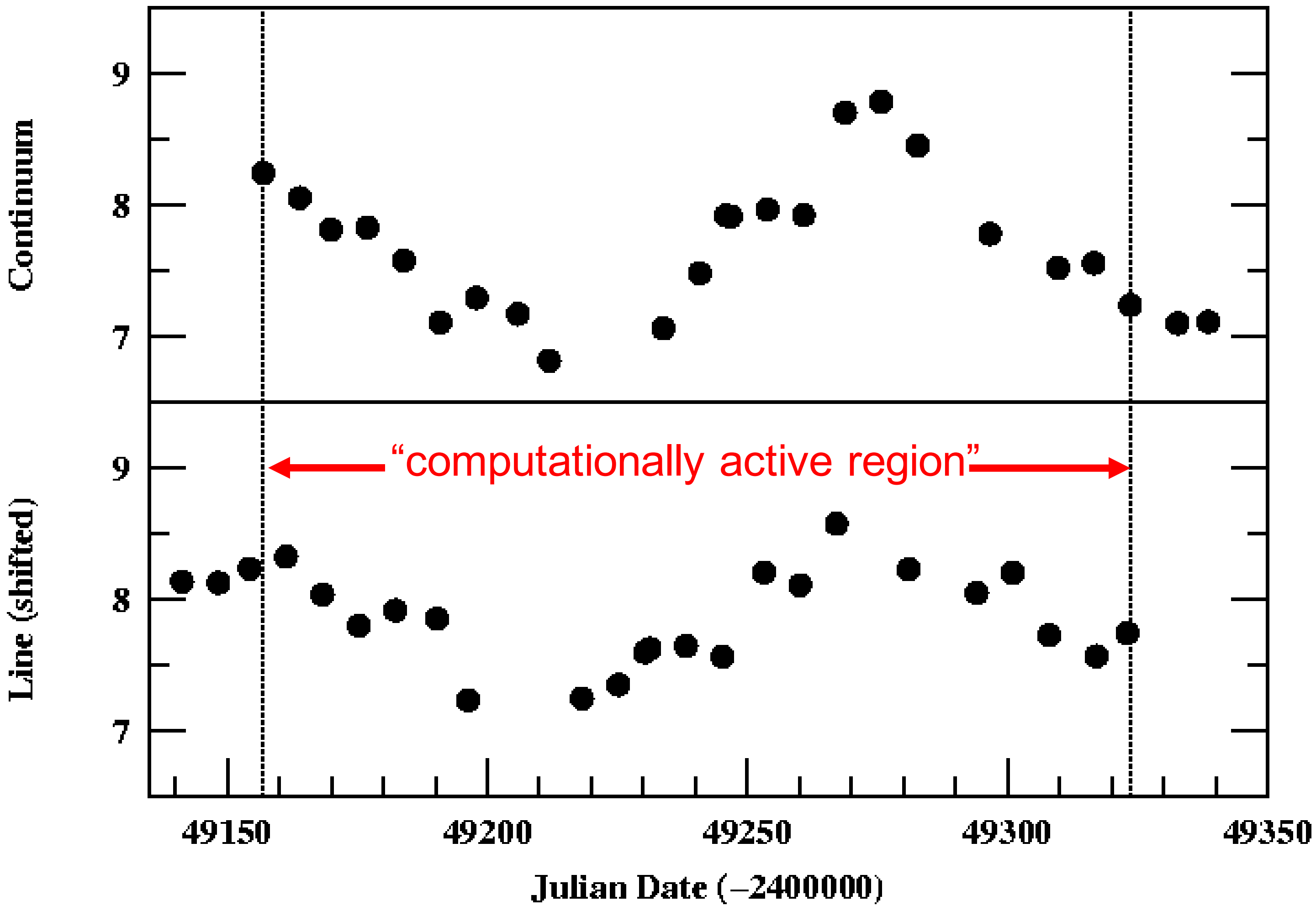


Computational Subtleties

- Can the results be improved with higher-order (non-linear) interpolation?
  - Doesn’t seem to be any advantage, and non-linear functions are hard to control.
- Can accuracy greater than average time interval be obtained?
  - Yes, provided that that light curves are reasonably well sampled.

Computational Subtleties

- If you have time series with N points, you are using all N points in the calculation only at $\tau = 0$. Otherwise, points at the end of the series drop out of the calculation.
- This has two consequences:
  - – Time series are almost always “non-stationary”.
  - – The significance of any lag must be assessed in terms of the actual number of point contributing to the correlation coefficient at the measured lag.



While the original time series has N = 24 points, at the shift τ = 15.6 days, the number of actual point contributing are N = 22 in the continuum and N = 21 in the line. This effect can be important for larger values of the lag.

Non-Stationary Series

- Statistically speaking, a series is “stationary” if the mean and standard deviation do not change when individual points at the ends of the series drop out.
- Since AGN time series are short, this is *never* true.
- For each lag, means and standard deviations must be recomputed to correctly normalize the CCF.

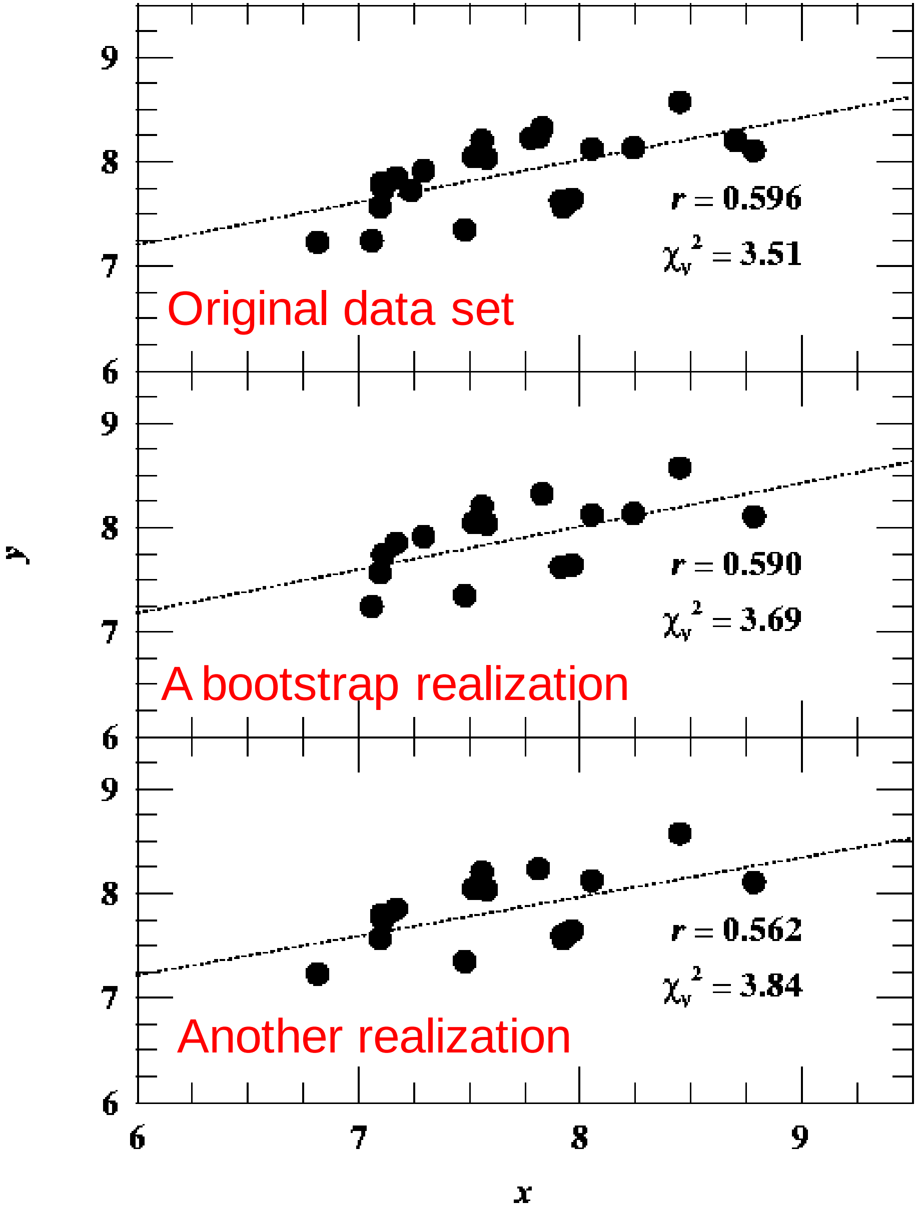
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\left(\sqrt{\sum_i (x_i - \bar{x})^2}\right)\left(\sqrt{\sum_i (y_i - \bar{y})^2}\right)}$$

Uncertainties in Cross-Correlation Lags

- Determining errors in cross-correlation lags has been a vexing problem for over a decade.
- At present, the best method is a model-independent Monte-Carlo method called "FR/RSS":
  - FR: Flux redistribution
    - accounts for the effects of uncertainties in flux measurement
  - RSS: Random subset selection
    - accounts for effects of sampling in time

Bootstrap Method

- RSS is based on a computationally intensive method for evaluating significance of linear correlation known as the “bootstrap method”.
- Bootstrap method:
  - for *N* real data points, select at random *N* points without regard to whether or not they have been previously selected.
  - Determine *r* for this subset
  - Repeat many times to obtain a distribution in the value of *r*. From this distribution, compute the mean and standard deviation for *r*.

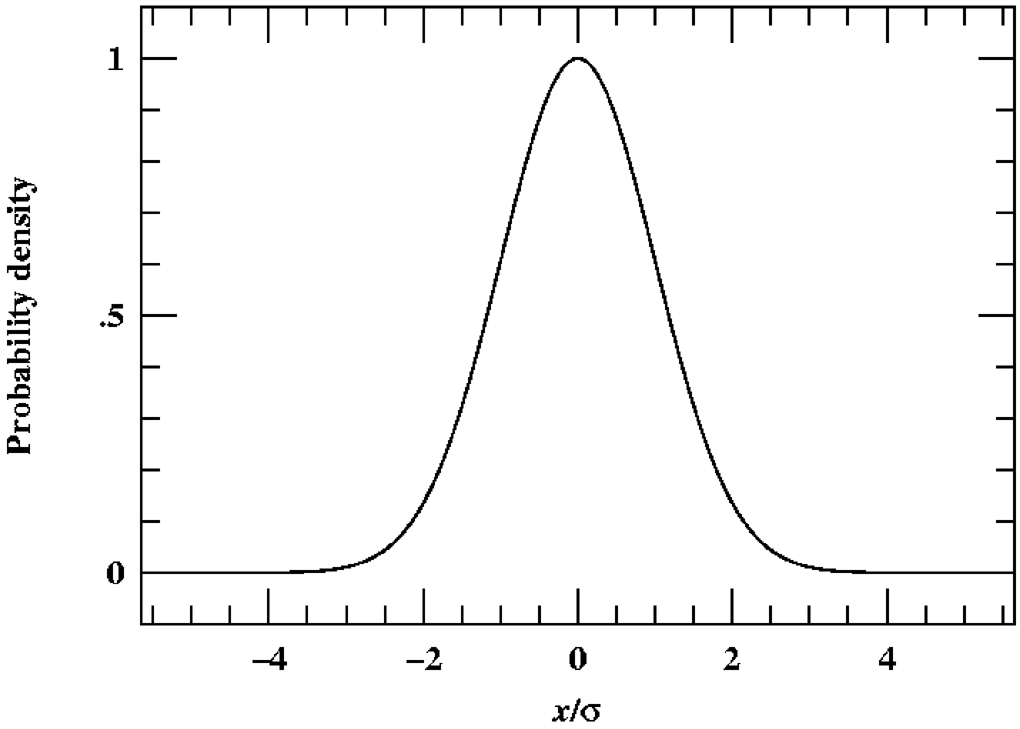


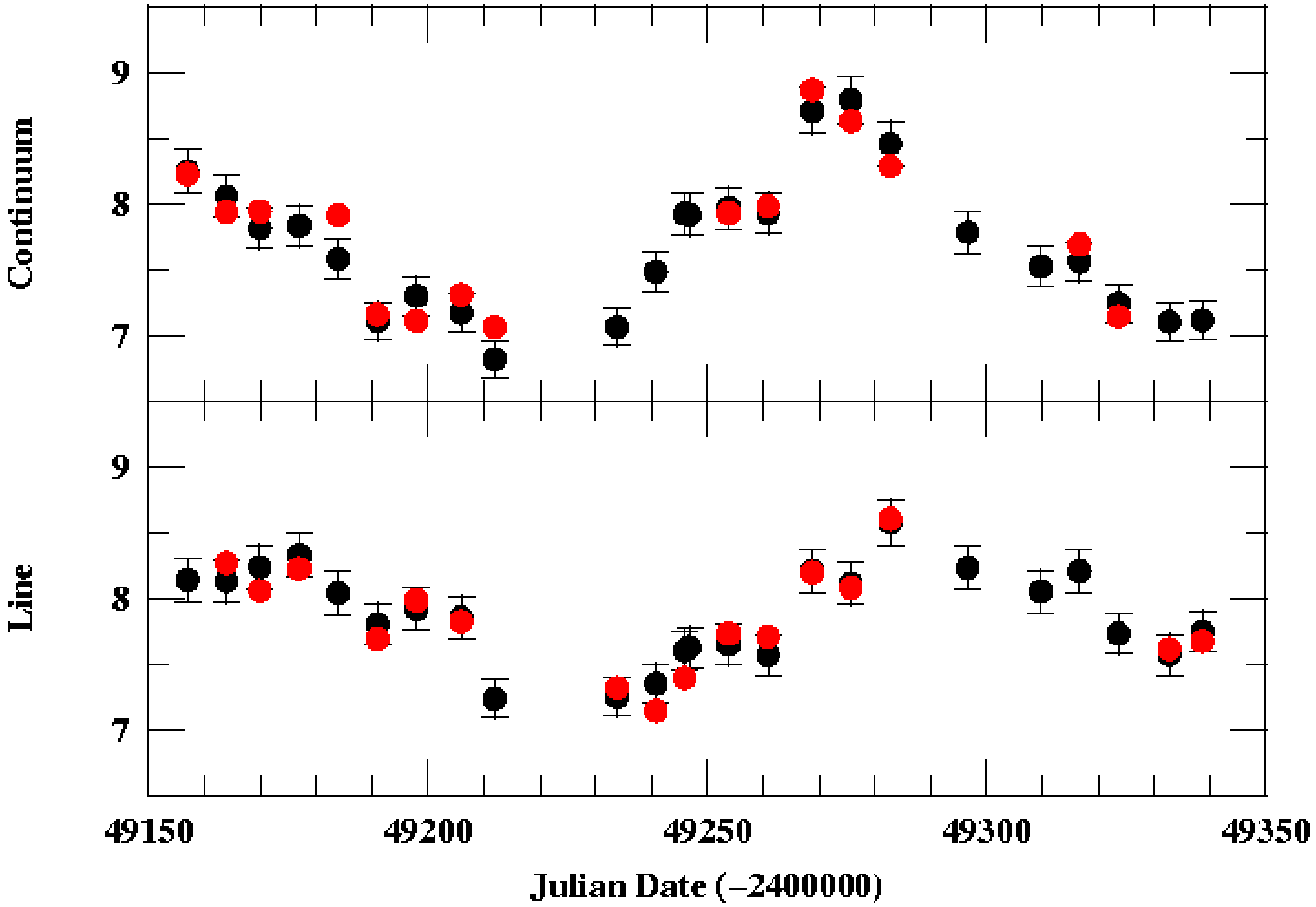
Random Subset Selection

- How do you deal with redundant selections in a time series, where order matters? Either:
  - Ignore redundant selections
    - Each realization has typically 1/*e* fewer points than the original (origin of the name RSS).
    - Numerical experiments show that this then gives a conservative error on the lag (the real uncertainty may be somewhat smaller).
  - Weight each datum according to number of times selected
    - Philosophically closer to original bootstrap.

Flux Redistribution

- Assume that flux uncertainties are Gaussian distributed about measured value, with uncertainty σ.
- Take each measured flux value and alter it by a random Gaussian deviate.
- Decrease σ by factor $n^{1/2}$, where n is the number of times the point is selected.

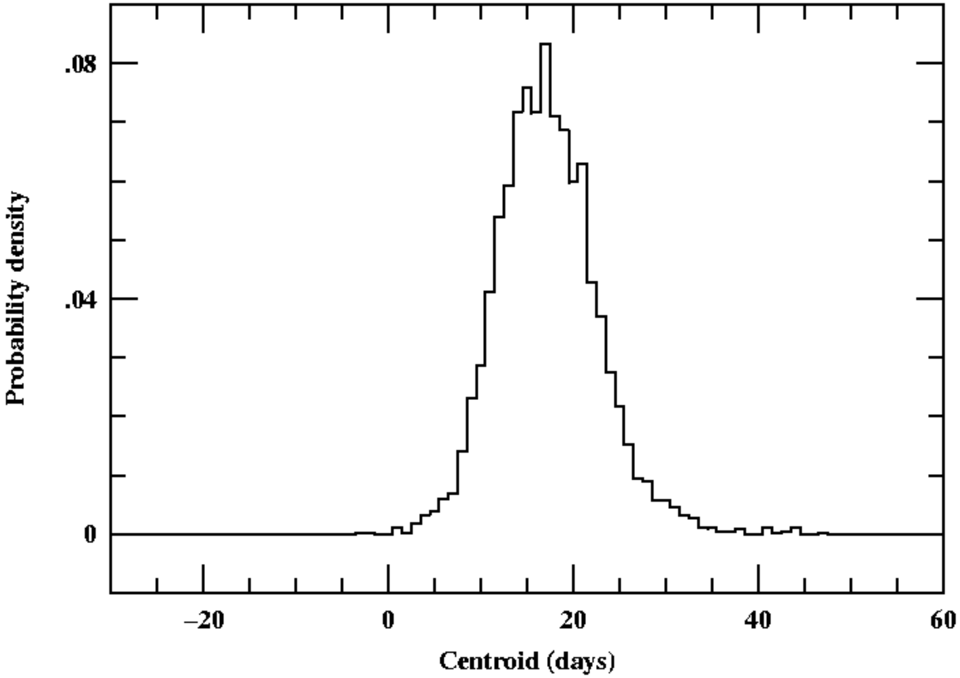




A single FR/RSS realization. Red points are selected at random from among the real (black) points, redundant points are discarded, and surviving points redistributed in flux using random Gaussian deviates scaled by the quoted uncertainty for each point. The realization shown here gives τ_{cent} = 17.9 days (value for original data is 15.6 days)

Cross-Correlation Centroid Distribution

- Many FR/RSS realizations are used to build up the "cross-correlation centroid distribution" (CCCD).
- The rms width of this distribution (which can be non-Gaussian) can be used as an estimate of the lag.



Mrk 335 FR/RSS result:

$$\tau_{\rm cent} = 15.6^{+7.2}_{-3.1} \text{ days}$$