

Environmental Data Analysis with *MatLab*

Lecture 17:

Covariance and Autocorrelation

SYLLABUS

Lecture 01	Using MatLab
Lecture 02	Looking At Data
Lecture 03	Probability and Measurement Error
Lecture 04	Multivariate Distributions
Lecture 05	Linear Models
Lecture 06	The Principle of Least Squares
Lecture 07	Prior Information
Lecture 08	Solving Generalized Least Squares Problems
Lecture 09	Fourier Series
Lecture 10	Complex Fourier Series
Lecture 11	Lessons Learned from the Fourier Transform
Lecture 12	Power Spectral Density
Lecture 13	Filter Theory
Lecture 14	Applications of Filters
Lecture 15	Factor Analysis
Lecture 16	Orthogonal functions
Lecture 17	Covariance and Autocorrelation
Lecture 18	Cross-correlation
Lecture 19	Smoothing, Correlation and Spectra
Lecture 20	Coherence; Tapering and Spectral Analysis
Lecture 21	Interpolation
Lecture 22	Hypothesis testing
Lecture 23	Hypothesis Testing continued; F-Tests
Lecture 24	Confidence Limits of Spectra, Bootstraps

purpose of the lecture

apply the idea of covariance

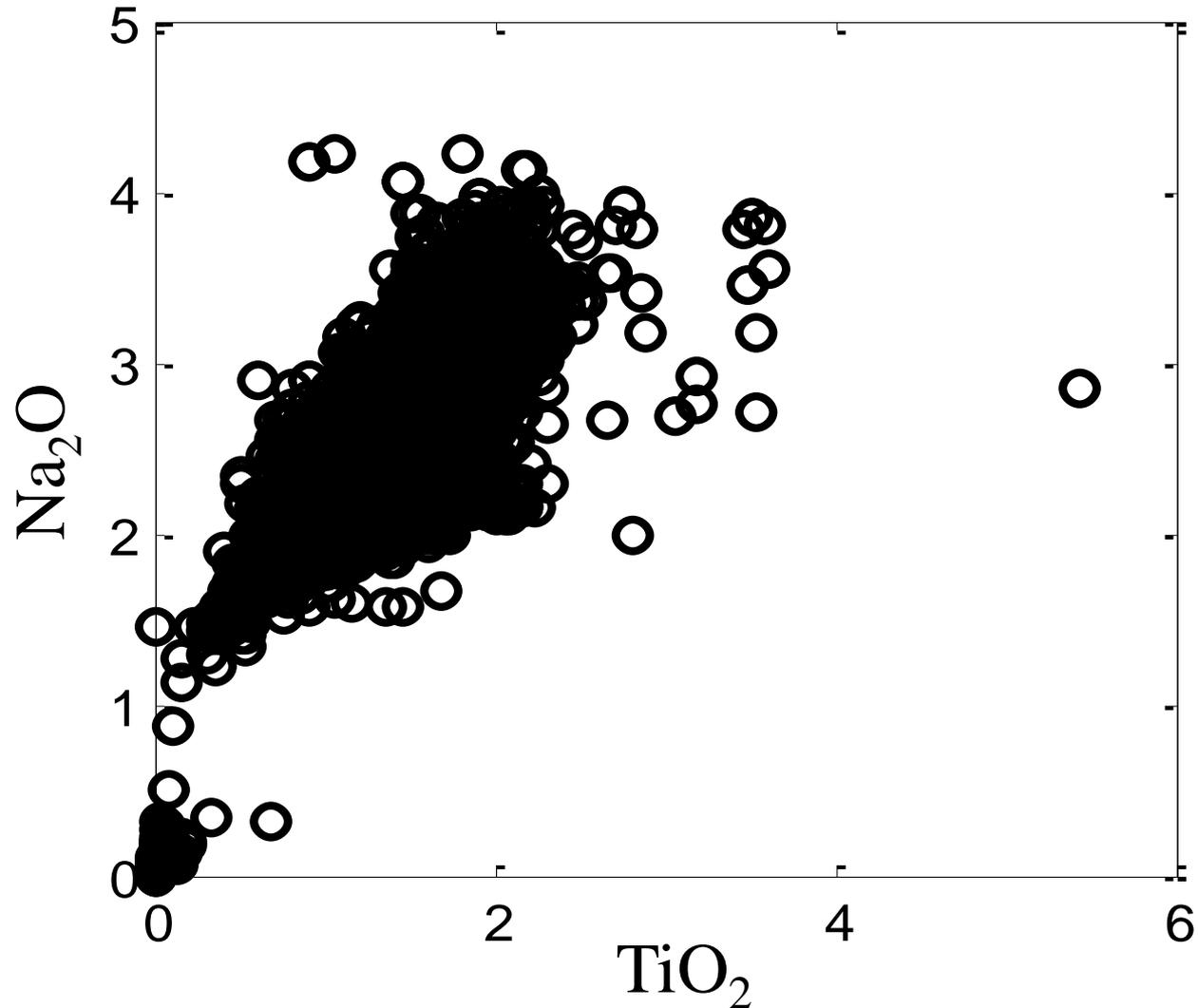
to time series

Part 1

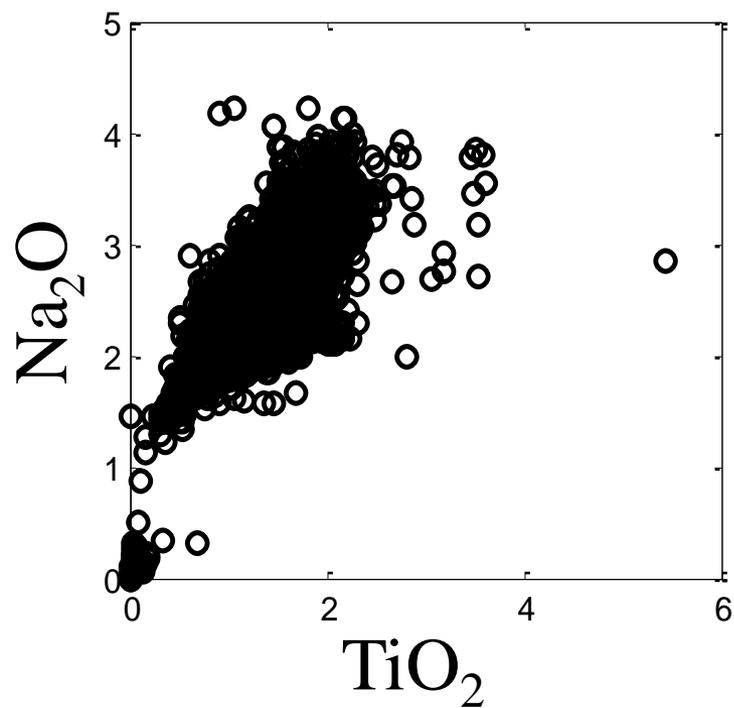
correlations between random variables

Atlantic Rock Dataset

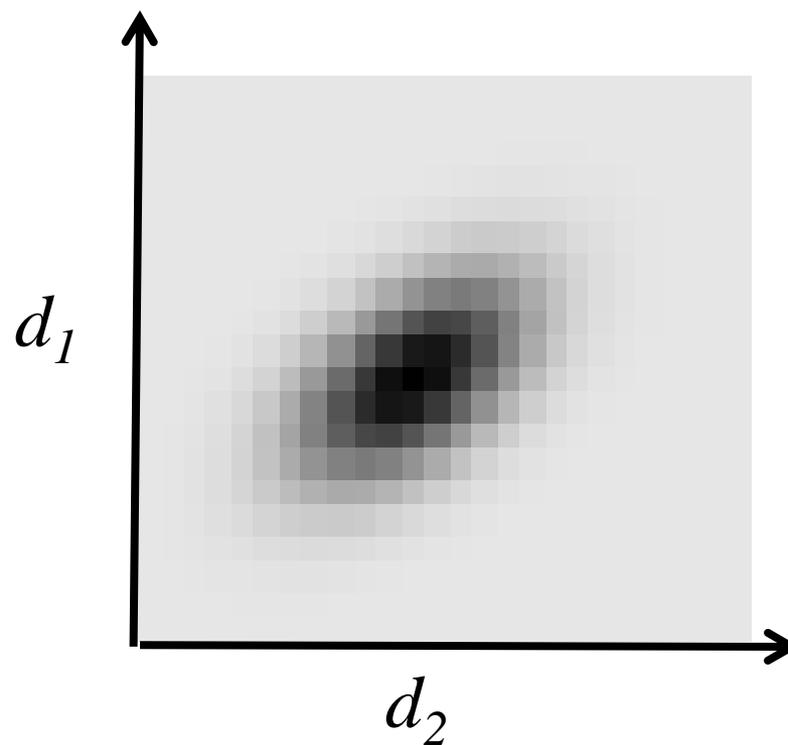
Scatter plot of TiO_2 and Na_2O



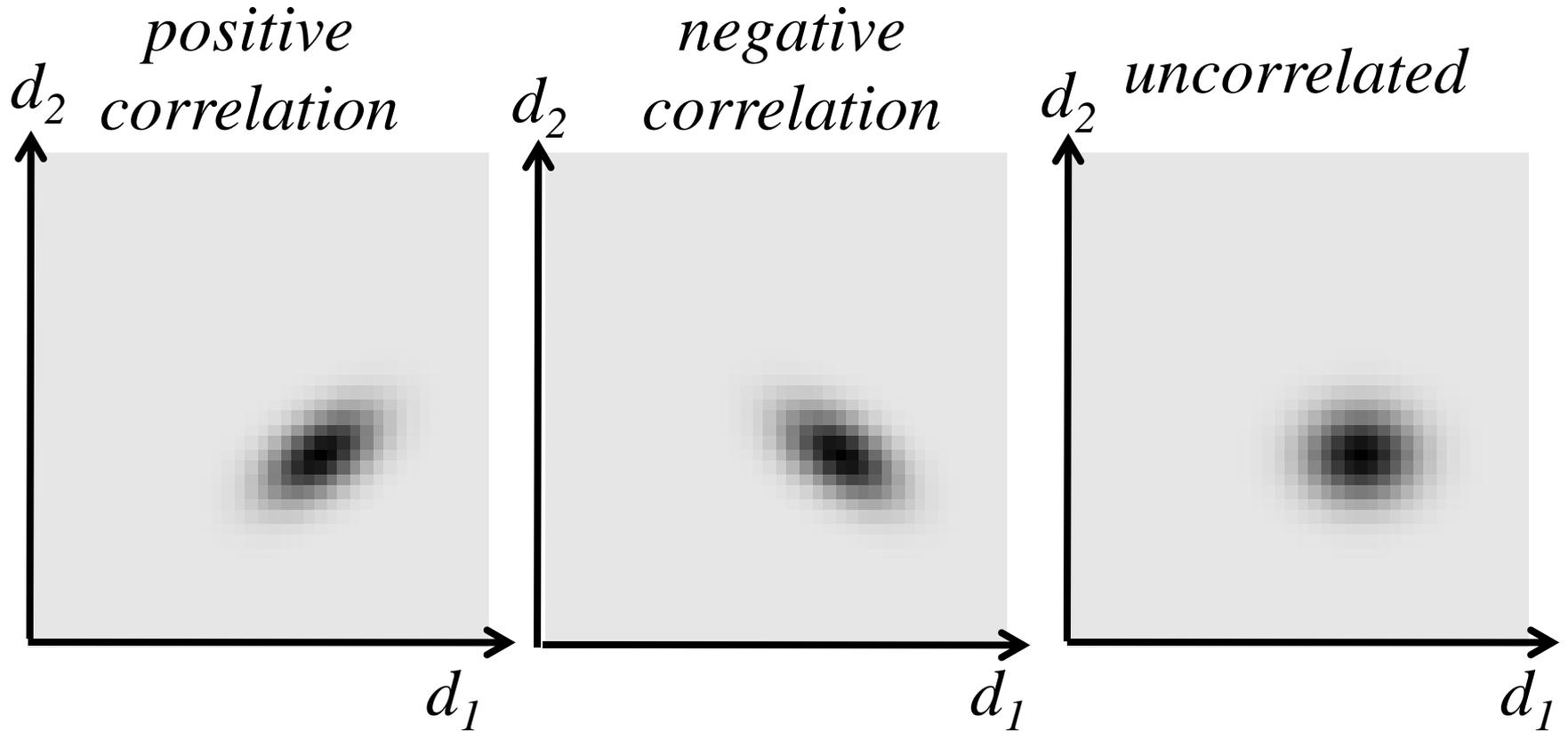
scatter plot



idealization as
a p.d.f.



types of correlations



the covariance matrix

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \dots \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \dots \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

$$C_{ij} = \iint (d_i - \bar{d}_i) (d_j - \bar{d}_j) p(d_1, d_2) dd_1 dd_2$$

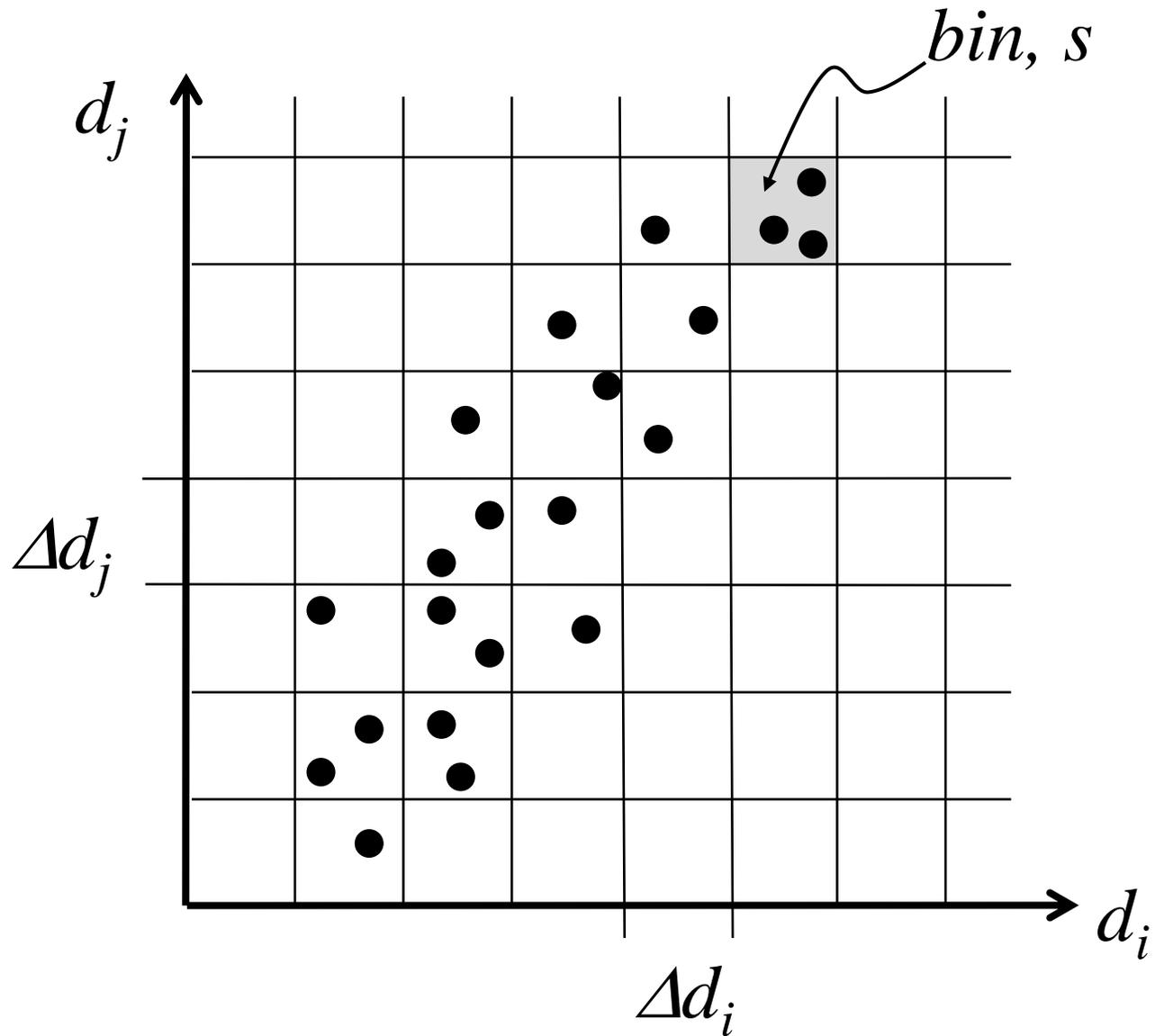
recall the covariance matrix

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \dots \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \dots \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

covariance of d_1 and d_2

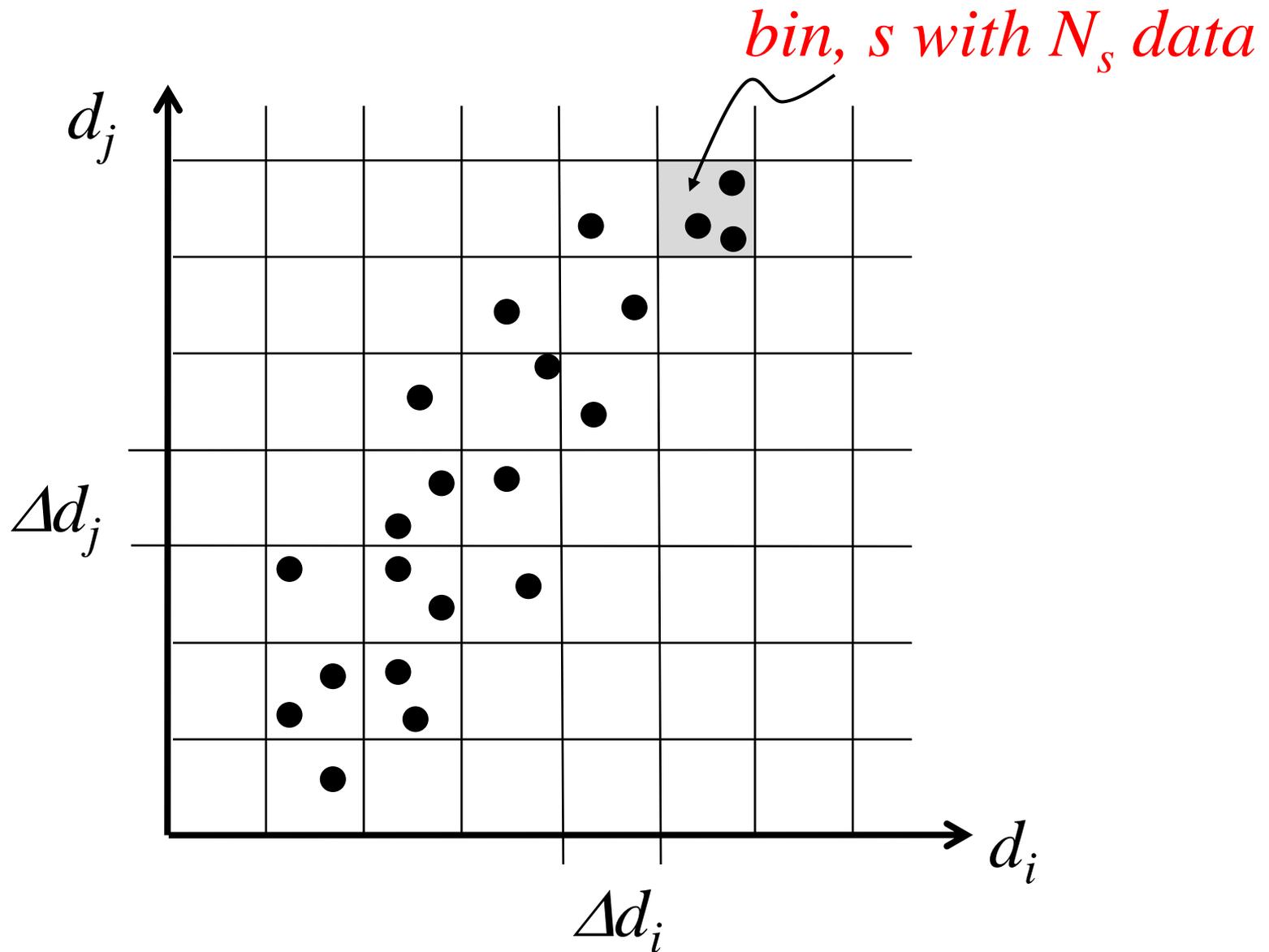
estimating covariance from data

divide (d_i, d_j) plane into bins



estimate total probability in bin from the data:

$$P_s = p(d_i, d_j) \Delta d_i \Delta d_j \approx N_s/N$$



$$C_{ij} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (d_i - \bar{d}_i) (d_j - \bar{d}_j) p(d_i, d_j) dd_i dd_j$$

$$\approx \frac{1}{N} \sum_s \left[d_i^{(s)} - \bar{d}_i \right] \left[d_j^{(s)} - \bar{d}_j \right] N_s$$

$$\approx \frac{1}{N} \sum_{k=1}^N \left[d_i^{(k)} - \bar{d}_i^{(k)} \right] \left[d_j^{(k)} - \bar{d}_j^{(k)} \right]$$

$$C_{ij} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (d_i - \bar{d}_i) (d_j - \bar{d}_j) p(d_i, d_j) dd_i dd_j$$

approximate integral with sum
and use $p(d_i, d_j) \Delta d_i \Delta d_j \approx N_s/N$

$$\approx \frac{1}{N} \sum_s \left[d_i^{(s)} - \bar{d}_i \right] \left[d_j^{(s)} - \bar{d}_j \right] N_s$$

$$\approx \frac{1}{N} \sum_{k=1}^N \left[d_i^{(k)} - \bar{d}_i^{(k)} \right] \left[d_j^{(k)} - \bar{d}_j^{(k)} \right]$$

$$C_{ij} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (d_i - \bar{d}_i) (d_j - \bar{d}_j) p(d_i, d_j) dd_i dd_j$$

approximate integral with sum
and use $p(d_i, d_j) \Delta d_i \Delta d_j \approx N_s/N$

$$\approx \frac{1}{N} \sum_s \left[d_i^{(s)} - \bar{d}_i \right] \left[d_j^{(s)} - \bar{d}_j \right] N_s$$

shrink bins so no more than one
data point in each bin

$$\approx \frac{1}{N} \sum_{k=1}^N \left[d_i^{(k)} - \bar{d}_i^{(k)} \right] \left[d_j^{(k)} - \bar{d}_j^{(k)} \right]$$

“sample” covariance

$$C_{ij} \approx \frac{1}{N} \sum_{k=1}^N \left[d_i^{(k)} - \bar{d}_i^{(k)} \right] \left[d_j^{(k)} - \bar{d}_j^{(k)} \right]$$

normalize to range ± 1

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}$$

“sample” correlation coefficient

R_{ij}

-1

perfect negative correlation

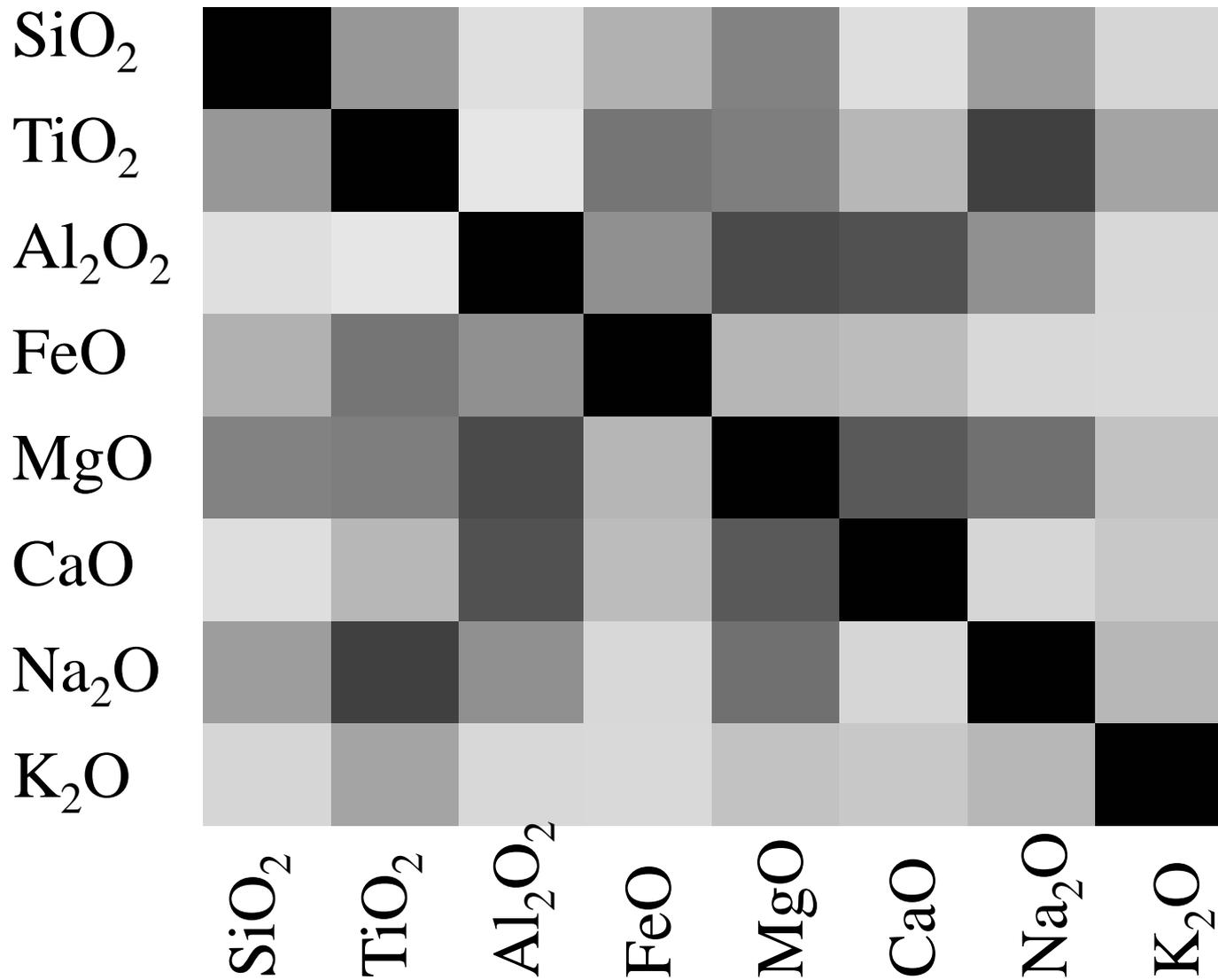
0

no correlation

1

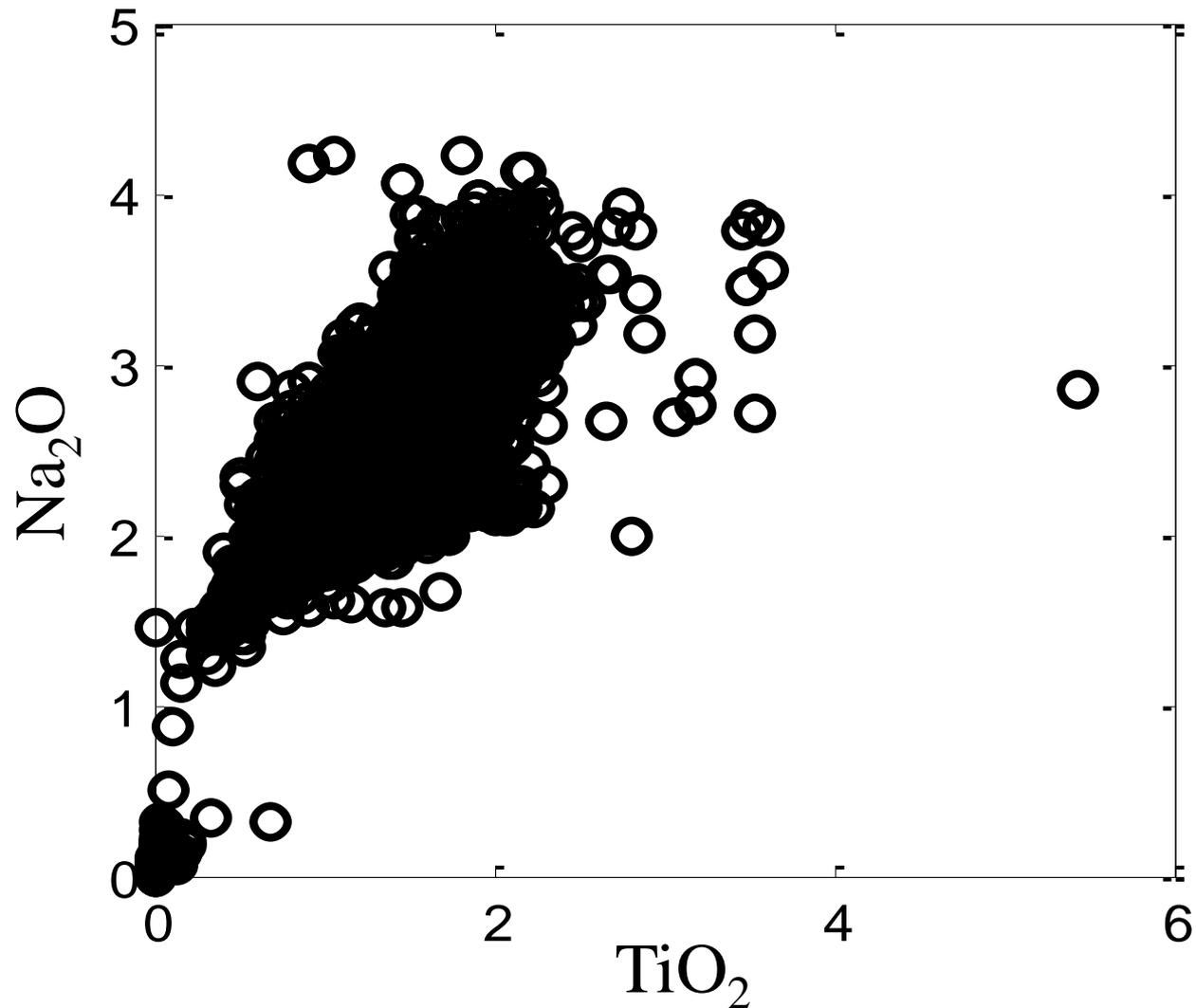
perfect positive correlation

$|R_{ij}|$ of the Atlantic Rock Dataset



Atlantic Rock Dataset

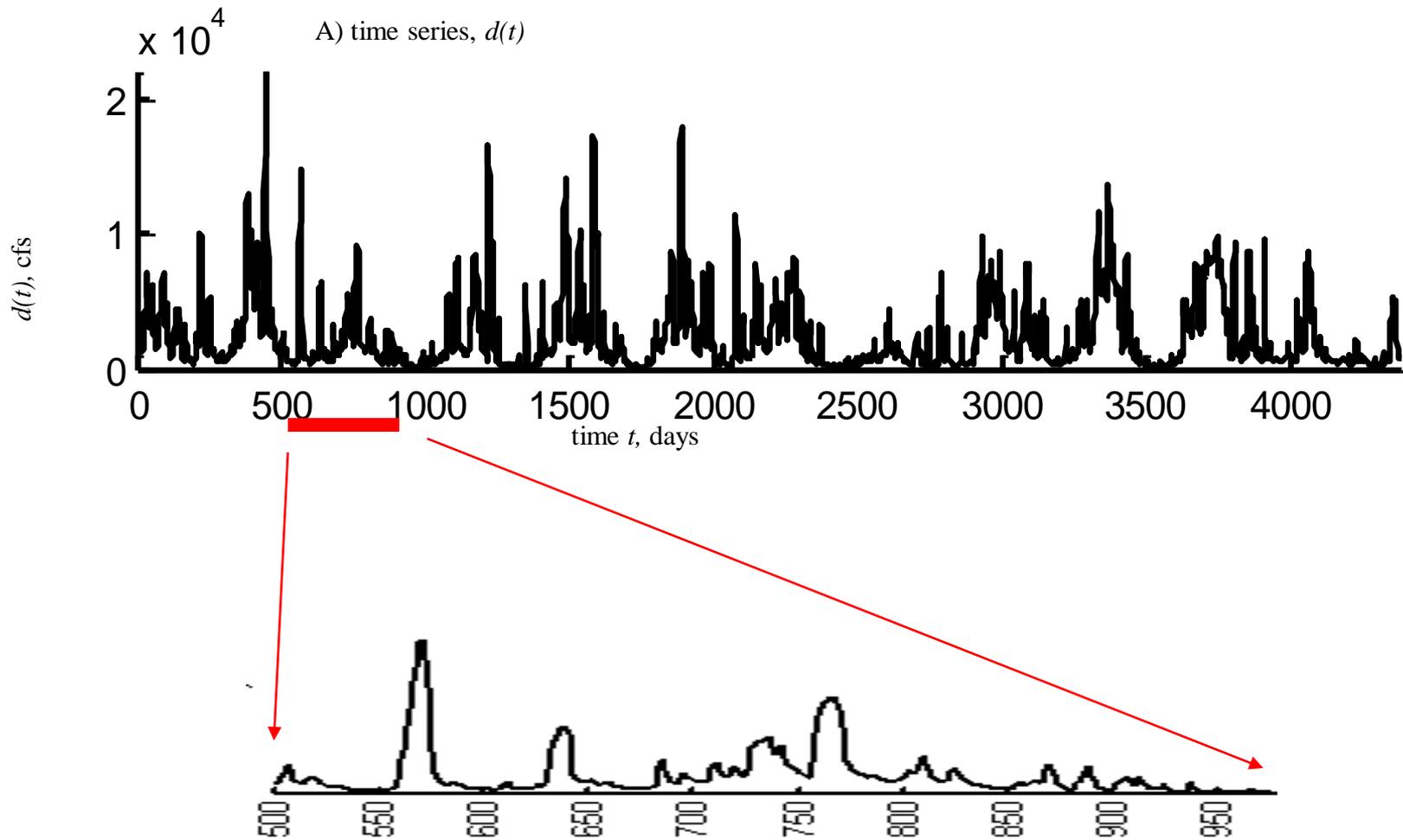
Scatter plot of TiO_2 and Na_2O



Part 2

correlations between samples
within a time series

Neuse River Hydrograph

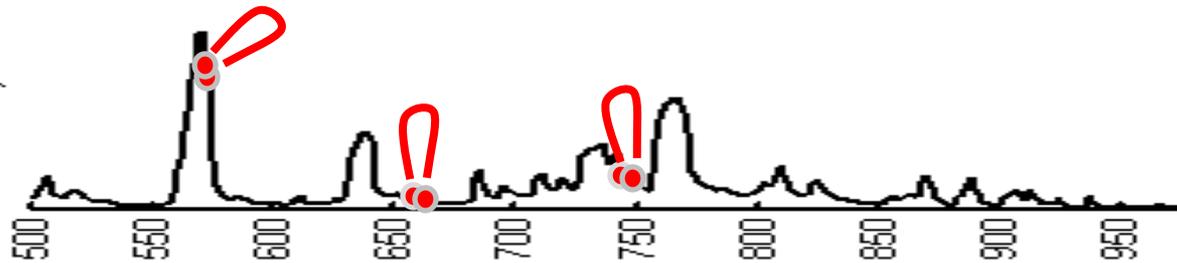
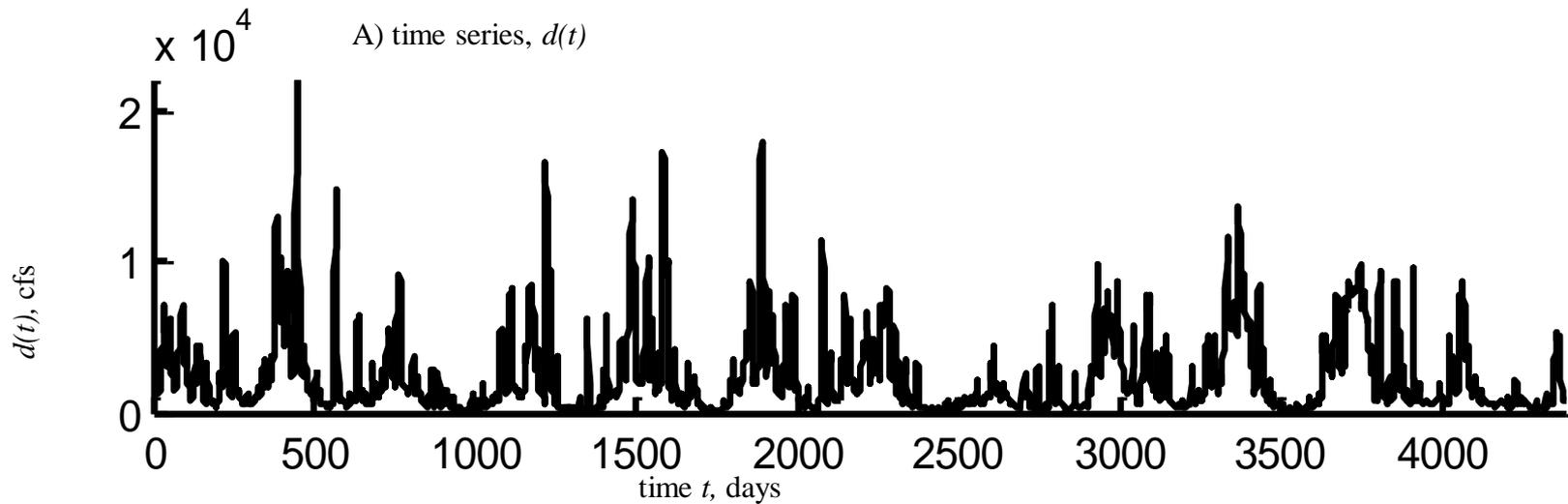


high degree of short-term correlation

*what ever the river was doing yesterday, its probably
doing today, too*

because water takes time to drain away

Neuse River Hydrograph

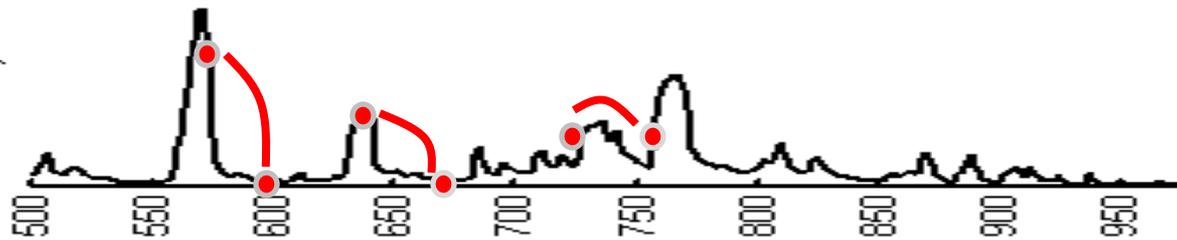
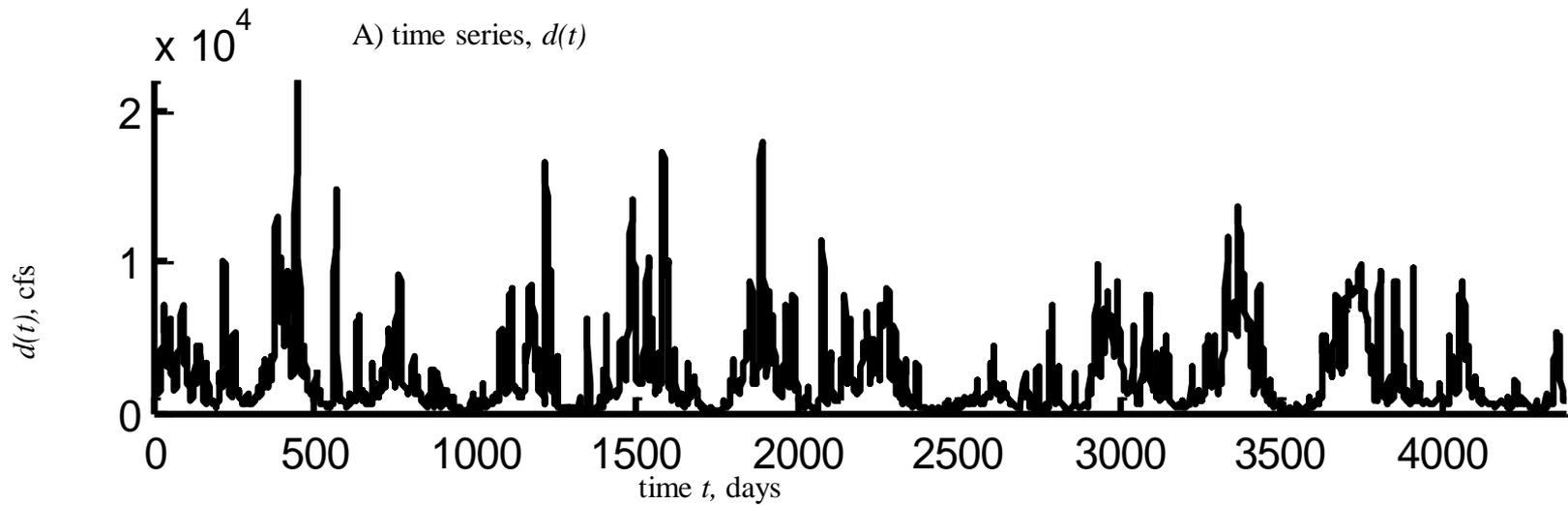


low degree of intermediate-term correlation

*what ever the river was doing last month, today it could
be doing something completely different*

because storms are so unpredictable

Neuse River Hydrograph

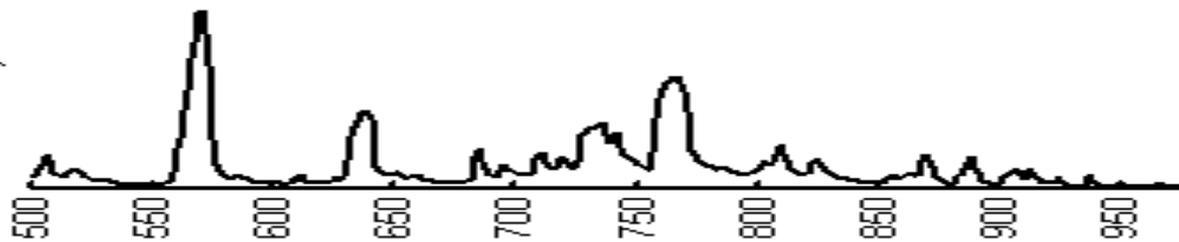
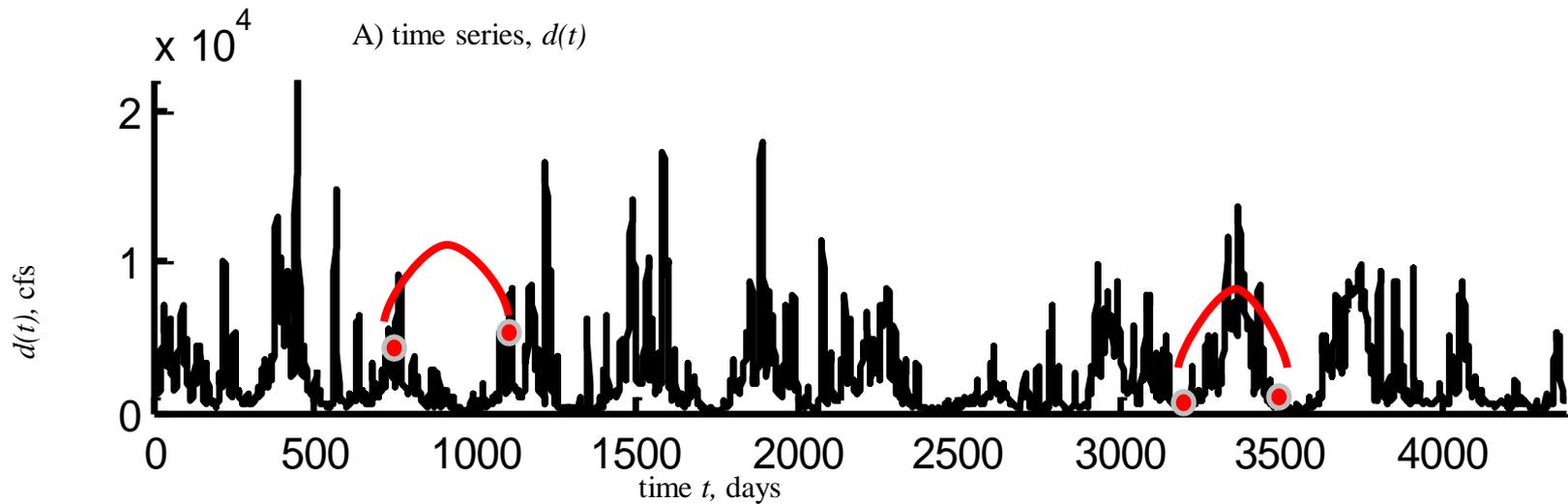


moderate degree of long-term correlation

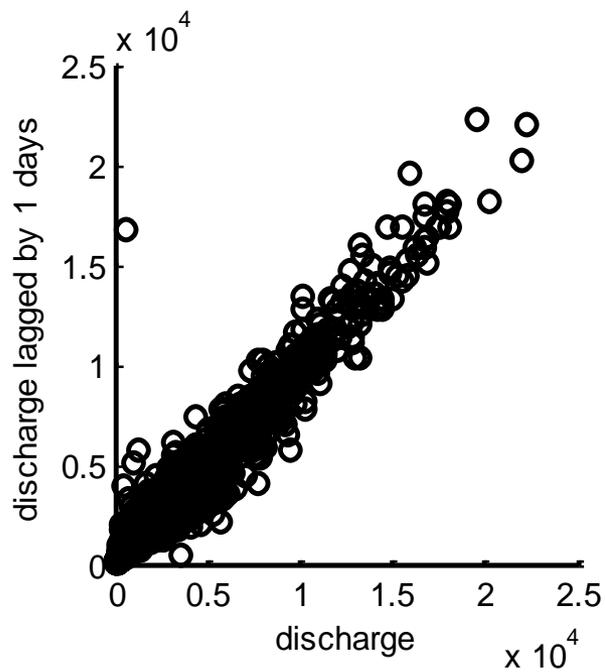
*what ever the river was doing this time last year, its
probably doing today, too*

because seasons repeat

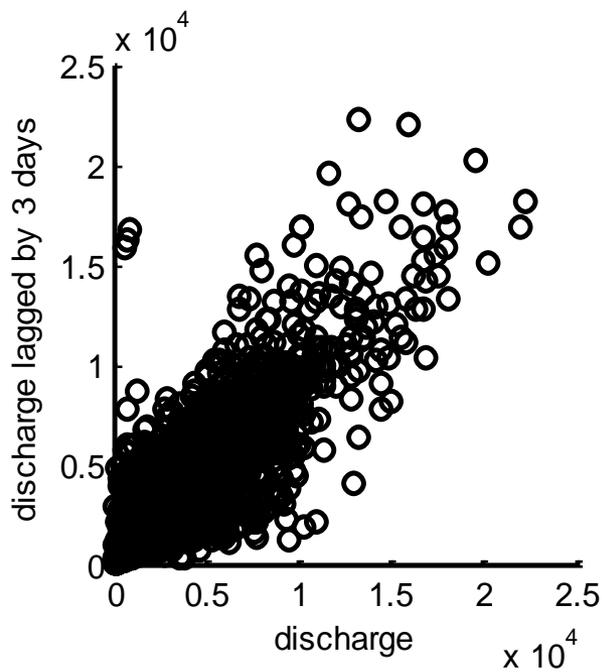
Neuse River Hydrograph



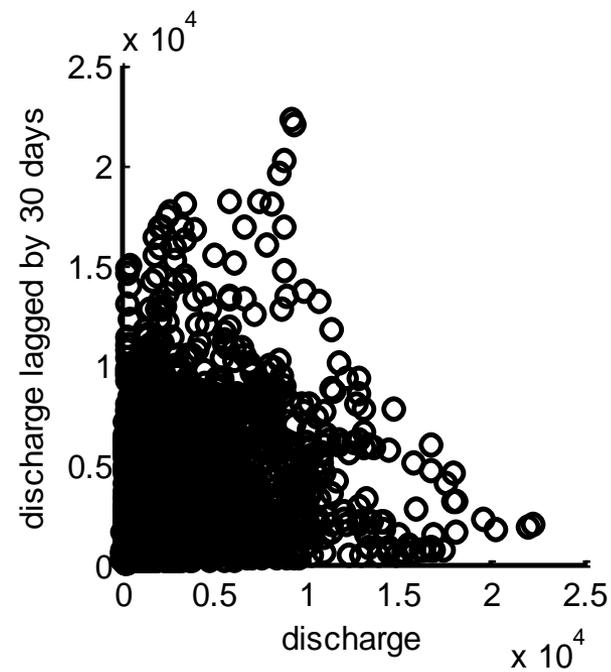
1 day



3 days



30 days



Let's assume different samples in time series are random variables and calculate their covariance

usual formula for the covariance

$$C_{ij} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (d_i - \bar{d}) (d_j - \bar{d}) p(d_i, d_j) dd_i dd_j$$

assuming time series has zero mean

$$A_{ij} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} d_i d_j p(d_i, d_j) dd_i dd_j$$

now assume that the time series is
stationary

(statistical properties don't vary with time)

so that covariance depends only on
time lag between samples

time series of length N

time lag of $(k-1)\Delta t$

using the same approximation for the sample covariance as before

$$A_{i,k+i-1} \approx \frac{1}{N - |k - 1|} \sum_{i=1}^{N-k+1} d_i d_{k+i-1} = \frac{a_k}{N - |k - 1|}$$

$$\text{with } a_k = \sum_{i=1}^{N-k+1} d_i d_{k+i-1}$$

using the same approximation for the sample covariance as before

$$A_{i,k+i-1} \approx \frac{1}{N - |k - 1|} \sum_{i=1}^{N-k+1} d_i d_{k+i-1} = \frac{a_k}{N - |k - 1|}$$

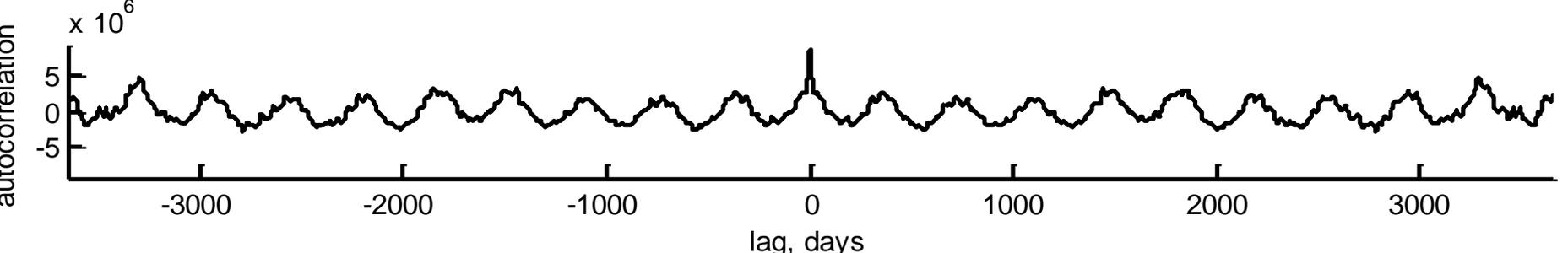
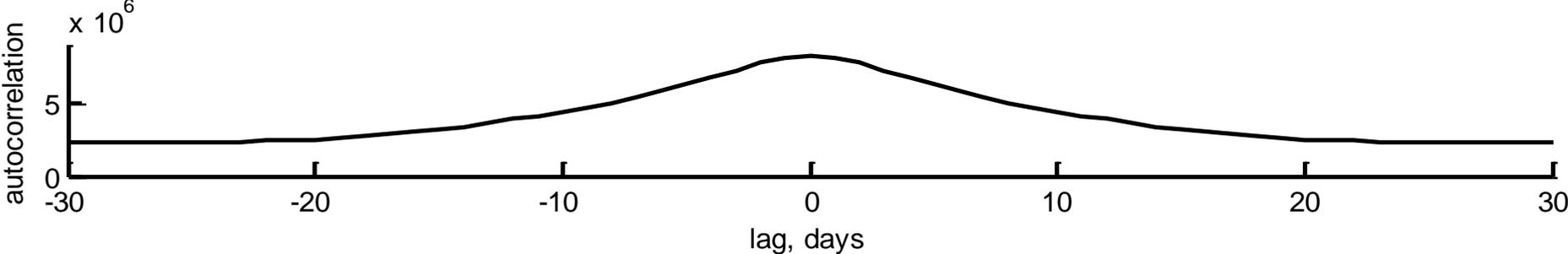
with $a_k = \sum_{i=1}^{N-k+1} d_i d_{k+i-1}$

autocorrelation
at lag $(k-1)\Delta t$

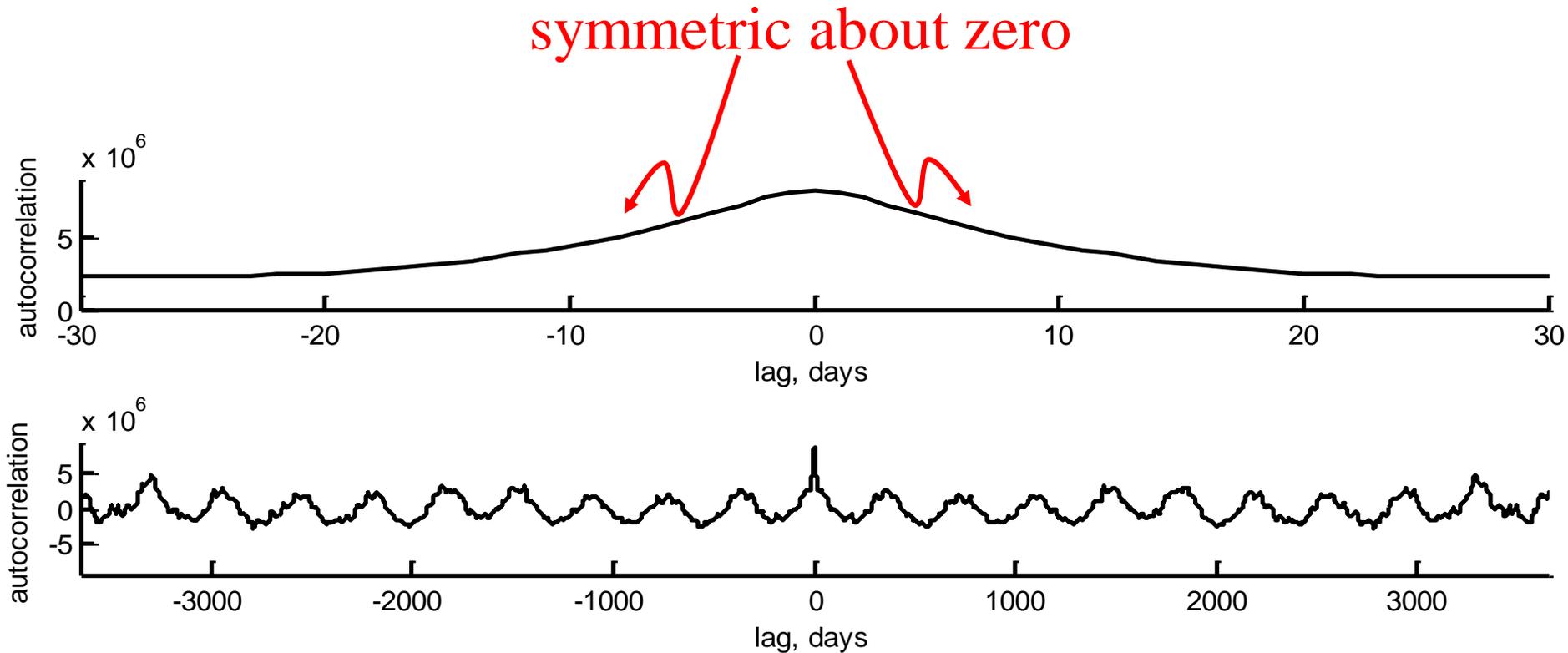
autocorrelation in MatLab

```
a = xcorr(d);
```

Autocorrelation on Neuse River Hydrograph

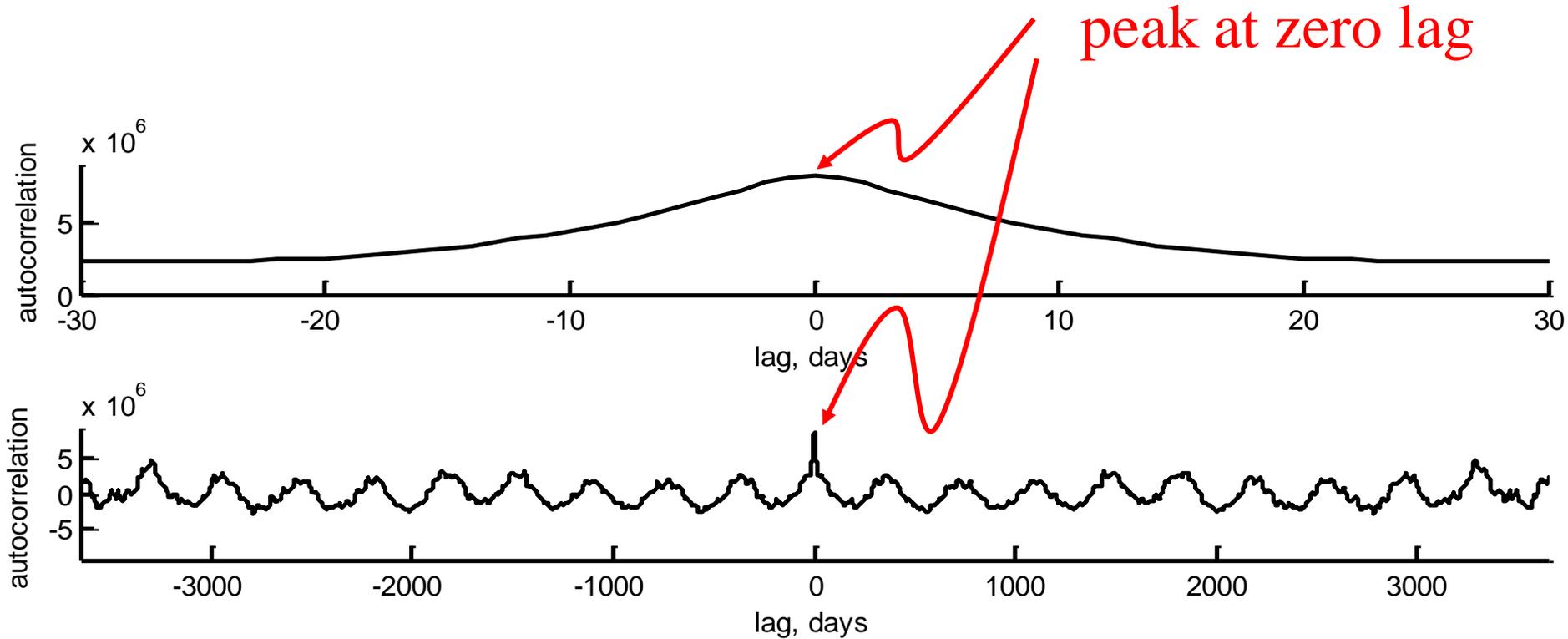


Autocorrelation on Neuse River Hydrograph



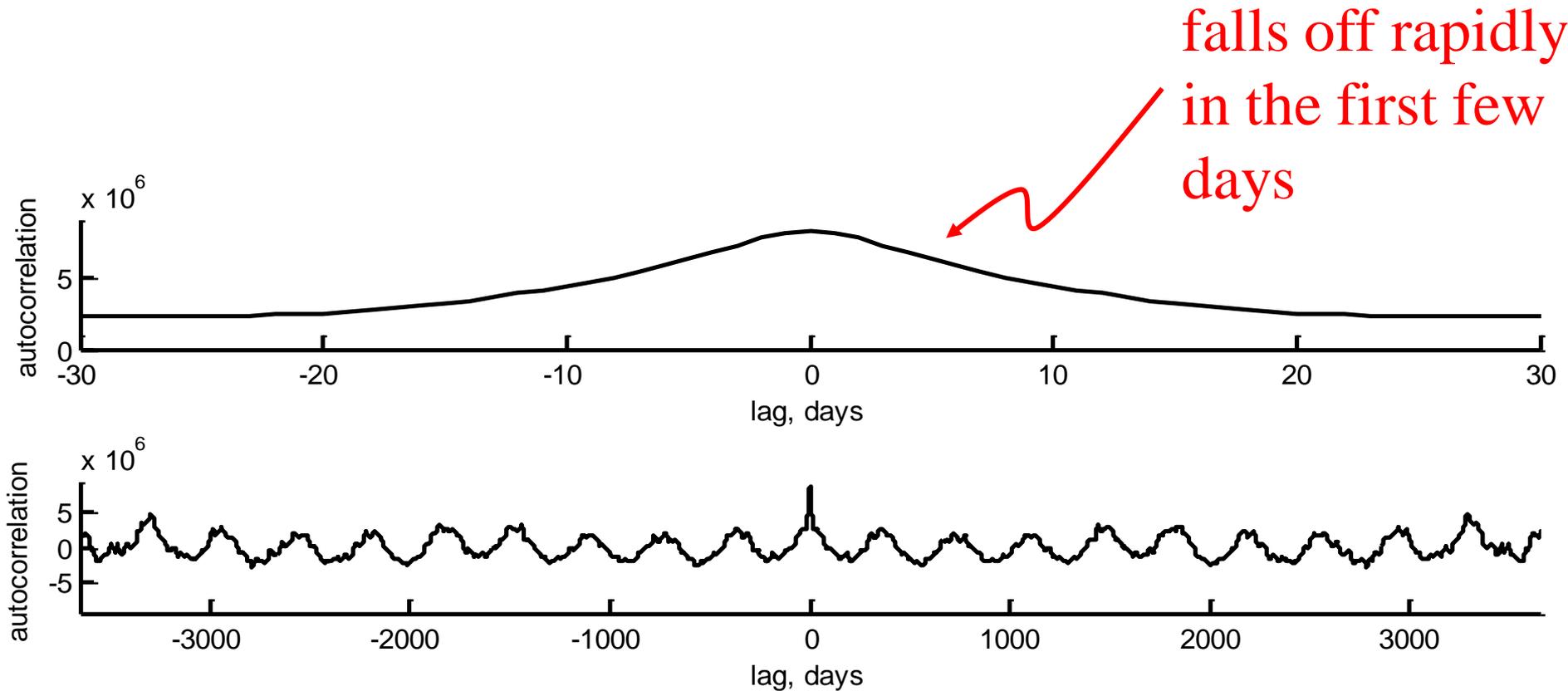
a point in a time series correlates
equally well with another in the
future and another in the past

Autocorrelation on Neuse River Hydrograph



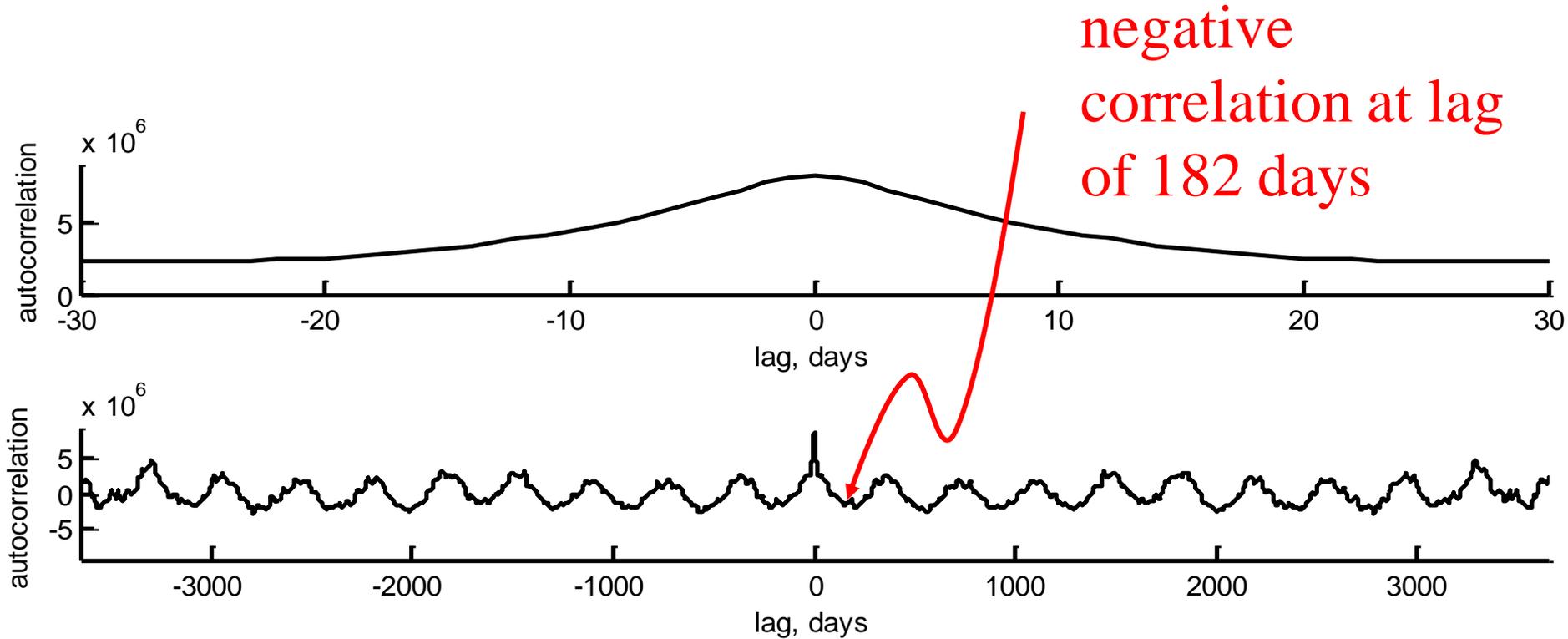
a point in time series is perfectly correlated with itself

Autocorrelation on Neuse River Hydrograph



lags of a few days are highly correlated
because the river drains the land over
the course of a few days

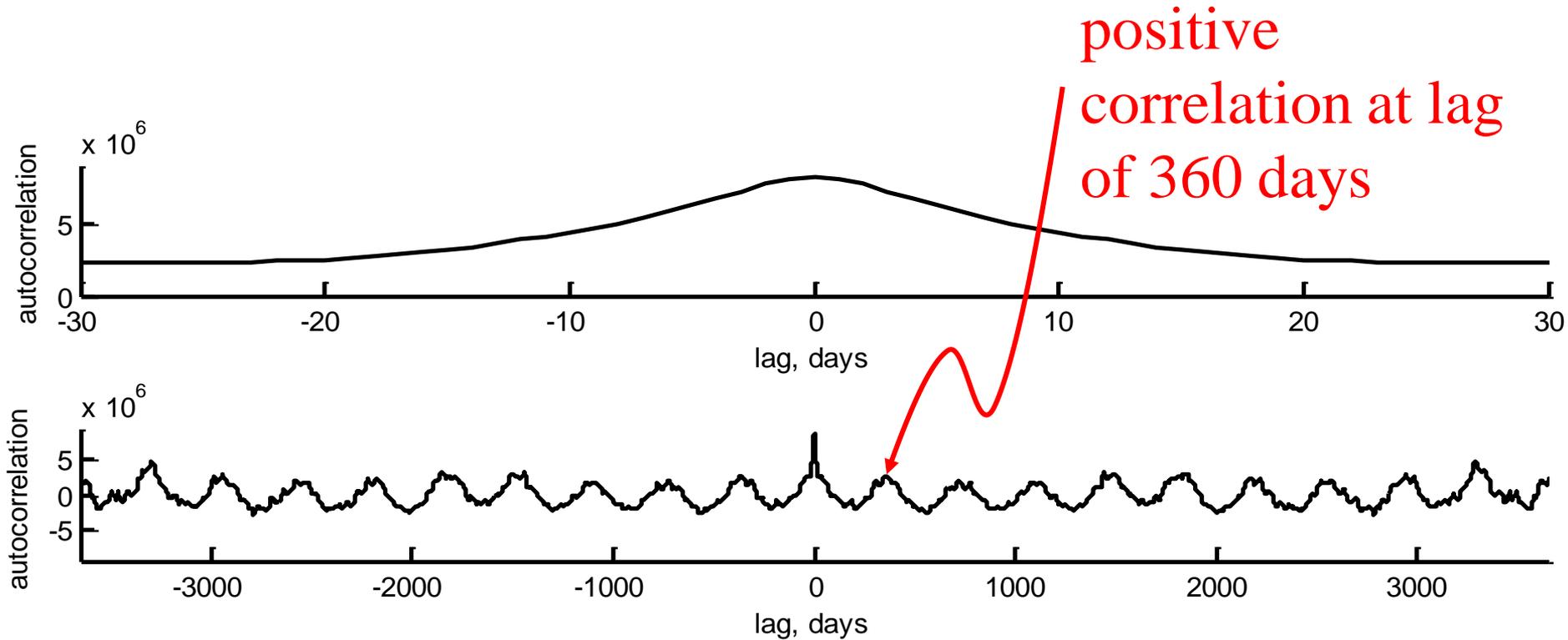
Autocorrelation on Neuse River Hydrograph



negative
correlation at lag
of 182 days

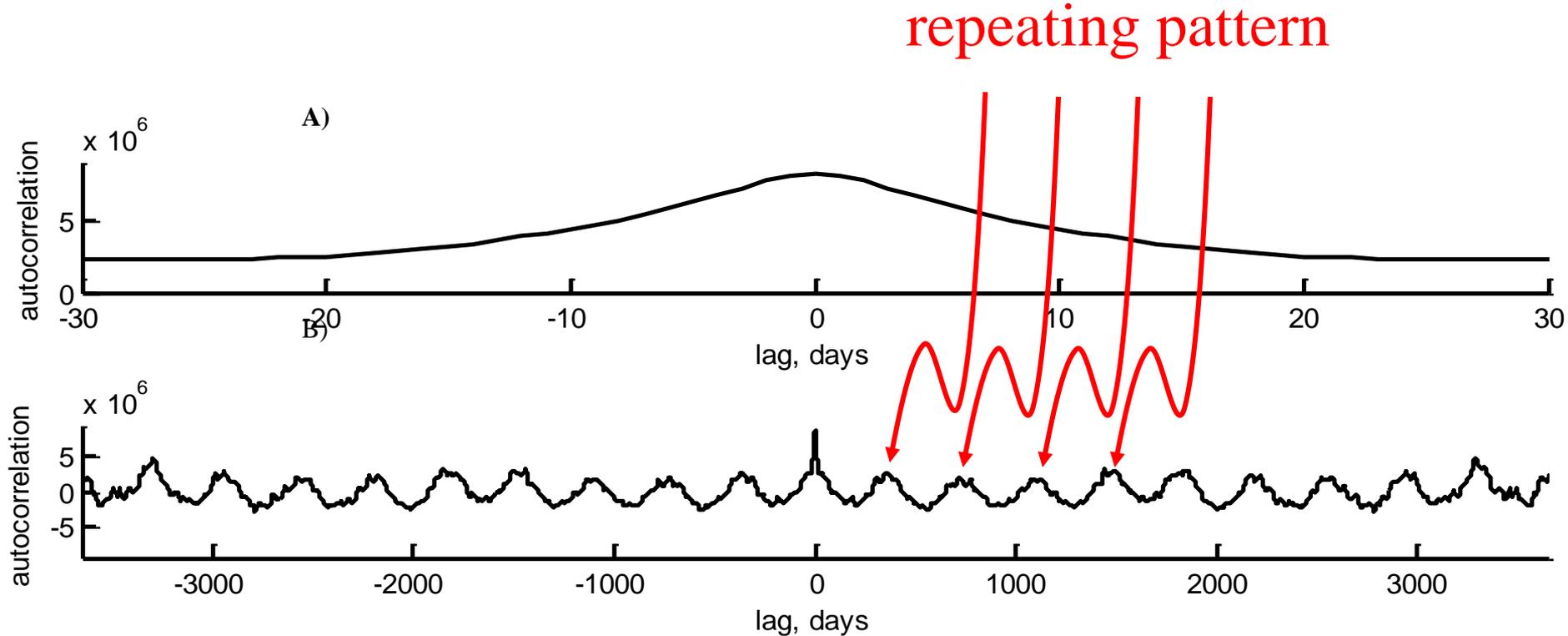
points separated by a half year are
negatively correlated

Autocorrelation on Neuse River Hydrograph



points separated by a year are positively correlated

Autocorrelation on Neuse River Hydrograph



the pattern of rainfall
approximately repeats annually

autocorrelation similar to convolution

autocorrelation

$$a_k = \sum_i d_i d_{k+i-1}$$

$$a(t) = \int_{-\infty}^{+\infty} d(\tau) d(t + \tau) d\tau$$
$$a = d \star d$$

convolution

$$\theta_k = \sum_i g_i h_{k-i+1}$$

$$\theta(t) = \int_{-\infty}^{+\infty} g(\tau) h(t - \tau) d\tau$$
$$\theta = g * h$$

autocorrelation similar to convolution

autocorrelation

$$a_k = \sum_i d_i d_{k+i-1}$$


$$a(t) = \int_{-\infty}^{+\infty} d(\tau) d(t + \tau) d\tau$$


$a = d \star d$

convolution

$$\theta_k = \sum_i g_i h_{k-i+1}$$


$$\theta(t) = \int_{-\infty}^{+\infty} g(\tau) h(t - \tau) d\tau$$


$\theta = g * h$

note difference in sign

Important Relation #1

autocorrelation is the convolution of a time series with its time-reversed self

$$a(t) = d(t) \star d(t) = d(-t) * d(t)$$

$$a(t) = d(t) \star d(t)$$

$$= \int_{-\infty}^{+\infty} d(\tau) d(t + \tau) d\tau$$

$$= \int_{-\infty}^{+\infty} d(-\tau') d(t - \tau) d\tau'$$

$$= d(-t) * d(t)$$

$$a(t) = d(t) \star d(t)$$

$$= \int_{-\infty}^{+\infty} d(\tau) d(t + \tau) d\tau$$

integral form of
autocorrelation

$$= \int_{-\infty}^{+\infty} d(-\tau') d(t - \tau) d\tau'$$

$$= d(-t) * d(t)$$

$$a(t) = d(t) \star d(t)$$

$$= \int_{-\infty}^{+\infty} d(\tau) d(t + \tau) d\tau$$

integral form of
autocorrelation

$$= \int_{-\infty}^{+\infty} d(-\tau') d(t - \tau) d\tau'$$

change of
variables, $\tau' = -\tau$

$$= d(-t) * d(t)$$

$$a(t) = d(t) \star d(t)$$

$$= \int_{-\infty}^{+\infty} d(\tau) d(t + \tau) d\tau$$

integral form of
autocorrelation

$$= \int_{-\infty}^{+\infty} d(-\tau') d(t - \tau) d\tau'$$

change of
variables, $\tau' = -\tau$

$$= d(-t) * d(t)$$

write as
convolution

Important Relationship #2

Fourier Transform of an autocorrelation
is proportional to the
Power Spectral Density of time series

$$\tilde{a}(\omega) = \tilde{d}^*(\omega) \tilde{d}(\omega) = |\tilde{d}(\omega)|^2$$

$$a(t) = d(-t) * d(t)$$

so

$$\tilde{a}(\omega) = \mathcal{F}\{d(-t)\} \tilde{d}(\omega)$$

since

$$\begin{aligned} \mathcal{F}\{d(-t)\} &= \int_{-\infty}^{+\infty} d(-t) \exp(i\omega t) dt \\ &= \int_{-\infty}^{+\infty} d(t') \exp(i(-\omega)t') dt' = \tilde{d}(-\omega) = \tilde{d}^*(\omega) \end{aligned}$$

$$a(t) = d(-t) * d(t)$$

so

$$\tilde{a}(\omega) = \mathcal{F}\{d(-t)\} \tilde{d}(\omega)$$

since

$$\begin{aligned} \mathcal{F}\{d(-t)\} &= \int_{-\infty}^{+\infty} d(-t) \exp(i\omega t) dt \\ &= \int_{-\infty}^{+\infty} d(t') \exp(i(-\omega)t') dt' = \tilde{d}(-\omega) = \tilde{d}^*(\omega) \end{aligned}$$

$$a(t) = d(-t) * d(t)$$

so

$$\tilde{a}(\omega) = \mathcal{F}\{d(-t)\} \tilde{d}(\omega)$$

since

$$\mathcal{F}\{d(-t)\} = \int_{-\infty}^{+\infty} d(-t) \exp(i\omega t) dt$$

$$= \int_{-\infty}^{+\infty} d(t') \exp(i(-\omega)t') dt' = \tilde{d}(-\omega) = \tilde{d}^*(\omega)$$

Fourier Transform of a convolution is product of the transforms

$$a(t) = d(-t) * d(t)$$

so

$$\tilde{a}(\omega) = \mathcal{F}\{d(-t)\} \tilde{d}(\omega)$$

since

$$\mathcal{F}\{d(-t)\} = \int_{-\infty}^{+\infty} d(-t) \exp(i\omega t) dt$$

$$= \int_{-\infty}^{+\infty} d(t') \exp(i(-\omega)t') dt' = \tilde{d}(-\omega) = \tilde{d}^*(\omega)$$

Fourier Transform of a convolution is product of the transforms

Fourier Transform integral

$$a(t) = d(-t) * d(t)$$

so

$$\tilde{a}(\omega) = \mathcal{F}\{d(-t)\} \tilde{d}(\omega)$$

since

$$\mathcal{F}\{d(-t)\} = \int_{-\infty}^{+\infty} d(-t) \exp(i\omega t) dt$$

$$= \int_{-\infty}^{+\infty} d(t') \exp(i(-\omega)t') dt' = \tilde{d}(-\omega) = \tilde{d}^*(\omega)$$

transform of
variables, $t' = -t$

Fourier Transform of a
convolution is product of the
transforms

Fourier
Transform
integral

$$a(t) = d(-t) * d(t)$$

so

$$\tilde{a}(\omega) = \mathcal{F}\{d(-t)\} \tilde{d}(\omega)$$

since

$$\mathcal{F}\{d(-t)\} = \int_{-\infty}^{+\infty} d(-t) \exp(i\omega t) dt$$

$$= \int_{-\infty}^{+\infty} d(t') \exp(i(-\omega)t') dt' = \tilde{d}(-\omega) = \tilde{d}^*(\omega)$$

Fourier Transform of a convolution is product of the transforms

Fourier Transform integral

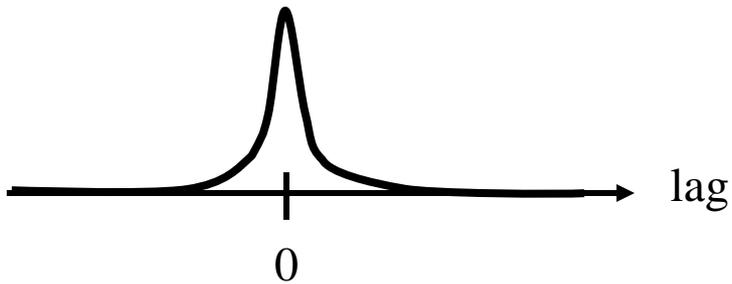
transform of variables, $t' = -t$

symmetry properties of Fourier Transform

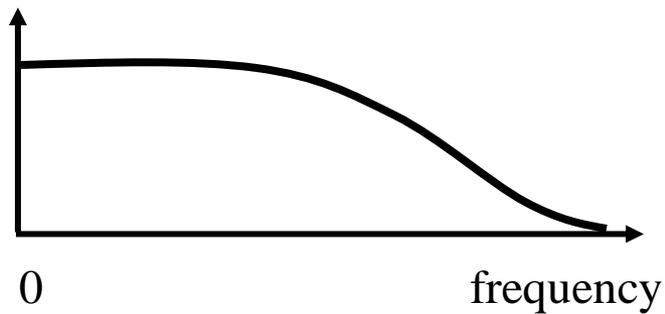
Summary



rapidly fluctuating
time series

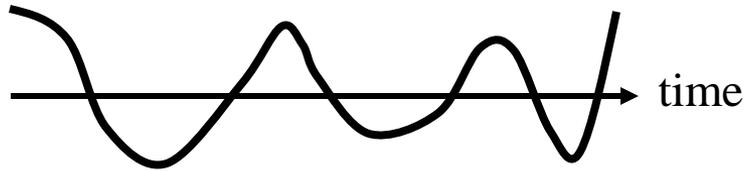


narrow
autocorrelation
function

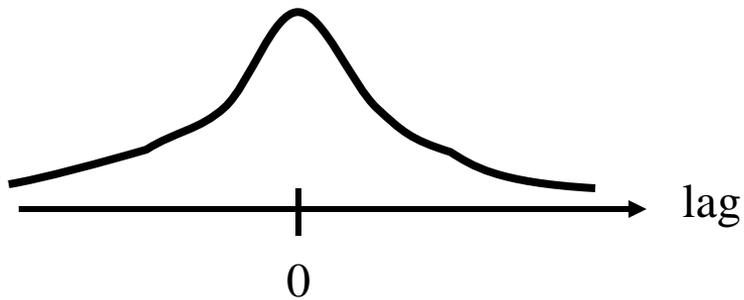


wide spectrum

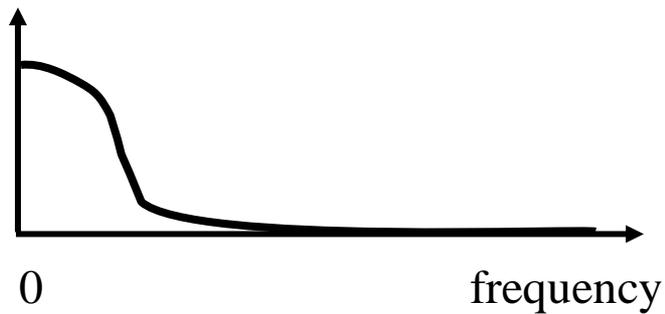
Summary



slowly fluctuating
time series



wide
autocorrelation
function



narrow spectrum