Tues, 4-09-19

10. Lagged Correlation

* Lightning talk
* Feedback on A9

1. **Lagged relationships between time series**
2. **Cross-correlation function**
3. **Significance of cross-correlation**
4. **Alternative ways to a confidence interval on cross-correlations**

Read notes_10.pdf

A9 Feedback

1. Download A9x.pdf from D2L
2. Automatic points, for running assignment and having uploaded by due time, is already marked in parentheses at top of first page
3. Each assignment has maximum possible 10 points; if you make no deductions, score is 10/10
4. A9x is color coded for points; purple=1; yellow=0.5; blue=0.5
5. Open your copy of the same assignment pdf you uploaded
6. In Acrobat Reader, using “Add text box,” mark in right margin for deductions only, with deduction and segment reference : (eg., -0.5 A); round to tenths in deductions (e.g., no -0.25)
7. At top of your pdf, mark grade like this : 9.5/10
8. If necessary, put any comments at top near the grade
9. Upload your self-graded pdf to folder A9_graded in D2L

Lagged relationships; example from climatology (p 1 of 2)

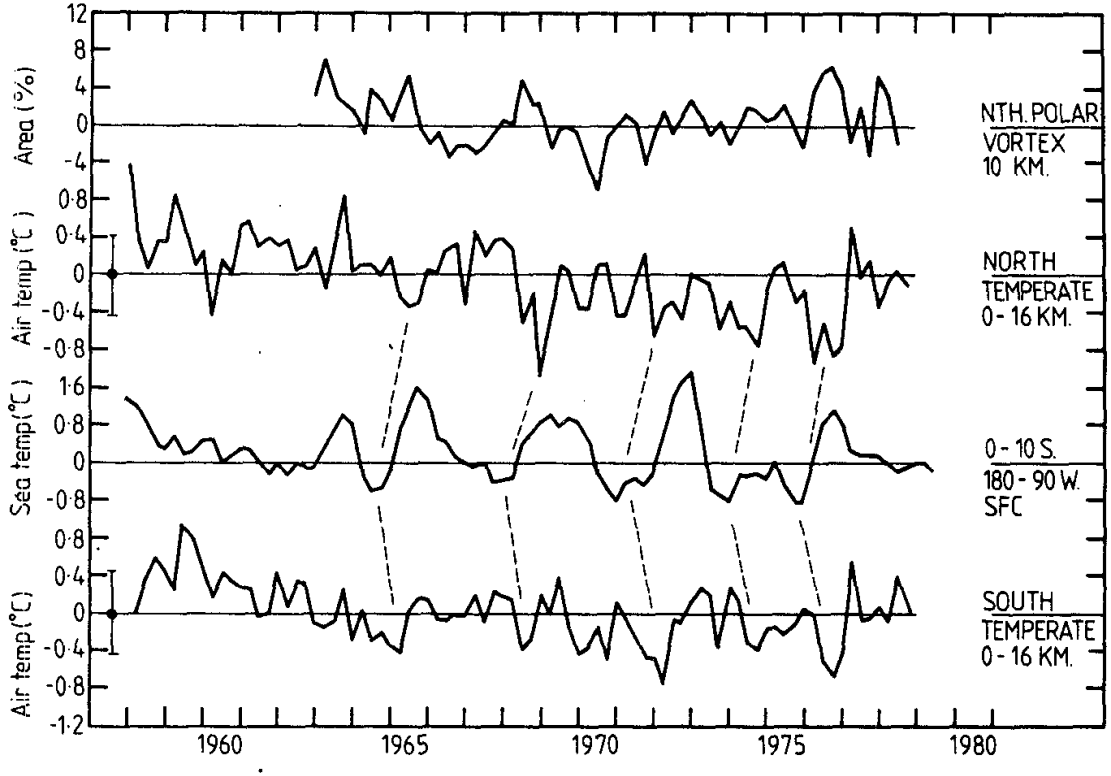


FIG. 5. Comparison of time variations in SST*, zonally averaged temperature in the 0–16 km layer of north and south temperate latitudes, and area of the north circumpolar vortex at a height of 10 km. There is about a 5% chance that the true mean value of seasonal air temperature lies outside the extent of the vertical bars at left.

Angell, J. K., 1981: Comparison of variations in atmospheric quantities with sea surface temperature variations in the equatorial eastern Pacic. Mon. Wea. Rev., 109, 230-243.

Example from climatology (p 2 of 2)

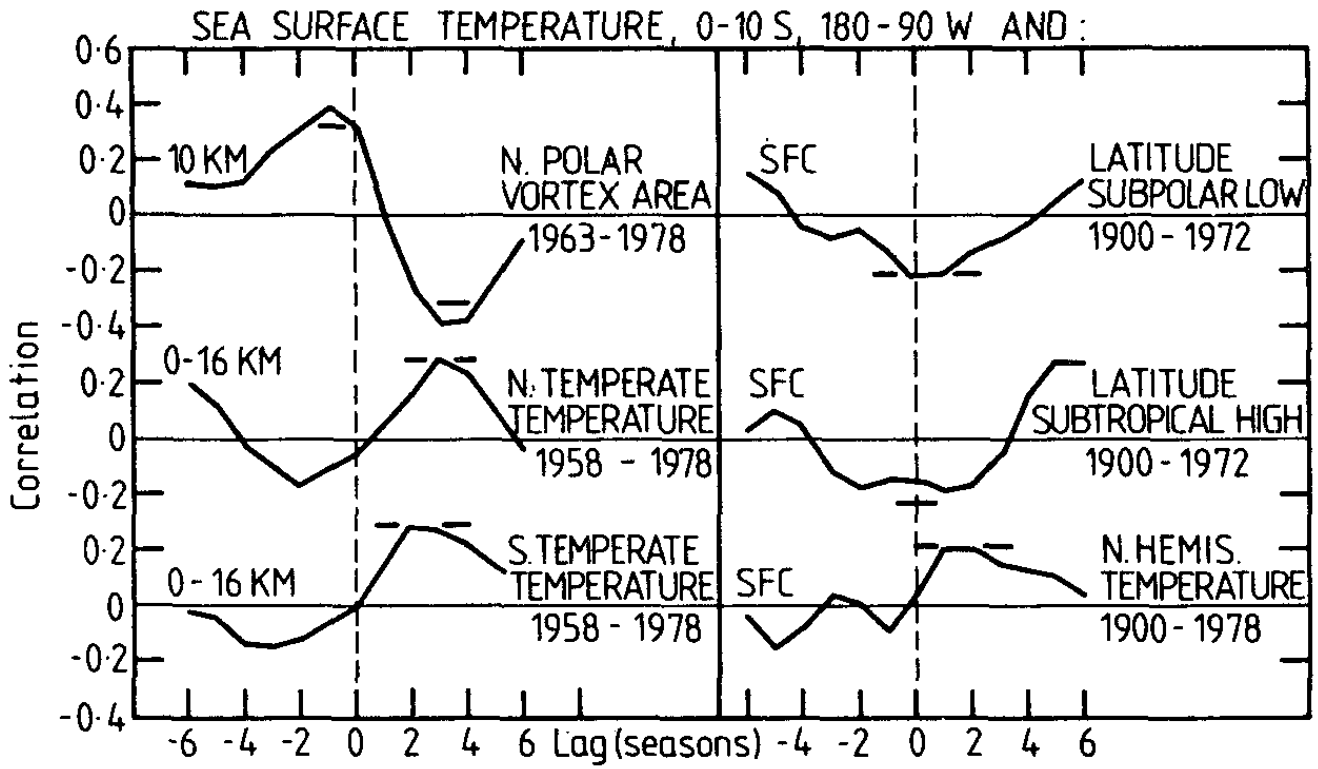
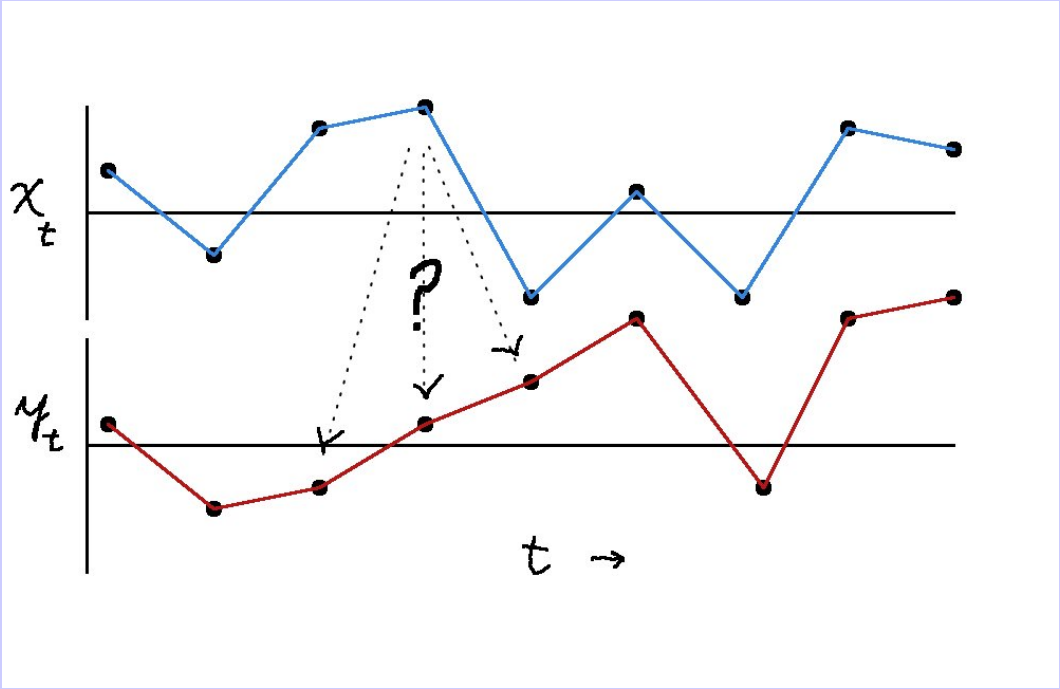


FIG. 6. Lag correlation between SST* and the quantities of Fig. 5 (left), as well as estimated latitude of subpolar low and subtropical high (right) for given years of record. The horizontal bars represent correlations significant at the 95% level taking into account the serial correlation in the data.

Angell, J. K., 1981: Comparison of variations in atmospheric quantities with sea surface temperature variations in the equatorial eastern Pacic. Mon. Wea. Rev., 109, 230-243.

The Problem



- Two time series, x_t and y_t
- How is x_t related to y_{t+k}, where k is a lag, and the lag can in general be negative or positive?
- Till now we have looked only at contemporaneous correlation (k=0)
- General case can be studied with cross-correlation function , $r_{x,y}(k)$

Cross-correlation function

$$r_{x,y}(k) \equiv \text{correlation of } x_t \text{ with } y_{t+k}$$

- Useful for studying lag-lead relationships
- Confidence intervals complicated by autocorrelation in x_t and y_t
- Computed from cross-covariance

Cross-covariance function (ccvf)

$x_t, y_t, \quad t = 1, N$ are time series of length N

Two equations -- one for positive lags and one for negative lags

$$c_{x,y}(k) = \frac{1}{N}\sum_{t=1}^{N-k}\left(x_t - \bar{x}\right)\left(y_{t+k} - \bar{y}\right) \qquad [k = 0,1,\ldots,(N-1)]$$

$$c_{x,y}(k) = \frac{1}{N}\sum_{t=1-k}^{N}\left(x_t - \bar{x}\right)\left(y_{t+k} - \bar{y}\right) \qquad [k = -1,-2,\ldots,-(N-1)]$$

- Average product of departures lagged in time
- Unlike acvf, non-symmetrical
- Scaled by variances to get cross-correlation

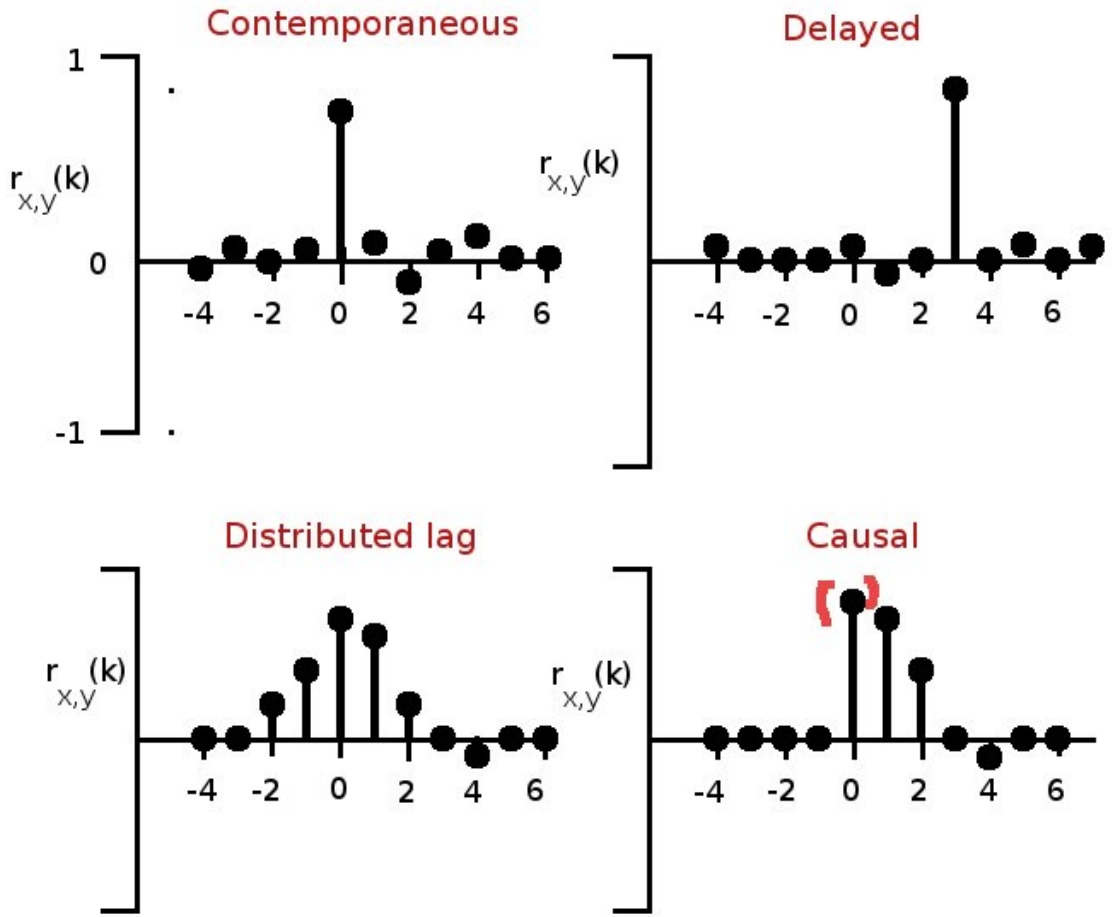
Cross-correlation function (ccf)

$c_{x,y}(k)$: sample cross-covariance between at lag k

$$r_{x,y}(k) = \frac{c_{x,y(k)}}{\sqrt{c_{x,x(0)} c_{y,y}(o)}}$$: sample cross-correlation

$c_{x,x}(0)$ and $c_{y,y}(0)$ are sample variances of x_t and y_t

Some Idealized Patterns of Cross-correlation



Statistical significance
Is the ccf significantly different from zero?

- Two time series x_t and y_t, $t = 1,N$
- Can estimate confidence interval given an estimate of the variance of $r_{x,y}(k)$, the sample cross-covariance at lag k
- But, this variance is a complicated function of the unknown population autocorrelations of x_t and y_t, as well as the population cross-correlations at lags other than lag k
- "Bartlett's equation" (Box and Jenkins, p. 376) shows just how complicated:

Bartlett's equation

$$\mathrm{var}[r_{x,y}(k)] \simeq \frac{1}{N-k}\sum_{v=-\infty}^{\infty}[\rho_{x,x}(v)\,\rho_{y,y}(v)+\rho_{x,y}(k+v)\rho_{x,y}(k-v)+$$

$$\rho_{x,y}^2(k)\left\{\rho_{x,y}^2(v)+\frac{1}{2}\rho_{x,x}^2(v)+\frac{1}{2}\rho_{y,y}^2(v)\right\}-$$

$$2\rho_{x,y}(k)\left\{\rho_{x,x}(v)\rho_{x,y}(k+v)+\rho_{x,y}(-v)\rho_{y,y}(v+k)\right\}]$$

$N =$ sample size

$k =$ lag

$\rho_{x,x}$ and $\rho_{y,y}$ are population autocorrelations

$\rho_{x,y}$ are population cross-correlations

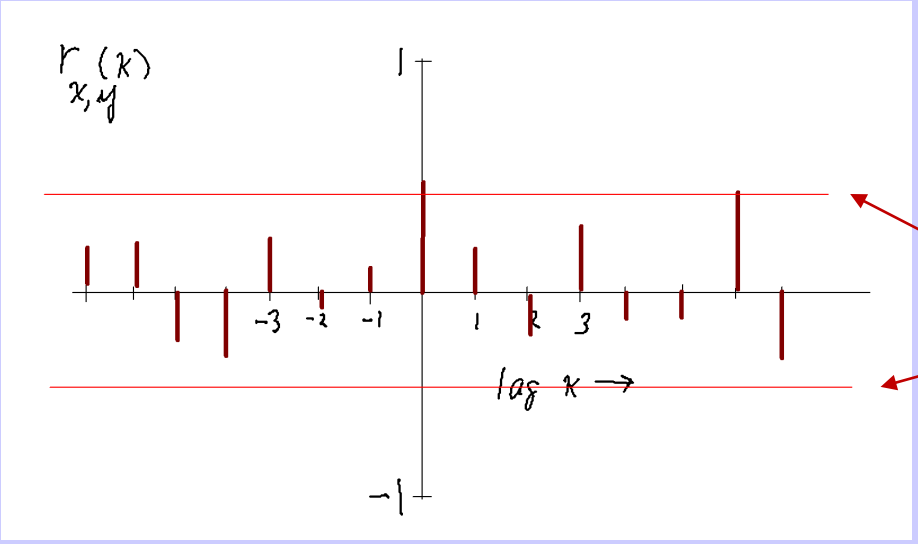
Confidence interval for $r_{x,y}(k)$ under simple conditions

If

1) Populations X and Y normal

2) $\rho_{x,y}(k) = 0$ for all k ← *Populations not cross-correlated*

3) $\rho_{x,x}(k) = 0,\ \rho_{y,y}(k) = 0$ for all k ← *Populations not autocorrelated*



Under these assumptions, could draw 95% CI as horizontal lines at

$$0 \pm \frac{2}{\sqrt{N}},$$ where N is sample size

But generally, processes are autocorrelated

- Need estimated variance of sample cross-correlations to get estimate of confidence interval
- Can use Bartlett's equation, but can drop many terms

Must resort to Bartlett's equation, but can simplify

Assume $\rho_{x,y}(\nu) = 0$ for all lags ν

- Processes assumed not cross-correlated
- Still allows for each to be autocorrelated

Resulting simplified version…

$$\mathrm{var}[r_{x,y}(k)] \simeq \frac{1}{N-k} \sum_{v=-\infty}^{\infty} \rho_{x,x}(v)\rho_{y,y}(v)$$

The variance of the sample cross-correlation at lag k therefore now depends on a summation of products of the population autocorrelations of the two processes

Box and Jenkins, p 377 (see ref in notes)

Simplification leads to “Quenouille’s equation”

- Substitute sample autocorrelations for population autocorrelations
- Truncate summation at some lag when terms become small
- Use estimated variance of cross-correlations, and assumption that they are normally distributed, to get 95% confidence interval.
- Can get horizontal lines as an expanded confidence interval that takes into account reduced number of independent observations due to autocorrelation in the two individual series

"Quenouille's equation" for effective sample size

$$\mathrm{var}[r_{x,y}(k)] \simeq \frac{1}{N-k}\sum_{v=-\infty}^{\infty}\rho_{x,x}(v)\rho_{y,y}(v)$$

$$\mathrm{var}[r_{x,y}(k)] \simeq \frac{1}{N-k}\sum_{j=-\infty}^{\infty}r_{x,x}(j)r_{y,y}(j)$$

Let r_j stand for $r_{x,x}(j)$

Let r'_j stand for $r_{\mathrm{y,y}}(j)$

From Quenouille

$$N' = N/\left(1+2r_1r'_1+2r_2r'_2+\ldots\right)$$

Truncate summation at some lag when product becomes small (e.g., 4 terms in Angell (1981))

Approximate 95% CI for $r_{x,\mathrm{y}}(k)$: $0 \pm \frac{2}{\sqrt{N'}}$

Three alternative ways to a confidence interval for

$$r_{x,y}(k)$$

1. Quenouille's formula (see eq (1) in Angell (1981))
    1. Substitute sample autocorrelations in the previous eqn
    2. Simplify the equation, and compute an effective sample size
    3. Use the effective sample size and normal assumption for horizontal CI that takes autocorrelation into account
2. "EQUAL FOOTING APPROACH" : Prewhiten the two time series to get rid of autocorrelation, and assume approximate 95% CI of 2/sqrt(N), where N is sample size
3. "SYSTEMS APPROACH": Assume linear system, with one series as input and the other as output; whiten the input and filter the output by the whitening model; compute cross-correlation of whitened input and filtered output and assume approximate 95% CI of 2/sqrt(N), where N is sample size

Equal-footing approach

1) Prewhiten x_t and y_t

2) Compute ccf of prewhitened series and use relationships that apply to non-autocorrelated series

$$x_t \rightarrow e_{t,x}$$

$$y_t \rightarrow e_{t,\mathrm{y}}$$

Can AR model the two original time series to get these residual series

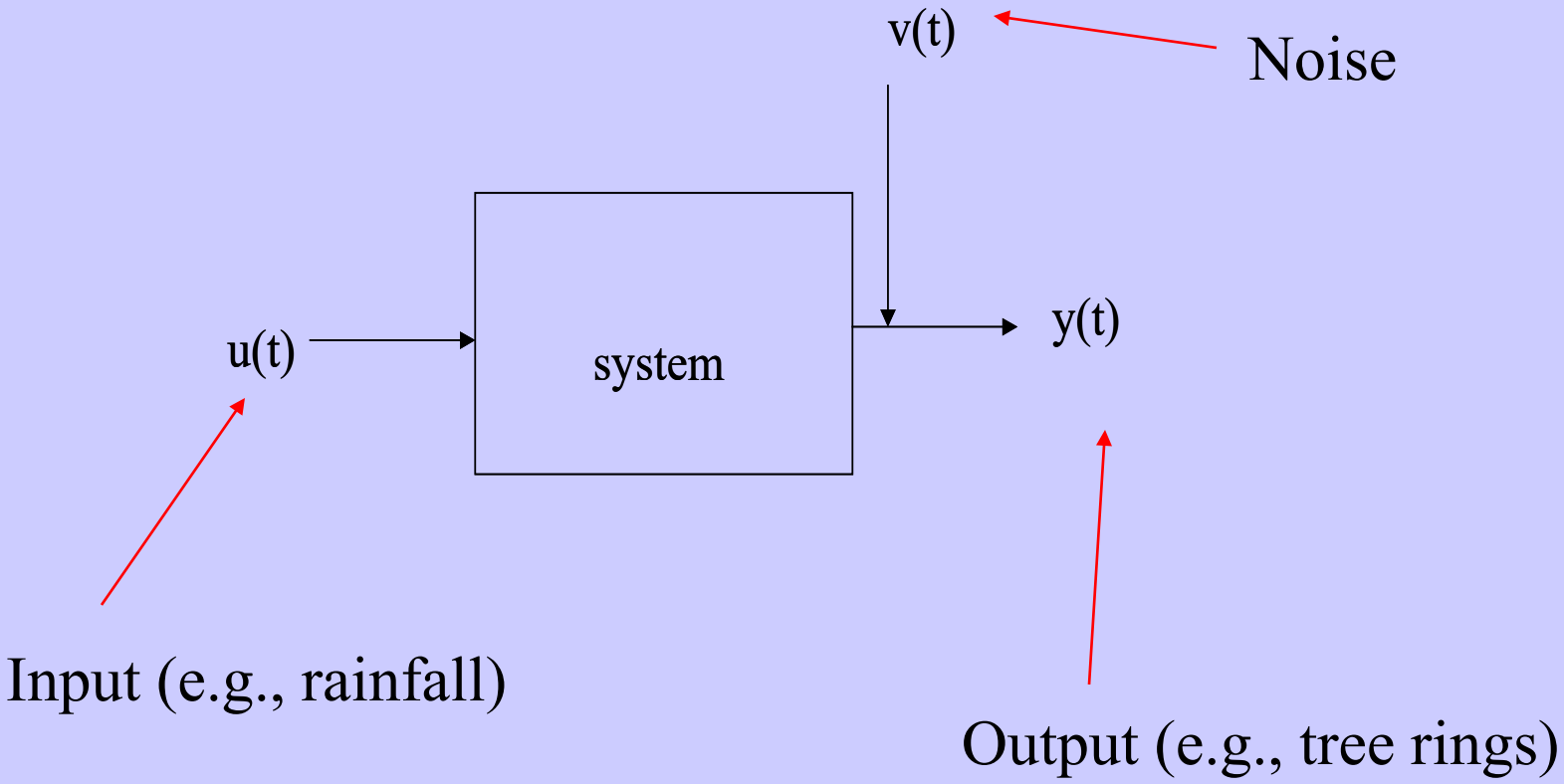
$$r_{e_{t,x},e_{t,x}}(k) \sim \mathrm{N}\left(0, \frac{1}{N}\right)$$

95% CI: $0 \pm \frac{2}{\sqrt{N}}$

Systems approach

- Regard the two time series as input to and output from some black-box linear system
- Let the time series be u_t and y_t (for consistency with Matlab's System Identification Toolbox, using u here instead of x)
- Output conceptually is corrupted by noise

Systems approach: concept is linear input-output system corrupted by noise



Systems approach: steps

1) Prewhiten the input with an AR model, $u_t \rightarrow \alpha_t$

2) Filter the output with same AR model

$y_t \rightarrow \beta_t$

3) Cross-correlate the prewhitened input and filtered output

$r_{\alpha,\beta}(k)$

4) Apply normal assumption as if dealing with non-autocorrelated series to get 95% CI

$$0 \pm \frac{2}{\sqrt{N}}$$