Lecture 2:

Covariance and correlation

Shane Elipot

The Rosenstiel School of Marine and Atmospheric Science,
University of Miami

Created with {Remark.js} using {Markdown} + {MathJax}

References

[1] Bendat, J. S., & Piersol, A. G. (2011). *Random data: analysis and measurement procedures* (Vol. 729). John Wiley & Sons.

[2] Thomson, R. E., & Emery, W. J. (2014). *Data analysis methods in physical oceanography*. Newnes. dx.doi.org/10.1016/B978-0-12-387782-6.00003-X

[3] Taylor, J. (1997). *Introduction to error analysis, the study of uncertainties in physical measurements*.

[4] Press, W. H. et al. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

[5] Kanji, G. K. (2006). *100 statistical tests*. Sage.

[6] von Storch, H. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research*, Cambridge University Press

Lecture 2: Outline

1. Covariance & correlation
2. Lagged covariance & correlation
3. A quick look at "A leisurely look at the bootstrap, the jackknife, and cross-validation"
4. Covariance and correlation of bivariate variables

1. Covariance & Correlation

Covariance (definitions)

Whereas we previously dealt with a single r.v. x, we now deal with two r.vs., x and y. In particular, we are interested in evaluating how much they covary, possibly to make some statement about a causation from one variable to the other, perhaps explained by a dynamical relationship.

Perhaps the first quantity to consider is the *covariance of x and y*:

$$\mathrm{Cov}(x, y) = C_{xy} \equiv E[(x - \mu_x)(y - \mu_y)]$$

The *variance of x* is a particular case of covariance when y and x are the same r.v.:

$$\mathrm{Cov}(x, x) = C_{xx} \equiv E[(x - \mu_x)(x - \mu_x)] = E[(x - \mu_x)^2] = \mathrm{Var}(x) = \sigma_x^2$$

Covariance (definitions)

$$
\begin{aligned}
C_{xy} &= E[(x - \mu_x)(y - \mu_y)] = E[xy] - \mu_x \mu_y \\
&= \int_{-\infty}^{+\infty} (x - \mu_x)(y - \mu_y) p(x, y) \, dx dy
\end{aligned}
$$

$p(x, y)$ is called the *joint probability density function* or *joint PDF*.

If $p(x, y) = p(x)p(y)$ where $p(x)$ and $p(y)$ are the PDFs of x and y, respectively, then it is said that x and y are *independent*.

Just like we were able to build histograms from samples X_n of a single r.v. x, we can also build 2D histograms from pairs of samples (X_n, Y_n) in order to estimate the joint PDF.

Correlation

Since the r.v. x and y can be of different nature or magnitude, we can consider the normalized covariance, that is the *correlation* between x and y

$$\rho_{xy} = \frac{C_{xy}}{\sqrt{C_{xx}C_{yy}}} = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\{E[(x-\mu_x)^2]E[(y-\mu_y)^2]\}^{1/2}}$$

Since we have the property that $|C_{xy}| \leq \sigma_x \sigma_y$, the correlation is a number between -1 and 1.

If $\rho_{xy} = 0$ it is said that the variable x and y are *uncorrelated*.

If two r.vs. are independent $[p(x,y) = p(x)p(y)]$ **then** they are also uncorrelated, but if two r.vs. are uncorrelated, they are not necessarily independent (i.e. maybe $p(x,y) \neq p(x)p(y)$)

Covariance & Correlation: estimation

Just like s_x^2 is an unbiased estimate of σ_x^2, an unbiased estimate of the covariance is

$$s_{xy} = \widehat{C}_{xy} = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \overline{X})(Y_n - \overline{Y})$$

and an estimate of ρ_{xy} is

$$r_{xy} = \hat{\rho}_{xy} = \frac{s_{xy}}{\sqrt{s_x^2\, s_y^2}} = \frac{\sum_{n=1}^{N}(X_n - \overline{X})(Y_n - \overline{Y})}{\left[\sum_{n=1}^{N}(X_n - \overline{X})^2 \sum_{n=1}^{N}(Y_n - \overline{Y})^2\right]^{1/2}}$$

r_{xy} is called the *Pearson's correlation coefficient*. It measures the relative strength of a *linear relationship* between x and y (see Lecture 3). A *nonlinear relationship* or *noise* will make r_{xy} tends to zero.

Correlation: significance

Since x and y are r.vs., r_{xy} is also a r.v. with a given distribution.

Typically, one wants to test if r_{xy} is different from zero. Kanji (2006), reference [5] recommends to calculate the following test statistic (test 12)

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{N - 2} \sim t(0, N - 2)$$

which follows the Student's t-distribution with $N - 2$ degrees of freedom.

Correlation: confidence interval

Alternatively, one may want to derive CIs for r_{xy} by using the Fisher transformed variable:

$$w = \tanh^{-1} r_{xy} = \frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right) \sim \mathcal{N}\left[\frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right), \frac{1}{\sqrt{N-3}}\right]$$

Since $r_{xy} = \tanh w$, the CI derived for w can be used to calculate CIs for r_{xy}.

The test statistic

$$z = \frac{w - w_0}{1/\sqrt{N-3}} \sim \mathcal{N}(0,1)$$

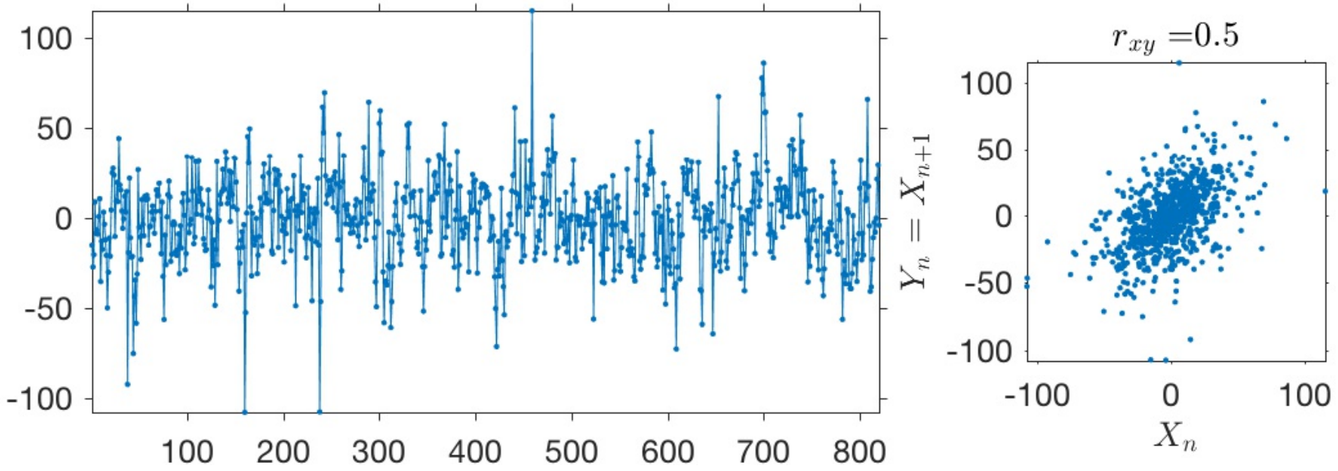
can be used to test the null hypotheses that $\rho_{xy} = \tanh w_0$, see test 13 of Kanji (2006), reference [5].

Correlation: effective degrees of freedom

A crucial aspect of testing for the significance of ρ_{xy} is to use the right value for N, which is the size of your sample if your data points are independent. Otherwise, a lower value of N, known as the *effective degree of freedom*, needs to be considered. We will explore these aspects later and during the practical this afternoon.

Serial correlation: example

Let's consider a realization $X_1, X_2, \ldots, X_N$ of a time series $x(t)$, and calculate the correlations $\rho(X_n, X_{n+1})$:



Since $r_{xy} \neq 0$, X_n and X_{n+1} are correlated, hence they are not independent samples. It is said that x is *serially correlated*, thus the number of degrees of freedom is less than N. The variance and bias of correlation estimates involving x are likely to be affected.

Spearman correlation

The Pearson's correlation coefficient is not the only correlation coefficient. The *rank correlation coefficient* or *Spearman correlation coefficient* is

$$r_s = 1 - \frac{6 \sum_{n=1}^{N} (R_{x;n} - R_{y;n})^2}{N(N^2 - 1)}$$

where $R_{x;n}$ is the rank of the data sample X_n.

Spearman correlation

The Pearson's correlation coefficient is not the only correlation coefficient. The *rank correlation coefficient* or *Spearman correlation coefficient* is

$$r_s = 1 - \frac{6 \sum_{n=1}^{N} (R_{x;n} - R_{y;n})^2}{N(N^2 - 1)}$$

where $R_{x;n}$ is the rank of the data sample X_n. As an example, with $N = 3$

X_n	$R_{x;n}$	Y_n	$R_{y;n}$	$R_{x;n} - R_{y;n}$
4	3	1	1	2
-1	1	12	3	-2
3	2	3	2	0

Spearman correlation

The Pearson's correlation coefficient is not the only correlation coefficient. The *rank correlation coefficient* or *Spearman correlation coefficient* is

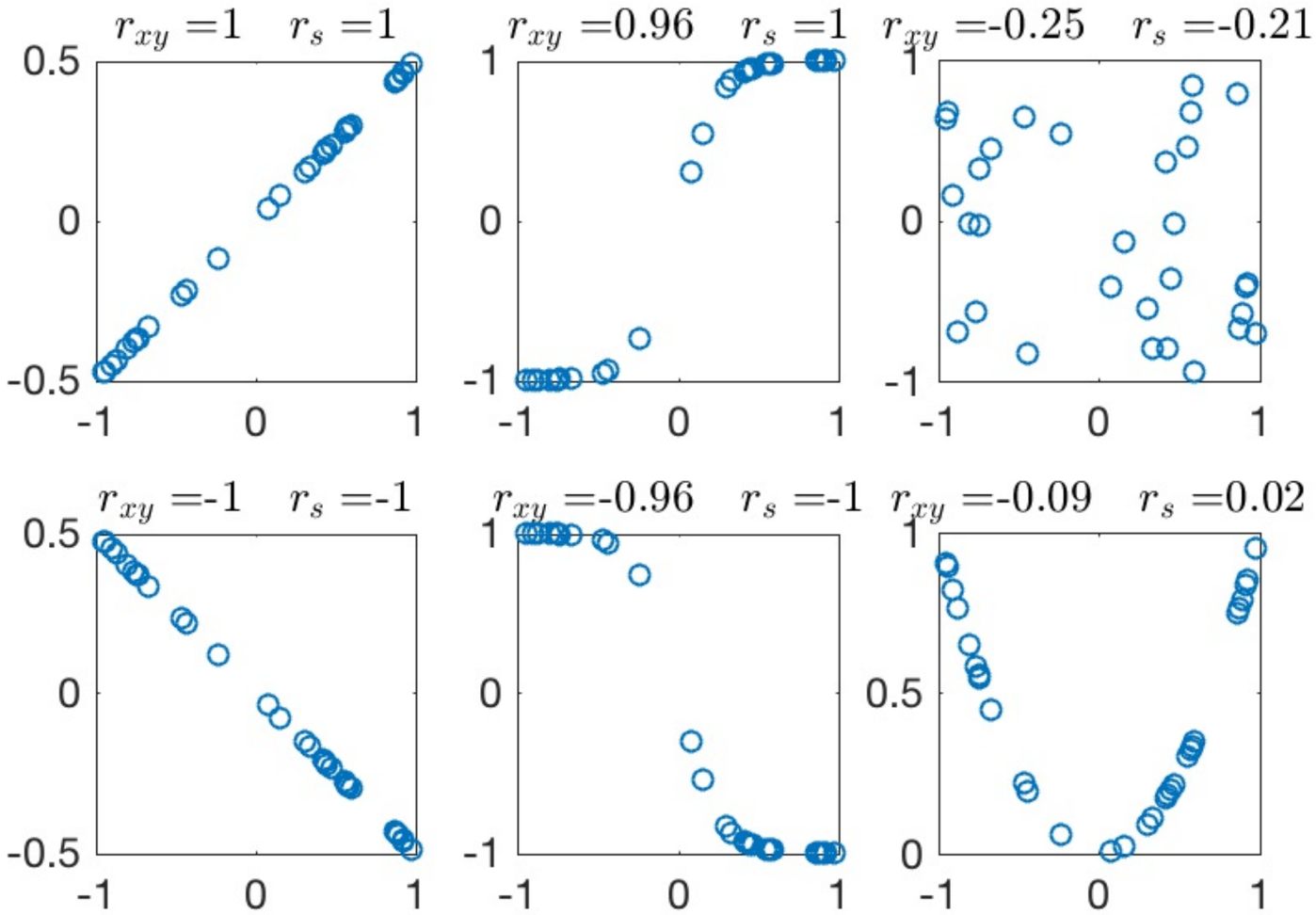
$$r_s = 1 - \frac{6 \sum_{n=1}^{N} (R_{x;n} - R_{y;n})^2}{N(N^2 - 1)}$$

where $R_{x;n}$ is the rank of the data sample X_n. As an example, with $N = 3$

X_n	$R_{x;n}$	Y_n	$R_{y;n}$	$R_{x;n} - R_{y;n}$
4	3	1	1	2
-1	1	12	3	-2
3	2	3	2	0

r_s measures if the relationship between x and y is monotonically increasing ($r_s > 0$) or decreasing ($r_s < 0$). For large N, $r_s \approx \mathcal{N}(0, 1/\sqrt{N-1})$ under $H_0 : r_s = 0$.

Pearson vs Spearman



Covariance: multivariate case

If you are dealing with more than two r.vs., let's say P variables $x_1, x_2, \ldots, x_P$ for which you have $n = 1, 2, \ldots, N$ samples, you need to build the *covariance matrix*

$$\mathbf{C}_{xx} \equiv \begin{bmatrix} C_{x_1 x_1} & C_{x_1 x_2} & \cdots & C_{x_1 x_P} \\ C_{x_2 x_1} & C_{x_2 x_2} & \cdots & C_{x_2 x_P} \\ \vdots & \vdots & \cdots & \vdots \\ C_{x_P x_1} & C_{x_P x_2} & \cdots & C_{x_P x_P} \end{bmatrix}$$

or the *correlation matrix*

$$\rho_{xx} \equiv \begin{bmatrix} \rho_{x_1 x_1} & \rho_{x_1 x_2} & \cdots & \rho_{x_1 x_P} \\ \rho_{x_2 x_1} & \rho_{x_2 x_2} & \cdots & \rho_{x_2 x_P} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{x_P x_1} & \rho_{x_P x_2} & \cdots & \rho_{x_P x_P} \end{bmatrix}$$

Methods to study these matrices will be covered in lecture 5 (Eigen techniques).

Covariance: multivariate case; estimate

If you have P r.vs. $x_1, x_2, \ldots, x_P$ for which you have $n = 1, 2, \ldots, N$ samples, typically arranged in a $N \times P$ data matrix

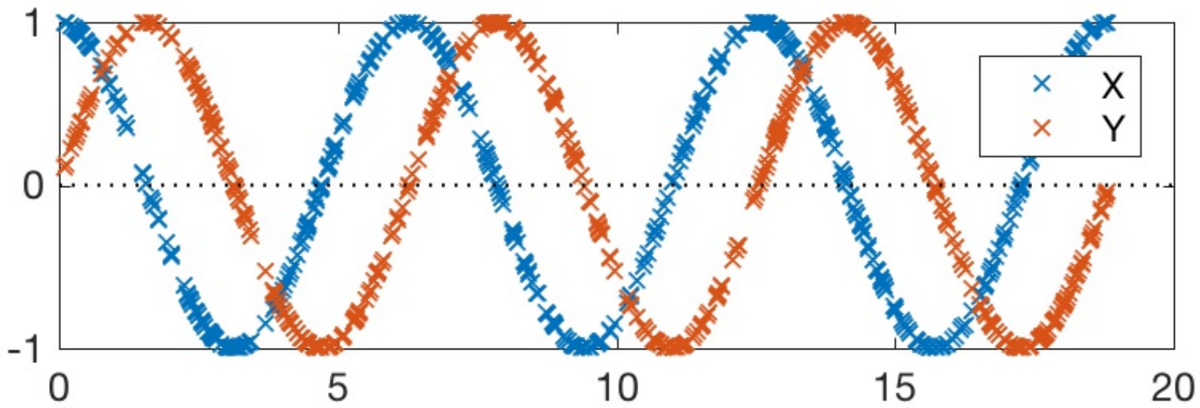
$$\mathbf{X} = [\mathbf{X}_1,\, \mathbf{X}_2,\, \cdots,\, \mathbf{X}_P] = \begin{bmatrix} X_1(1) & X_2(1) & \ldots & X_P(1) \\ X_1(2) & X_2(2) & \cdots & X_P(2) \\ \vdots & \vdots & \ldots & \vdots \\ X_1(N) & X_2(N) & \cdots & X_P(N) \end{bmatrix}$$

The estimate of the covariance matrix is

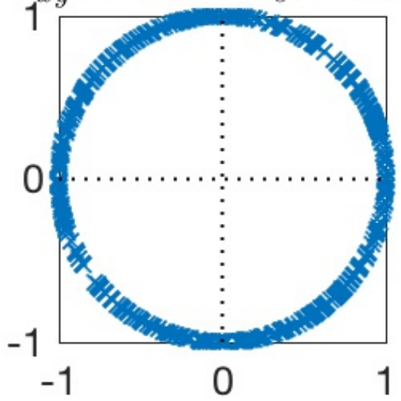
$$\widehat{\mathbf{C}}_{xx} = \tfrac{1}{N-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} \widehat{C}_{x_1x_1} & \widehat{C}_{x_1x_2} & \ldots & \widehat{C}_{x_1x_P} \\ \widehat{C}_{x_2x_1} & \widehat{C}_{x_2x_2} & \ldots & \widehat{C}_{x_2x_P} \\ \vdots & \vdots & \ldots & \vdots \\ \widehat{C}_{x_Px_1} & \widehat{C}_{x_Px_2} & \ldots & \widehat{C}_{x_Px_P} \end{bmatrix}$$

Correlation: beware!

It is not because you find a zero or non significant correlation between two r.vs. that there is no formal or dynamical relationship between the two!

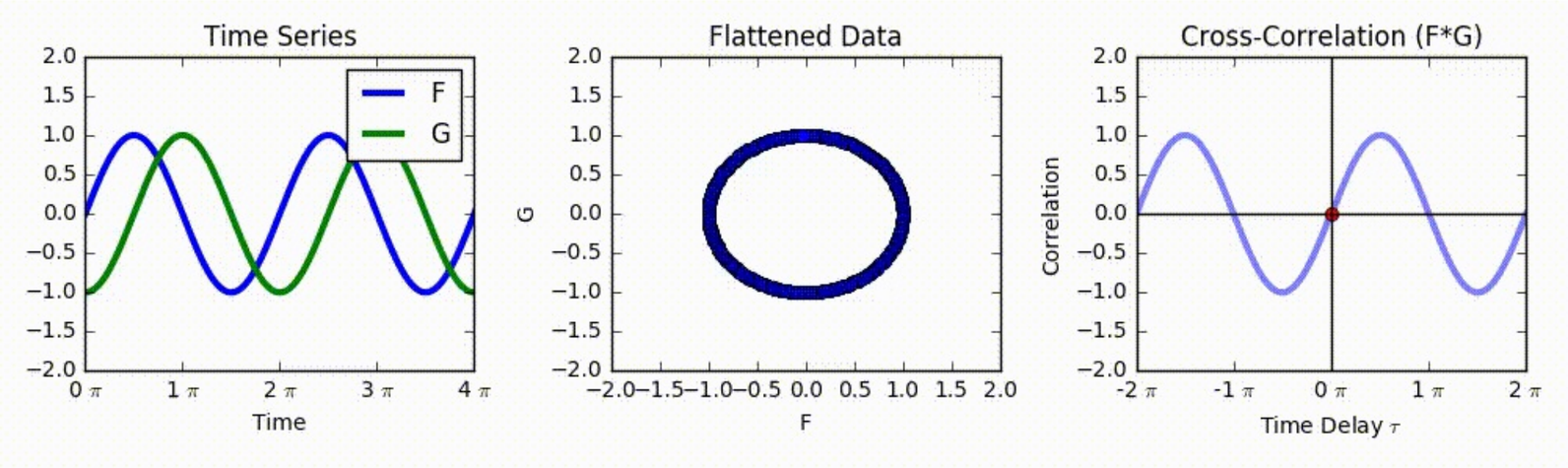


r_{xy} =-0.07 r_s =-0.05



Correlation: beware!

This animation shows the correlation coefficient between two sinusoid signals F and G as a function of phase lag.



Gif by Divergentdata - Own work, CC BY-SA 4.0, Link

2. Lagged Covariance & Correlation

Lagged covariance & correlation functions

We now generalize the concept of covariance by considering two r.vs. for which the samples are ordered, maybe as a function of time t (or of space). In this case, the samples are realizations of time series.

Lagged covariance & correlation functions

We now generalize the concept of covariance by considering two r.vs. for which the samples are ordered, maybe as a function of time t (or of space). In this case, the samples are realizations of time series.

The covariance statistic presented earlier is a special case of the *(cross-)covariance function*, function of *lag* τ

$$C_{xy}(\tau) = E\{[x(t) - \mu_x][y(t+\tau) - \mu_y]\}$$

If $y \equiv x$, $C_{xx}(\tau)$ is called the *auto-covariance function* of x.

Lagged covariance & correlation functions

We now generalize the concept of covariance by considering two r.vs. for which the samples are ordered, maybe as a function of time t (or of space). In this case, the samples are realizations of time series.

The covariance statistic presented earlier is a special case of the *(cross-)covariance function*, function of *lag* τ

$$C_{xy}(\tau) = E\{[x(t) - \mu_x][y(t+\tau) - \mu_y]\}$$

If $y \equiv x$, $C_{xx}(\tau)$ is called the *auto-covariance function* of x.

Note that here we have assumed that the population means μ_x and μ_y are constant with time, in which case it is said that $x(t)$ and $y(t)$ are stationary time series.

Lagged covariance & correlation functions

In statistics and engineering sciences the names can be different. In particular, in Matlab, the following function is called the *cross-correlation function*

$$R_{xy}(\tau) = E\{x(t)y(t+\tau)\}$$

which is similar to the covariance function but without subtracting the means. As such

$$C_{xy}(\tau) = R_{xy}(\tau) - \mu_x\mu_y$$

Lagged covariance & correlation functions

In statistics and engineering sciences the names can be different. In particular, in Matlab, the following function is called the *cross-correlation function*

$$R_{xy}(\tau) = E\{x(t)y(t+\tau)\}$$

which is similar to the covariance function but without subtracting the means. As such

$$C_{xy}(\tau) = R_{xy}(\tau) - \mu_x \mu_y$$

With such naming convention, $R_{xx}(\tau)$ is the *auto-correlation function*.

Lagged covariance & correlation functions

The auto-covariance and auto-correlation functions are even functions of τ (i.e. symmetric around 0):

$$C_{xx}(-\tau) = C_{xx}(\tau)$$
$$R_{xx}(-\tau) = R_{xx}(\tau)$$

The cross-covariance and cross-correlation functions are neither odd nor even, but satisfies

$$C_{xy}(-\tau) = C_{yx}(\tau)$$
$$R_{xy}(-\tau) = R_{yx}(\tau)$$

Lagged covariance & correlation functions

The auto-covariance and auto-correlation functions are even functions of τ (i.e. symmetric around 0):

$$C_{xx}(-\tau) = C_{xx}(\tau)$$
$$R_{xx}(-\tau) = R_{xx}(\tau)$$

The cross-covariance and cross-correlation functions are neither odd nor even, but satisfies

$$C_{xy}(-\tau) = C_{yx}(\tau)$$
$$R_{xy}(-\tau) = R_{yx}(\tau)$$

Beware! Check the conventions of the softwares you use! In Matlab, the cross-correlation function is defined as
$E[x(t+\tau)y(t)] = R_{xy}(-\tau) = R_{yx}(\tau)$.

Lagged correlation coefficient

The *lagged correlation coefficient* is

$$\rho_{xy}(\tau) = \frac{C_{xy}(\tau)}{\sqrt{C_{xx}(0)C_{yy}(0)}} = \frac{C_{xy}(\tau)}{\sigma_x \sigma_y}$$

Lagged correlation coefficient

The *lagged correlation coefficient* is

$$\rho_{xy}(\tau) = \frac{C_{xy}(\tau)}{\sqrt{C_{xx}(0)C_{yy}(0)}} = \frac{C_{xy}(\tau)}{\sigma_x \sigma_y}$$

It can be challenging but let's try not to confuse correlation function and lagged correlation.

Lagged correlation coefficient

The *lagged correlation coefficient* is

$$\rho_{xy}(\tau) = \frac{C_{xy}(\tau)}{\sqrt{C_{xx}(0)C_{yy}(0)}} = \frac{C_{xy}(\tau)}{\sigma_x \sigma_y}$$

It can be challenging but let's try not to confuse correlation function and lagged correlation.

You may want to call $\rho_{xx}(\tau) = \frac{C_{xx}(\tau)}{C_{xx}(0)}$ the *lagged auto-correlation coefficient*, but it is usually called the autocorrelation function (which should be used for $R_{xx}(\tau)$).

Lagged correlation coefficient

The *lagged correlation coefficient* is

$$\rho_{xy}(\tau) = \frac{C_{xy}(\tau)}{\sqrt{C_{xx}(0)C_{yy}(0)}} = \frac{C_{xy}(\tau)}{\sigma_x \sigma_y}$$

It can be challenging but let's try not to confuse correlation function and lagged correlation.

You may want to call $\rho_{xx}(\tau) = \frac{C_{xx}(\tau)}{C_{xx}(0)}$ the *lagged auto-correlation coefficient*, but it is usually called the autocorrelation function (which should be used for $R_{xx}(\tau)$).

Distributions under null hypotheses for $\rho_{xy}(\tau)$ or $\rho_{xx}(\tau)$ are difficult to come by and require to make several assumptions (see practical this afternoon).

Lagged covariance function estimates

Consider two r.vs. x and y for which we have N samples separated by constant intervals Δt. One estimate of the cross-covariance function at lags $\tau_k = k\Delta t$ is

$$\widehat{C}_{xy}^{(1)}(k\Delta t) = \frac{1}{N-|k|}\sum_{n=1}^{N-|k|}(X_n - \overline{X})(Y_{n+k} - \overline{Y})$$

where $k = 0, \pm 1, \pm 2, \ldots, \pm(N-1)$ are the possible lags. It can be shown that this estimator is unbiased (if the population means are known rather than estimated).

Lagged covariance function estimates

Consider two r.vs. x and y for which we have N samples separated by constant intervals Δt. One estimate of the cross-covariance function at lags $\tau_k = k\Delta t$ is

$$\widehat{C}_{xy}^{(1)}(k\Delta t) = \frac{1}{N-|k|}\sum_{n=1}^{N-|k|}(X_n - \overline{X})(Y_{n+k} - \overline{Y})$$

where $k = 0, \pm 1, \pm 2, \ldots, \pm(N-1)$ are the possible lags. It can be shown that this estimator is unbiased (if the population means are known rather than estimated). There exists another estimator of $C_{xy}(\tau)$ which is

$$\widehat{C}_{xy}^{(2)}(k\Delta t) = \frac{1}{N}\sum_{n=1}^{N-|k|}(X_n - \overline{X})(Y_{n+k} - \overline{Y})$$

which is typically called the "biased" estimator but has a smaller random error compared to $\widehat{C}_{xy}^{(1)}$. See practical this afternoon.

Lagged correlation estimate

An estimate of the lagged correlation coefficient is

$$\hat{\rho}_{xy}(k\Delta t) = \frac{\widehat{C}_{xy}^{(2)}(k\Delta t)}{\left[\widehat{C}_{xx}^{(2)}\,\widehat{C}_{yy}^{(2)}\right]^{1/2}} = \frac{\sum_{n=1}^{N-|k|}(X_n - \overline{X})(Y_{n+k} - \overline{Y})}{\left[\sum_{n=1}^{N}(X_n - \overline{X})^2 \sum_{n=1}^{N}(Y_n - \overline{Y})^2\right]^{1/2}}$$

where the use of $\widehat{C}_{xy}^{(2)}$ cancels out the factor N in the formulas for the variances. This estimator is accessed in Matlab using the scale option `'coeff'` as in

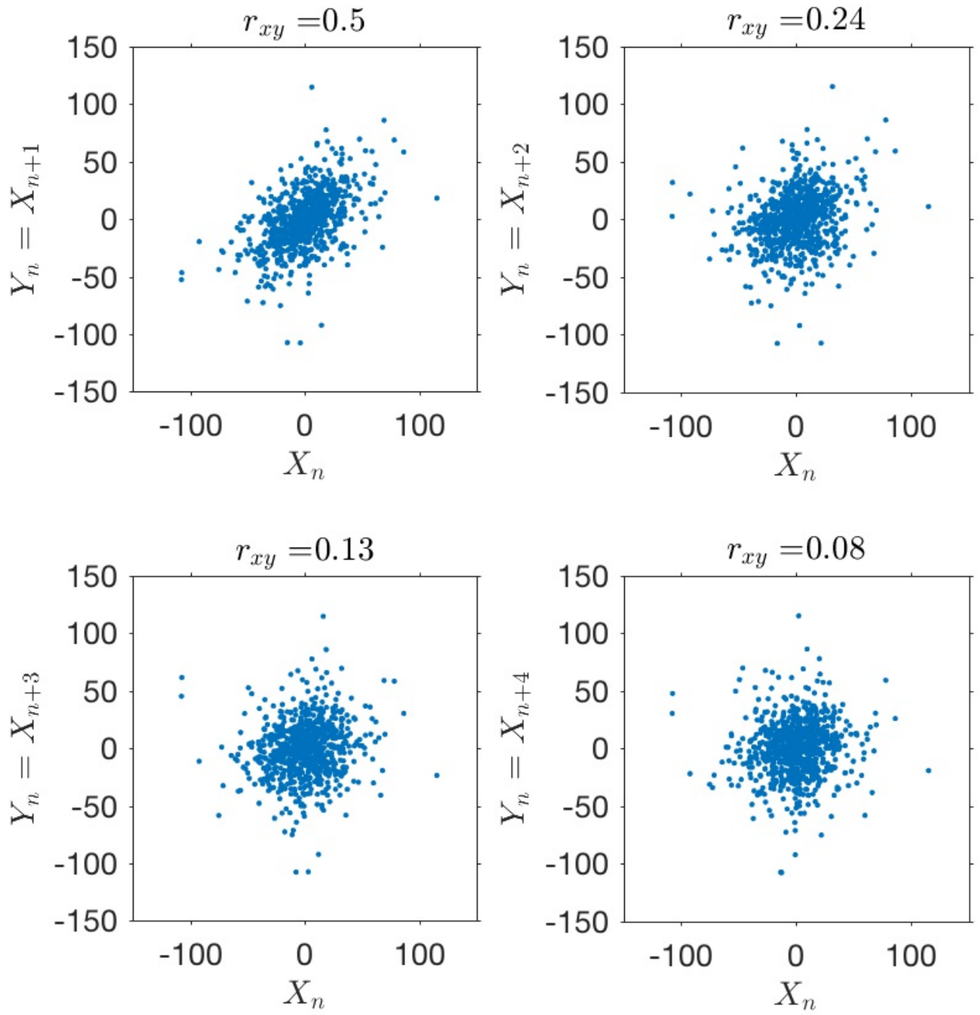
```
[rhox,lags] = xcov(Y,X,'coeff');
```

or

```
[rhox,lags] = xcorr(Y-mean(Y),X-mean(X),'coeff');
```

Auto-correlation as a measure of memory

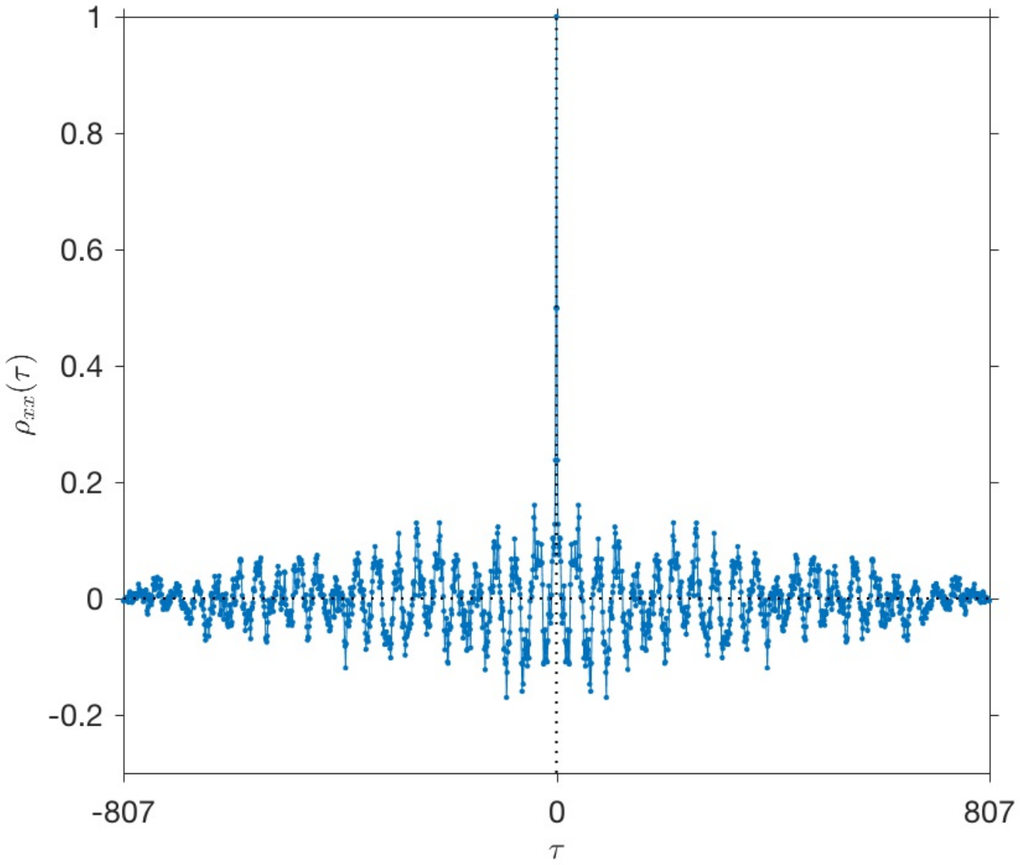
Lag m scatter plots and cor. coeff. $\rho(X_n, X_{n+m})$ for Agulhas



transport

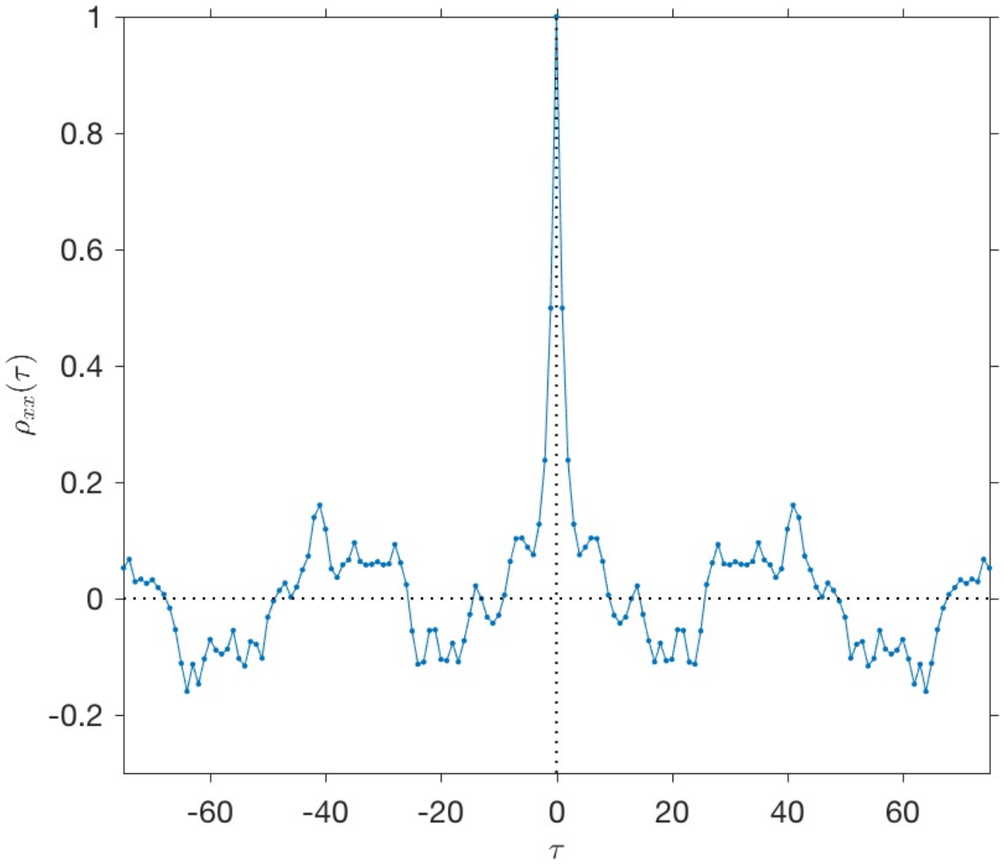
Auto-correlation as a measure of memory

Lagged autocorrelation coefficient for Agulhas transport time series. With $N = 808$, possible lags are $\tau = -807, \ldots, 807$



Auto-correlation as a measure of memory

Lagged autocorrelation coefficient for Agulhas transport time series. With $N = 808$, possible lags are $\tau = -807, \ldots, 807$



Lagged correlation for time delay

The lagged correlation or covariance function can be used to determine a time delay between two signals. Let's assume that a transmitted signal (a time series) $x(t)$ is received as

$$y(t) = \alpha x(t - \tau_0) + n(t)$$

where α is an *attenuation factor*, $\tau_0 = d/c$ is a constant time delay equal to let's say a distance d divided by a propagation velocity c, and $n(t)$ is an added noise uncorrelated with $x(t)$.

Lagged correlation for time delay

The lagged correlation or covariance function can be used to determine a time delay between two signals. Let's assume that a transmitted signal (a time series) $x(t)$ is received as

$$y(t) = \alpha x(t - \tau_0) + n(t)$$

where α is an *attenuation factor*, $\tau_0 = d/c$ is a constant time delay equal to let's say a distance d divided by a propagation velocity c, and $n(t)$ is an added noise uncorrelated with $x(t)$. It is easily shown that

$$R_{xy}(\tau) = \alpha R_{xx}(\tau - \tau_0)$$

Lagged correlation for time delay

The lagged correlation or covariance function can be used to determine a time delay between two signals. Let's assume that a transmitted signal (a time series) $x(t)$ is received as

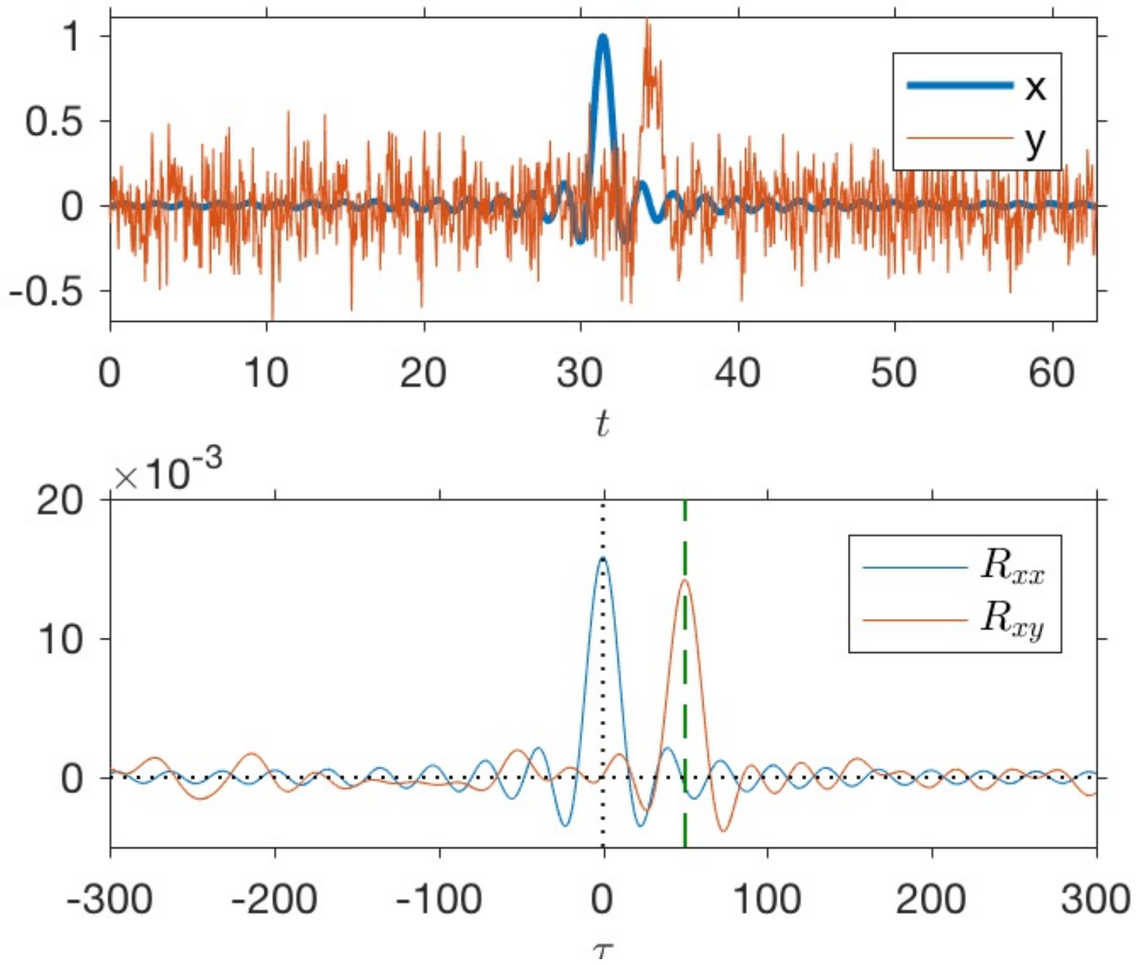
$$y(t) = \alpha x(t - \tau_0) + n(t)$$

where α is an *attenuation factor*, $\tau_0 = d/c$ is a constant time delay equal to let's say a distance d divided by a propagation velocity c, and $n(t)$ is an added noise uncorrelated with $x(t)$. It is easily shown that

$$R_{xy}(\tau) = \alpha R_{xx}(\tau - \tau_0)$$

Since the maximum of $R_{xx}(\tau)$ is for $\tau = 0$, the peak value of $R_{xy}(\tau)$ occurs when $\tau = \tau_0$ which allows us to determine the constant time delay.

Lagged correlation for time delay



Lagged correlation for time delay

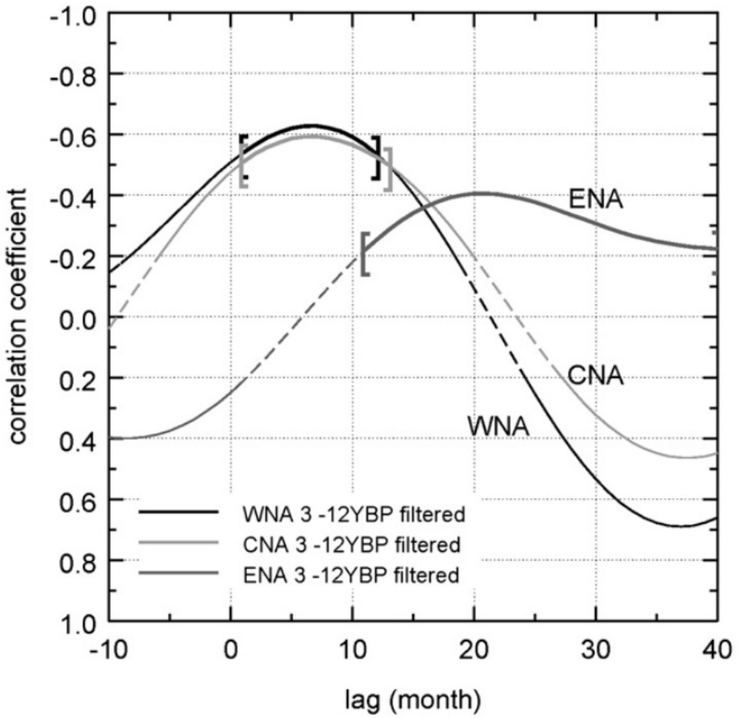


FIG. 8. Correlation coefficient between the 3–12-yr bandpass-filtered (3–12 YBP) FC transport time series and the 3–12 YBP filtered WSC time series corresponding to each forcing region, using a range of lag times in months. Correlation coefficients that are statistically significant at the 67% confidence level are connected by solid lines; dashed lines connect correlations that lack statistical significance. In addition, lags at which the low-frequency WSC–FC correlation is statistically indistinguishable from that of the peak with 67% confidence are bounded by square brackets, providing an estimate of the statistical significance of the peak lags.

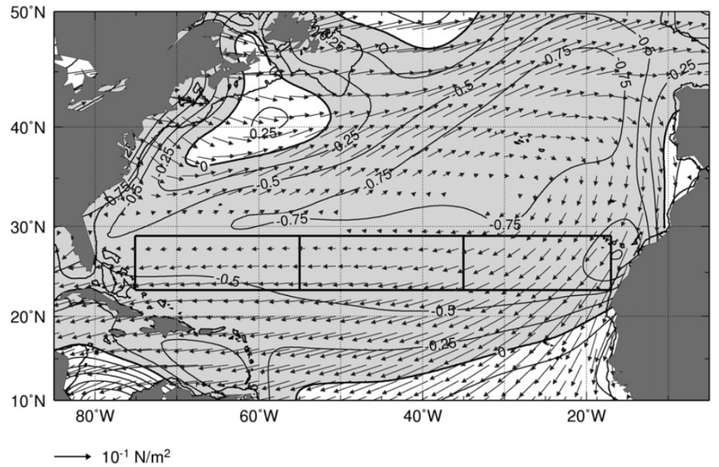


FIG. 2. Climatological mean surface wind stress from the NCEP–NCAR reanalysis project between 1982 and 2007. Arrows show the wind stress field, and the contour map is the vertical component of WSC. The WSC contours are in units of 10^{-7} Pa m^{-1}. White (gray) background is for positive (negative) mean WSC. Three regions are outlined across a latitude band centered at 27°N, corresponding approximately to the Straits of Florida: WNA, CNA, and ENA. WSC forcing over these regions is hypothesized to influence the FC transport on interannual time scales, via baroclinic adjustment. The solid line shown following continental coastlines corresponds to the 200-m isobath.

Example from DiNezio et al. 2009 investigating the relationship between Gulf Stream transport through the Florida Strait and wind stress curl.

Auto-correlation and effective degrees of freedom

The concept of *effective degrees of freedom* and *(de)correlation time scale* are intimately linked. If N is the number of evenly distributed samples at interval Δt, then the number of *effective degrees of freedom* is

$$N_{eff} = \frac{N\Delta t}{T_0} = \frac{T}{T_0}$$

where T is the length of your time series and T_0 is the *decorrelation time scale*, also referred to as an *integral time scale*. N_{eff} gives you the number of effectively independent samples in your data.

Auto-correlation and effective degrees of freedom

The concept of *effective degrees of freedom* and *(de)correlation time scale* are intimately linked. If N is the number of evenly distributed samples at interval Δt, then the number of *effective degrees of freedom* is

$$N_{eff} = \frac{N \Delta t}{T_0} = \frac{T}{T_0}$$

where T is the length of your time series and T_0 is the *decorrelation time scale*, also referred to as an *integral time scale*. N_{eff} gives you the number of effectively independent samples in your data.

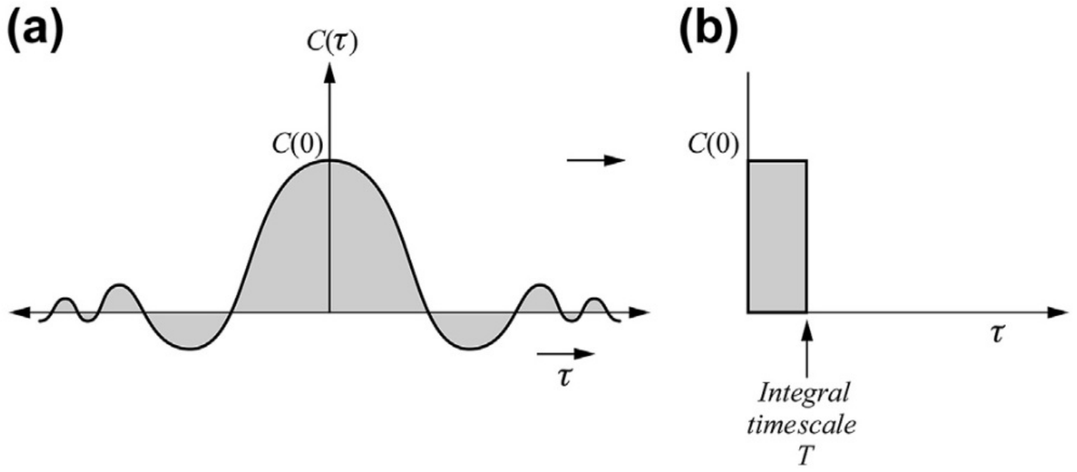
If your samples are independent, the decorrelation time scale is the time step of your time series Δt and $N_{eff} = N$. There is not only one way of computing N_{eff} or estimating T_0. In fact, it depends on the statistics for which these will be used.

Decorrelation time scale

One common definition of *the decorrelation or integral time scale* for the r.v. x is (e.g. Thompson and Emery (2014), section 3.15.2)

$$T_0 = \frac{1}{C_{xx}(0)} \int_{-\infty}^{+\infty} C_{xx}(\tau)\, d\tau = \frac{2}{C_{xx}(0)} \int_{0}^{+\infty} C_{xx}(\tau)\, d\tau$$

A graphical interpretation of T_0 is given in Figure 3.13 of Emery and Thompson (2014) : $T_0 \times C_{xx}(0) = \int_{-\infty}^{+\infty} C_{xx}(\tau)\, d\tau$



Decorrelation time scale

An estimate of T_0 is obtained by applying the trapezoidal integration formula for the integral, i.e.

$$\widehat{T_0} = \frac{2}{s_x^2} \sum_{n=0}^{M-1} \frac{\widehat{C}_{xx}[(n+1)\Delta t] + \widehat{C}_{xx}[n\Delta t]}{2} \Delta t$$

where $s_x^2 = \widehat{C}_{xx}(0)$ is the sample variance estimate of x and M is the index where the summation is stopped.

Decorrelation time scale

An estimate of T_0 is obtained by applying the trapezoidal integration formula for the integral, i.e.

$$\widehat{T_0} = \frac{2}{s_x^2} \sum_{n=0}^{M-1} \frac{\widehat{C}_{xx}[(n+1)\Delta t] + \widehat{C}_{xx}[n\Delta t]}{2} \Delta t$$

where $s_x^2 = \widehat{C}_{xx}(0)$ is the sample variance estimate of x and M is the index where the summation is stopped.

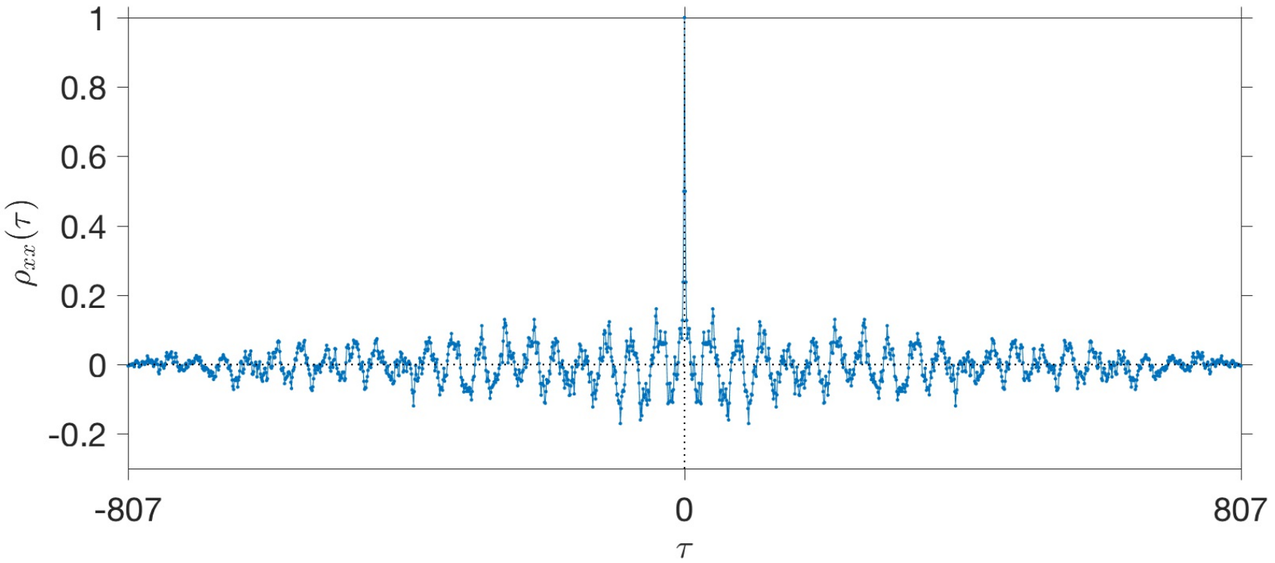
Note that Thompson and Emery (2014), section 5.3.5, give an alternate formula

$$\widehat{T_0} = \frac{1}{s_x^2} \sum_{n=0}^{M-1} \frac{\widehat{C}_{xx}[(n+1)\Delta t] + \widehat{C}_{xx}[n\Delta t]}{2} \Delta t$$

where they omitted the factor 2 and essentially integrated only one side of the auto-covariance function. This is also the formula typically used in Lagrangian studies. (And this is what I use).

Decorrelation time scale

There are several sources of error entering formulas for T_0: estimate of the variance, estimate of the autocovariance, and truncation of the trapezoidal integration. The truncation of the integration may be where the estimated autocovariance function reaches a constant value but this may not happen in practice. Another possibility is to integrate up to the first zero crossing or up to where the autocorrelation is insignificant. See practical session this afternoon.



Auto-correlation and effective degrees of freedom

Once you have obtained an integral time scale T_0 and calculated the effective degrees of freedom N_{eff}, you may use this parameter to calculate CIs for sample means.

Auto-correlation and effective degrees of freedom

Once you have obtained an integral time scale T_0 and calculated the effective degrees of freedom N_{eff}, you may use this parameter to calculate CIs for sample means.

The formulas given previously do not seem to be appropriate when N_{eff} is needed to assess estimators of auto-covariance/correlation or cross-covariance/correlation. In these cases, Von Storch and Zwiers (1999), reference [6], gives the respective two formulas

$$\frac{T_0}{\Delta t} = 1 + 2 \sum_{k=1}^{+\infty} \rho_{xx}^2 (k \Delta t)$$

$$\frac{T_0}{\Delta t} = 1 + 2 \sum_{k=1}^{+\infty} \rho_{xx} (k \Delta t) \rho_{yy} (k \Delta t)$$

assuming you have a constant time step Δt for your time series.

3. A quick look at "A leisurely look at the bootstrap, the jackknife, and cross-validation" by Efron and Gong (1983)

The bootstrap and the jackknife

So far, we have used theoretical or assumed distributions of our data in order to derive standard errors and CIs of our parameter estimates. However, in many cases, like for the correlation coefficient, it is impossible to express the variance in closed form. In addition, we are often stuck with one sample $(X_1, X_2, \ldots, X_n)$ from one experiment in order to investigate the x population. On top of this, our samples may not be independent and we run into the issue of estimating the effective degrees of freedom.

The bootstrap and the jackknife

So far, we have used theoretical or assumed distributions of our data in order to derive standard errors and CIs of our parameter estimates. However, in many cases, like for the correlation coefficient, it is impossible to express the variance in closed form. In addition, we are often stuck with one sample $(X_1, X_2, \ldots, X_n)$ from one experiment in order to investigate the x population. On top of this, our samples may not be independent and we run into the issue of estimating the effective degrees of freedom.

The idea of the *bootstrap* and *jackknife*, combined with Monte Carlo methods, is to resample your original sample with *replacements for the bootstrap*, and with *deletions for the jackknife*. These methods are relatively easy to implement with our fast and modern computers.

The bootstrap : principle (1)

Let's assume you are investigating a univariate or multivariate r.v. x for which you want to estimate a statistic ϕ with estimator $\widehat{\phi}$, using a sample $(X_1, X_2, \ldots, X_n)$.

The bootstrap : principle (1)

Let's assume you are investigating a univariate or multivariate r.v. x for which you want to estimate a statistic ϕ with estimator $\widehat{\phi}$, using a sample $(X_1, X_2, \ldots, X_n)$.

First, you draw randomly a *bootstrap sample* $(X_1^*, X_2^*, \ldots, X_N^*)$ of the same size as your original sample. As an example, if your original sample if of size $N = 3$, i.e. (X_1, X_2, X_3), a bootstrap sample with replacement may be $(X_1^*, X_2^*, X_3^*) = (X_1, X_2, X_1)$.

The bootstrap : principle (1)

Let's assume you are investigating a univariate or multivariate r.v. x for which you want to estimate a statistic ϕ with estimator $\widehat{\phi}$, using a sample $(X_1, X_2, \ldots, X_n)$.

First, you draw randomly a *bootstrap sample* $(X_1^*, X_2^*, \ldots, X_N^*)$ of the same size as your original sample. As an example, if your original sample if of size $N = 3$, i.e. (X_1, X_2, X_3), a bootstrap sample with replacement may be $(X_1^*, X_2^*, X_3^*) = (X_1, X_2, X_1)$.

From the bootstrap sample, you calculate a bootstrap replication estimate of your statistic $\widehat{\phi}^*$.

The bootstrap : principle (1)

Let's assume you are investigating a univariate or multivariate r.v. x for which you want to estimate a statistic ϕ with estimator $\widehat{\phi}$, using a sample $(X_1, X_2, \ldots, X_n)$.

First, you draw randomly a *bootstrap sample* $(X_1^*, X_2^*, \ldots, X_N^*)$ of the same size as your original sample. As an example, if your original sample if of size $N = 3$, i.e. (X_1, X_2, X_3), a bootstrap sample with replacement may be $(X_1^*, X_2^*, X_3^*) = (X_1, X_2, X_1)$.

From the bootstrap sample, you calculate a bootstrap replication estimate of your statistic $\widehat{\phi}^*$. You repeat this operation B times to obtain B bootstrap replications $\widehat{\phi}^{*(1)}, \widehat{\phi}^{*(2)}, \ldots, \widehat{\phi}^{*(B)}$.

The bootstrap : principle (1)

Let's assume you are investigating a univariate or multivariate r.v. x for which you want to estimate a statistic ϕ with estimator $\widehat{\phi}$, using a sample $(X_1, X_2, \ldots, X_n)$.

First, you draw randomly a *bootstrap sample* $(X_1^*, X_2^*, \ldots, X_N^*)$ of the same size as your original sample. As an example, if your original sample if of size $N = 3$, i.e. (X_1, X_2, X_3), a bootstrap sample with replacement may be $(X_1^*, X_2^*, X_3^*) = (X_1, X_2, X_1)$.

From the bootstrap sample, you calculate a bootstrap replication estimate of your statistic $\widehat{\phi}^*$. You repeat this operation B times to obtain B bootstrap replications $\widehat{\phi}^{*(1)}, \widehat{\phi}^{*(2)}, \ldots, \widehat{\phi}^{*(B)}$.

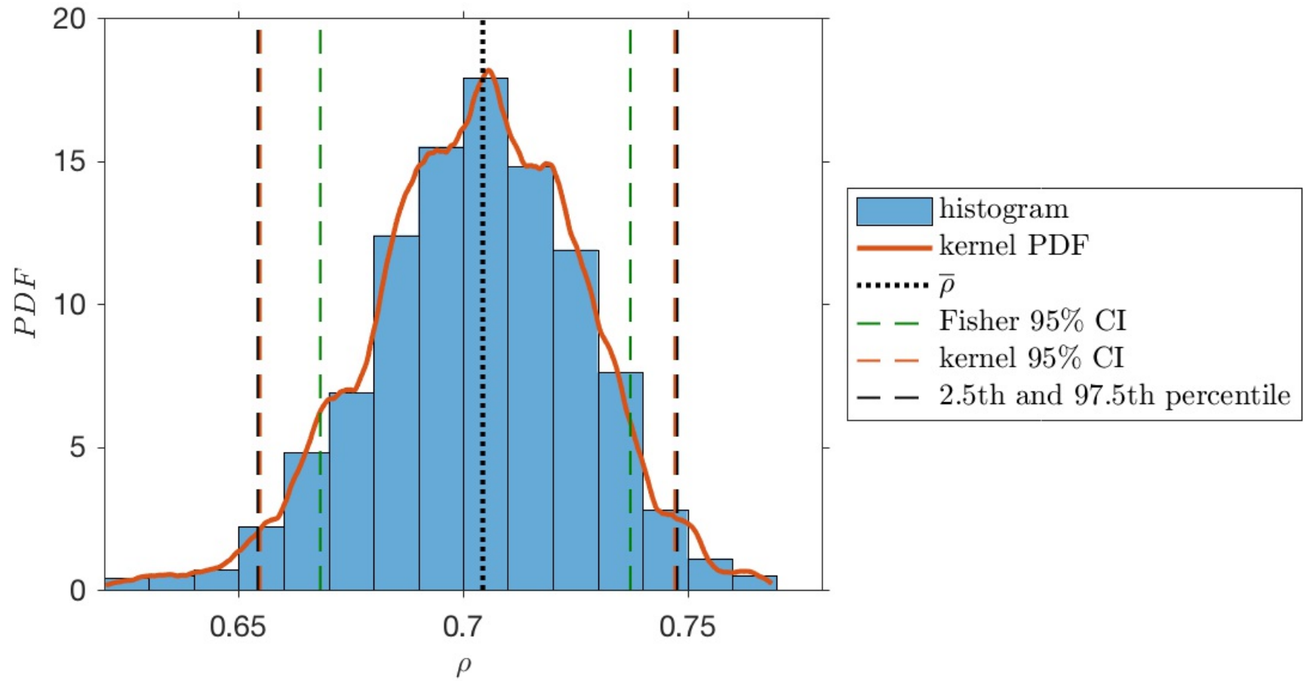
You finally calculate (estimate) the variance of your replications as

$$\mathrm{Var}[\widehat{\phi}^*] = \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\phi}^{*(b)} - \widehat{\phi}^{*(.)})^2 \quad \text{where} \quad \widehat{\phi}^{*(.)} = \frac{\sum_b \widehat{\phi}^{*(b)}}{B}$$

The bootstrap : principle (2)

You can now use $\mathrm{Var}[\widehat{\phi}^*]$ to derive a standard error of your estimate as $\sqrt{\mathrm{Var}[\widehat{\phi}^*]}$ or use the estimated distribution of your bootstrap replications $\widehat{\phi}^{*(1)}, \widehat{\phi}^{*(2)}, \ldots, \widehat{\phi}^{*(B)}$ to calculate CIs.

Example : Bootstrap correlations between between Agulhas jet and boundary transports with $B = 1000$.



The jackknife : principle (1)

Let's assume again that you are investigating a univariate or multivariate r.v. x for which you want to estimate a statistic ϕ with estimator $\widehat{\phi}$, using a sample $(X_1, X_2, \ldots, X_n)$.

The jackknife : principle (1)

Let's assume again that you are investigating a univariate or multivariate r.v. x for which you want to estimate a statistic ϕ with estimator $\widehat{\phi}$, using a sample $(X_1, X_2, \ldots, X_n)$. A *jackknife sample* is obtained by deleting J data points from your original sample of N points. The number of such sample that can be obtained is the number of permutations of N objects taken J at a time which is $N!/(N-J)!$. If you have N data points there are N permutations for the "delete-1" jackknife. The number of of possible jackknife samples can become very large so that a subset need to be choosen.

The jackknife : principle (1)

Let's assume again that you are investigating a univariate or multivariate r.v. x for which you want to estimate a statistic ϕ with estimator $\widehat{\phi}$, using a sample $(X_1, X_2, \ldots, X_n)$. A *jackknife sample* is obtained by deleting J data points from your original sample of N points. The number of such sample that can be obtained is the number of permutations of N objects taken J at a time which is $N!/(N-J)!$. If you have N data points there are N permutations for the "delete-1" jackknife. The number of of possible jackknife samples can become very large so that a subset need to be choosen.

The formula for the "delete-1" jackknife variance of your statistic estimate is

$$\mathrm{Var}[\widehat{\phi}^J] = \frac{N-1}{N}\sum_{j=1}^{N}(\widehat{\phi}^{*(j)} - \widehat{\phi}^{*(.)})^2 \quad \text{where} \quad \widehat{\phi}^{*(.)} = \frac{\sum_j \widehat{\phi}^{*(j)}}{N}$$

The jackknife : principle (2)

The jackknife method can be also used to estimate the bias of your estimator as

$$\hat{b}_J[\widehat{\phi}] = (N-1)(\widehat{\phi}^{*(.)} - \widehat{\phi})$$

so that the *jackknife estimate* of your statistic is

$$\widehat{\phi}_J = N\widehat{\phi} - (N-1)\widehat{\phi}^{*(.)}$$

The jackknife : principle (2)

The jackknife method can be also used to estimate the bias of your estimator as

$$\hat{b}_J[\widehat{\phi}] = (N-1)(\widehat{\phi}^{*(.)} - \widehat{\phi})$$

so that the *jackknife estimate* of your statistic is

$$\widehat{\phi}_J = N\widehat{\phi} - (N-1)\widehat{\phi}^{*(.)}$$

There are some important details to the bootstrap and the jackknife methods. Please see Thompson and Emery , Tichelaar and Ruff (1989) and Efron and Gong (1983).

4. Covariance and correlation of bivariate variables

Bivariate random variable

What if your process of interest, or your data, is a "vector variable"? It could be an ocean current, an atmospheric wind, the position of a drifter etc. Such variables are called *bivariate* variables in the statistics litterature, and are also treated as r.vs.

Bivariate random variable

What if your process of interest, or your data, is a "vector variable"? It could be an ocean current, an atmospheric wind, the position of a drifter etc. Such variables are called *bivariate* variables in the statistics litterature, and are also treated as r.vs. Here we will use the formalism of time series, that is we assume that we have samples indexed along an axis t that represents time.

Bivariate random variable

What if your process of interest, or your data, is a "vector variable"? It could be an ocean current, an atmospheric wind, the position of a drifter etc. Such variables are called *bivariate* variables in the statistics litterature, and are also treated as r.vs. Here we will use the formalism of time series, that is we assume that we have samples indexed along an axis t that represents time.

Let's consider a first bivariate variable $z(t)$ with Cartesian components $x(t)$ and $y(t)$ (east-west and north-south).

Bivariate random variable

What if your process of interest, or your data, is a "vector variable"? It could be an ocean current, an atmospheric wind, the position of a drifter etc. Such variables are called *bivariate* variables in the statistics litterature, and are also treated as r.vs. Here we will use the formalism of time series, that is we assume that we have samples indexed along an axis t that represents time.

Let's consider a first bivariate variable $z(t)$ with Cartesian components $x(t)$ and $y(t)$ (east-west and north-south).

We can arrange the components into a 1×2 vector function of time

$$\mathbf{z}(t) = \begin{bmatrix} x(t) & y(t) \end{bmatrix} \quad \text{or} \quad \mathbf{z}_n = \begin{bmatrix} x_n & y_n \end{bmatrix}$$

or alternatively use a complex-valued notation

$$z(t) = x(t) + iy(t) = |z(t)| e^{i \arg(z)},$$

where $i = \sqrt{-1}$ and $\arg(z)$ is the *complex argument* (or polar angle) of z in the interval $[-\pi, +\pi]$.

The mean of bivariate Data

The sample mean of the vector time series $\mathbf{z}(t)$ is also a vector,

$$\begin{aligned}\widehat{\mu}_{\mathbf{z}} = \overline{\mathbf{z}} \equiv \frac{1}{N}\sum_{n=1}^{N}\mathbf{z}_n &= \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} x_n & \frac{1}{N}\sum_{n=1}^{N} y_n \end{bmatrix} \\ &= \begin{bmatrix} \overline{x} & \overline{y} \end{bmatrix} = \begin{bmatrix} \widehat{\mu}_x & \widehat{\mu}_y \end{bmatrix}\end{aligned}$$

that consists of the *sample means* of the x and y components of z.

The mean of bivariate Data

The sample mean of the vector time series $\mathbf{z}(t)$ is also a vector,

$$\begin{aligned}\widehat{\mu}_{\mathbf{z}} = \overline{\mathbf{z}} \equiv \frac{1}{N}\sum_{n=1}^{N}\mathbf{z}_n &= \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} x_n & \frac{1}{N}\sum_{n=1}^{N} y_n \end{bmatrix} \\ &= \begin{bmatrix} \overline{x} & \overline{y} \end{bmatrix} = \begin{bmatrix} \widehat{\mu}_x & \widehat{\mu}_y \end{bmatrix}\end{aligned}$$

that consists of the *sample means* of the x and y components of z.

Using the complex notation, the sample mean is

$$\overline{z} = \overline{x} + i\,\overline{y}$$

which is complex number.

The variance of bivariate variable

The *variance* of the vector-valued times series $\mathbf{z}_n$ is not a scalar or a vector, it is a 2×2 *matrix*

$$\mathbf{\Sigma}_z \equiv \frac{1}{N-1} \sum_{n=1}^{N} \left(\mathbf{z}_n - \overline{\mathbf{z}}\right)^T \left(\mathbf{z}_n - \overline{\mathbf{z}}\right)$$

where "T" is the *matrix transpose*, $\mathbf{z}_n = \begin{bmatrix} x_n & y_n \end{bmatrix}$, $\mathbf{z}_n^T = \begin{bmatrix} x_n \\ y_n \end{bmatrix}$.

Carrying out the matrix multiplication leads to

$$\mathbf{\Sigma}_z = \frac{1}{N-1} \sum_{n=1}^{N} \begin{bmatrix} \left(x_n - \overline{x}\right)^2 & \left(x_n - \overline{x}\right)\left(y_n - \overline{y}\right) \\ \left(y_n - \overline{y}\right)\left(x_n - \overline{x}\right) & \left(y_n - \overline{y}\right)^2 \end{bmatrix} = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{yx} & s_y^2 \end{bmatrix}$$

The diagonal elements of $\mathbf{\Sigma}_z$ are the sample variances, while the off-diagonal gives the sample covariance between x_n and y_n. Note that the two off-diagonal elements are identical, $s_{yx} = s_{xy}$.

The variance of bivariate variable

The variance of a complex r.v., $z = x + iy$, needs a definition:

$$\mathrm{Var}[z] = E[(z - \mu_z)^*(z - \mu_z)]$$

where $(.)^*$ means the *complex conjuguate*, i.e. $z^* = x - iy$.

The variance of bivariate variable

The variance of a complex r.v., $z = x + iy$, needs a definition:

$$\mathrm{Var}[z] = E[(z - \mu_z)^*(z - \mu_z)]$$

where $(.)^*$ means the *complex conjuguate*, i.e. $z^* = x - iy$.

Substituting and expanding the previous expression gives

$$\begin{aligned}\mathrm{Var}[z] &= E[(x - \mu_x)^2] + E[(y - \mu_y)^2] \\ &= \mathrm{Var}[x] + \mathrm{Var}[y]\end{aligned}$$

The variance of a "physical vector" written as a complex r.v. is a single **real** number which is the sum of the variance of its components. It is different from the 2×2 matrix of variances and covariances of its components seen in the previous slide.

The variance of bivariate variable

The variance of a complex r.v., $z = x + iy$, needs a definition:

$$\mathrm{Var}[z] = E[(z - \mu_z)^*(z - \mu_z)]$$

where $(.)^*$ means the *complex conjuguate*, i.e. $z^* = x - iy$.

Substituting and expanding the previous expression gives

$$\begin{aligned}\mathrm{Var}[z] &= E[(x - \mu_x)^2] + E[(y - \mu_y)^2] \\ &= \mathrm{Var}[x] + \mathrm{Var}[y]\end{aligned}$$

The variance of a "physical vector" written as a complex r.v. is a single **real** number which is the sum of the variance of its components. It is different from the 2×2 matrix of variances and covariances of its components seen in the previous slide. If z represents an ocean current, $\mathrm{Var}[z]$ is proportional to the average eddy kinetic energy (density) $KE = (< x'^2 > + < y'^2 >)/2$

The covariance of bivariate variables

In addition to $z(t)$, let's consider a second bivariate variable $w(t)$ with Cartesian components $g(t)$ and $h(t)$.

The covariance of bivariate variables

In addition to $z(t)$, let's consider a second bivariate variable $w(t)$ with Cartesian components $g(t)$ and $h(t)$. The cross covariance function at zero lag, or simply the covariance between the components of z and w can be formed as

$$
\begin{aligned}
E\left\{\mathbf{z}^{\mathrm{T}}\mathbf{w}\right\} &= E\left\{\begin{bmatrix} (x-\mu_x)(g-\mu_g) & (x-\mu_x)(h-\mu_h) \\ (y-\mu_y)(g-\mu_g) & (y-\mu_y)(h-\mu_h) \end{bmatrix}\right\} \\
&= \begin{bmatrix} C_{xg} & C_{xh} \\ C_{yg} & C_{yh} \end{bmatrix}
\end{aligned}
$$

The covariance of bivariate variables

In addition to $z(t)$, let's consider a second bivariate variable $w(t)$ with Cartesian components $g(t)$ and $h(t)$. The cross covariance function at zero lag, or simply the covariance between the components of z and w can be formed as

$$
\begin{aligned}
E\left\{\mathbf{z}^{\mathrm{T}}\mathbf{w}\right\} &= E\left\{\begin{bmatrix} (x-\mu_x)(g-\mu_g) & (x-\mu_x)(h-\mu_h) \\ (y-\mu_y)(g-\mu_g) & (y-\mu_y)(h-\mu_h) \end{bmatrix}\right\} \\
&= \begin{bmatrix} C_{xg} & C_{xh} \\ C_{yg} & C_{yh} \end{bmatrix}
\end{aligned}
$$

The meaning of each individual entry of this matrix is obvious but the interpretation of all of them at once may be less so.

The covariance of bivariate variables

Alternatively, let's use the complex representations of z and w

$$z = x + iy$$
$$w = g + ih$$

but assume for simplicity that $\mu_z = \mu_w = 0$.

The covariance of bivariate variables

Alternatively, let's use the complex representations of z and w

$$z = x + iy$$
$$w = g + ih$$

but assume for simplicity that $\mu_z = \mu_w = 0$.

A definition of the covariance between these two complex r.v. is

$$\begin{aligned} C_{zw} &= E[z^* w] \\ &= E[(x - iy)(g + ih)] \\ &= E[xg] + E[yh] + i(E[xh] + E[-yg]). \end{aligned}$$

The covariance of bivariate variables

Alternatively, let's use the complex representations of z and w

$$z = x + iy$$
$$w = g + ih$$

but assume for simplicity that $\mu_z = \mu_w = 0$.

A definition of the covariance between these two complex r.v. is

$$\begin{aligned} C_{zw} &= E[z^* w] \\ &= E[(x - iy)(g + ih)] \\ &= E[xg] + E[yh] + i(E[xh] + E[-yg]). \end{aligned}$$

Beware that the definition of the covariance between complex r.v. may differ. Sometimes it is defined as $E[zw^*]$. Our convention here is the same as Matlab.

The covariance of bivariate variables

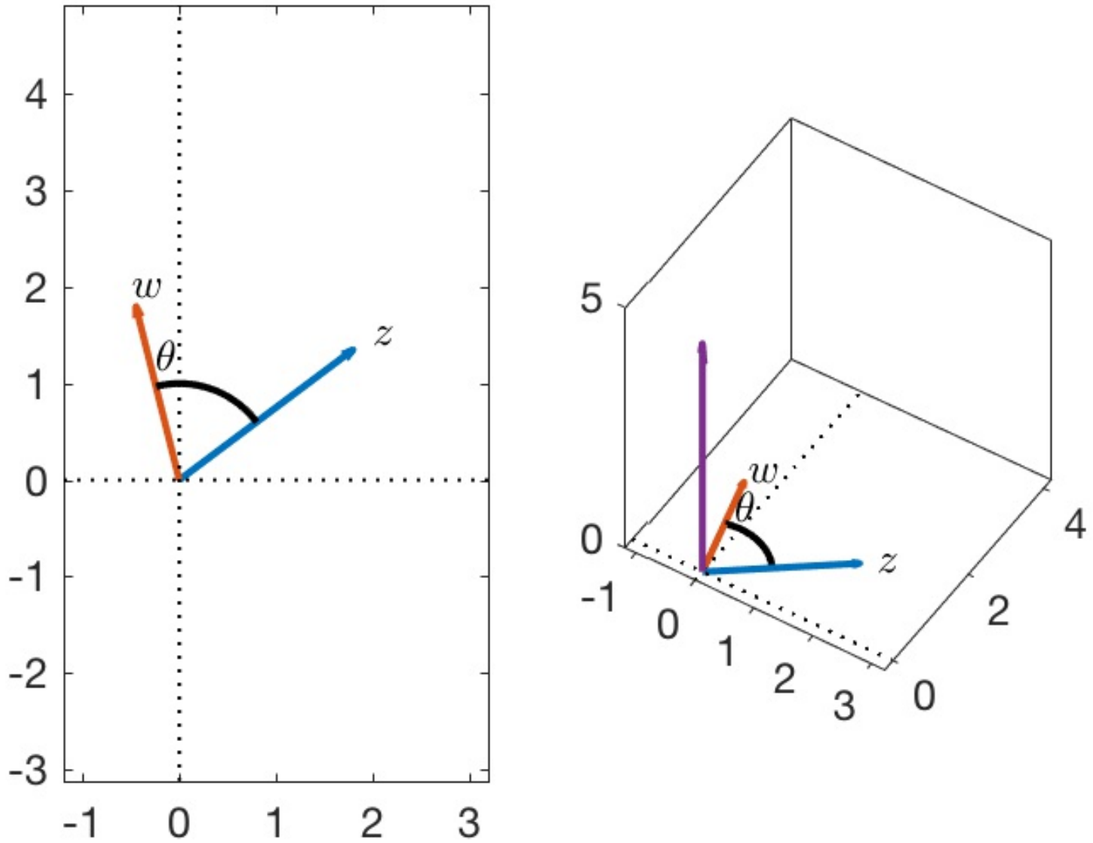
Simple geometry reveals that the real part of the complex covariance is the expectation of the dot product [or inner product, noted $(\cdot)$] while the imaginary part is the expectation of the magnitude of the vector cross product [or outer product, noted $(\times)$]:

$$\begin{aligned} C_{zw} &= E[xg] + E[yh] + i(E[xh] + E[-yg]) \\ &= E[\mathbf{z} \cdot \mathbf{w}] + iE[||\mathbf{z} \times \mathbf{w}||] \\ &= E[|z||w| \cos\theta] + iE[|z||w| \sin\theta] \end{aligned}$$

where θ is the instantaneous geometric angle in the Cartesian plane between $\mathbf{z}$ and $\mathbf{w}$, measured positively counterclockwise relative to the geometric angle of $\mathbf{z}$.

The covariance of bivariate variables

$$C_{zw} = E[|z||w|\cos\theta] + iE[|z||w|\sin\theta]$$



The covariance of bivariate variables

$$C_{zw} = E[|z||w| \cos\theta] + iE[|z||w| \sin\theta]$$

The covariance of z and w is a complex number which argument, or phase, is

$$\mathrm{Arg}[C_{zw}] = \mathrm{atan}\left\{ \frac{E[|z||w| \sin\theta]}{E[|z||w| \cos\theta]} \right\} \neq E[\theta].$$

The covariance of bivariate variables

$$C_{zw} = E[|z||w|\cos\theta] + iE[|z||w|\sin\theta]$$

The covariance of z and w is a complex number which argument, or phase, is

$$\mathrm{Arg}[C_{zw}] = \mathrm{atan}\left\{\frac{E[|z||w|\sin\theta]}{E[|z||w|\cos\theta]}\right\} \neq E[\theta].$$

The absolute value of C_{zw} is clearly a measure of the covariance between the two bivariate variables, but its phase is not $E[\theta(t)]$: it is not the expectation, or mean, of the geometric angle between $\mathbf{z}$ and $\mathbf{w}$.

The covariance of bivariate variables

$$C_{zw} = E[|z||w|\cos\theta] + iE[|z||w|\sin\theta]$$

The covariance of z and w is a complex number which argument, or phase, is

$$\mathrm{Arg}[C_{zw}] = \mathrm{atan}\left\{\frac{E[|z||w|\sin\theta]}{E[|z||w|\cos\theta]}\right\} \neq E[\theta].$$

The absolute value of C_{zw} is clearly a measure of the covariance between the two bivariate variables, but its phase is not $E[\theta(t)]$: it is not the expectation, or mean, of the geometric angle between $\mathbf{z}$ and $\mathbf{w}$.

However, $\mathrm{Arg}[C_{zw}]$ can still be seen as an indication of the relative angle of covariance. If $\mathrm{Arg}[C_{zw}] = 0$ this indicates that the covariance occurs with $\mathbf{z}$ and $\mathbf{w}$ aligned and pointing in the same direction, and if $\mathrm{Arg}[C_{zw}] = \pi/2$ this indicates that the covariance occurs with $\mathbf{w}$ at right angle counterclockwise from $\mathbf{z}$.

The correlation of bivariate variables

From the complex covariance, we can define the complex correlation between **z** and **w** as

$$\rho_{zw} \equiv \frac{C_{zw}}{\sqrt{C_{zz}C_{ww}}} = |\rho_{zw}|e^{i\text{Arg}[\rho_{zw}]}$$

Since C_{zz} and C_{ww} are real numbers, the phase of ρ_{zw} is the same phase as the phase of C_{zw}: $\text{Arg}[\rho_{zw}] = \text{Arg}[C_{zw}]$

The correlation of bivariate variables

From the complex covariance, we can define the complex correlation between **z** and **w** as

$$\rho_{zw} \equiv \frac{C_{zw}}{\sqrt{C_{zz}C_{ww}}} = |\rho_{zw}|e^{i\text{Arg}[\rho_{zw}]}$$

Since C_{zz} and C_{ww} are real numbers, the phase of ρ_{zw} is the same phase as the phase of C_{zw}: $\text{Arg}[\rho_{zw}] = \text{Arg}[C_{zw}]$

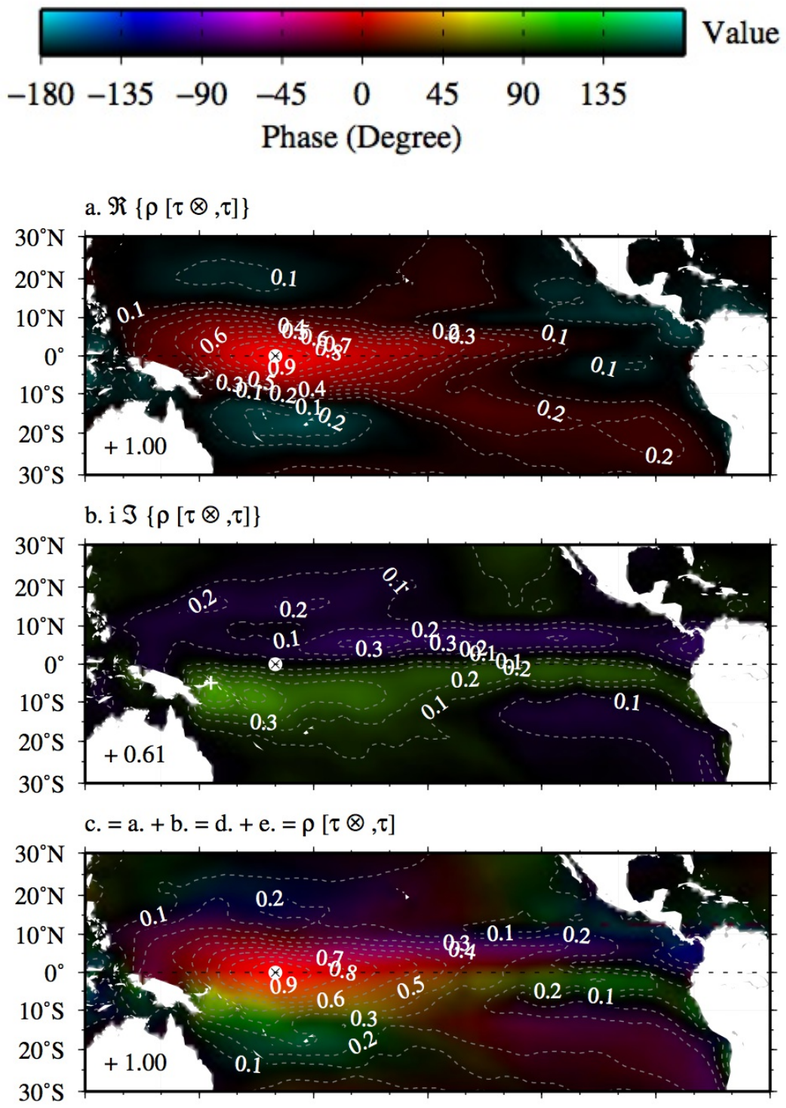
Note that the "vector" regression model for w based on the covariance between z and w is (see Lecture 3)

$$w_m(t) = \frac{C_{zw}}{C_{zz}}z(t) = \frac{|C_{zw}|}{C_{zz}}e^{i\text{Arg}[C_{zw}]}z(t)$$

which shows that the regressed vector $w_m(t)$ is rotated counterclockwise by $\text{Arg}[C_{zw}]$ from the direction of $z(t)$. This is valid after removing the means.

The correlation of bivariate variables

Panel c is the correlation ρ between the 10-m wind at 170E, 0N and 10-m winds at all other locations from ECMWF reanalyses. The real part of ρ is shown in panel a and the imaginary part is shown in panel b as a pure imaginary number so that c = a + b.



The complementary of a vector

The *complementary of a vector* $\mathbf{z}$, noted $\mathbf{z}^c$, is, using complex notation

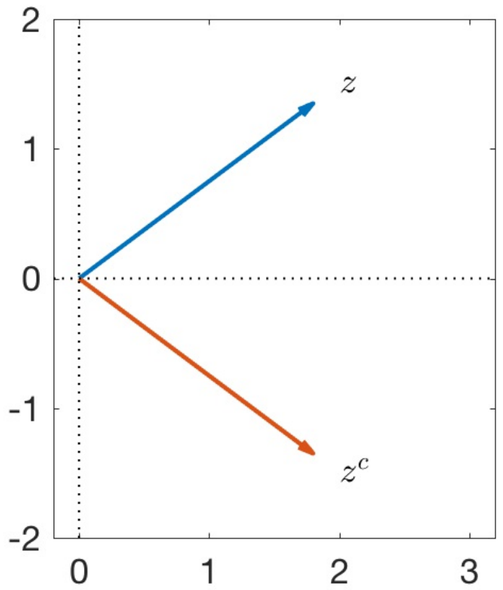
$$z^c \equiv z^* = x - iy$$

The complementary of a vector

The *complementary of a vector* $\mathbf{z}$, noted $\mathbf{z}^c$, is, using complex notation

$$z^c \equiv z^* = x - iy$$

Geometrically, it is the vector $\mathbf{z}$ flipped with respect to the real axis. If a bivariate time series represents a vector rotating in one direction, then its complementary is rotating in the opposite direction.



The complementary correlation of bivariate variables

The *complementary covariance* between $\mathbf{z}$ and $\mathbf{w}$ is

$$\begin{aligned} C_{z^c w} &= E[(z^*)^* w] \\ &= E[(x+iy)(g+ih)] \\ &= E[xg] - E[yh] + i(E[xh] + E[yg] \end{aligned}$$

The *complementary or reflectional correlation* is

$$\rho_{z^c w} \equiv \frac{C_{z^c w}}{\sqrt{C_{z^c z^c} C_{ww}}}$$

where $C_{z^c z^c} = C_{zz}$

Auto-covariance and variance ellipses

The special cases of standard and complementary auto covariances of a physical vector with itself can be related to the concept of *variance or standard deviation ellipses*.

Auto-covariance and variance ellipses

The special cases of standard and complementary auto covariances of a physical vector with itself can be related to the concept of *variance or standard deviation ellipses*.

For a bivariate r.v. z, the fixed angle θ_M between the real axis positive direction and the so-called major axis of the standard deviation ellipse is

$$\theta_M = \frac{1}{2}\mathrm{Arg}\left[C_{zz}\right] = \frac{1}{2}\arctan\left[\frac{2E[xy]}{E[x^2]-E[y^2]}\right],$$

while $\theta_M + \pi/2$ defines the direction of the so-called minor axis.

Auto-covariance and variance ellipses

The variances a_v^2 and b_v^2 of the bivariate variable along the major and minor axes, respectively, are given by

$$\begin{aligned} a_v^2 &= \frac{1}{2}[C_{zz} + |C_{z^c z}|] \\ &= \frac{1}{2}\left\{E[x^2] + E[y^2] + \sqrt{(E[x^2] - E[y^2])^2 + 4(E[xy])^2}\right\}, \end{aligned}$$

$$\begin{aligned} b_v^2 &= \frac{1}{2}[C_{zz} - |C_{z^c z}|] \\ &= \frac{1}{2}\left\{E[x^2] + E[y^2] - \sqrt{(E[x^2] - Ey^2)^2 + 4(E[xy])^2}\right\}. \end{aligned}$$

Anticipating Lecture 5, a_v^2 and b_v^2 are the 2 eigen values of the 2×2 cross covariance matrix of the Cartesian component, while the associated eigen vectors are $[\cos\theta_M, \sin\theta_M]$ and $[-\sin\theta_M, \cos\theta_M]$.

Auto-covariance and variance ellipses

The ratio of the absolute value of the complementary auto correlation to the absolute value of the auto correlation is a measure of the absolute *linearity* of the standard deviation ellipse

$$|\lambda| = \frac{C_{z^c z}}{|C_{zz}|} = \frac{a_v^2 - b_v^2}{a_v^2 + b_v^2}.$$

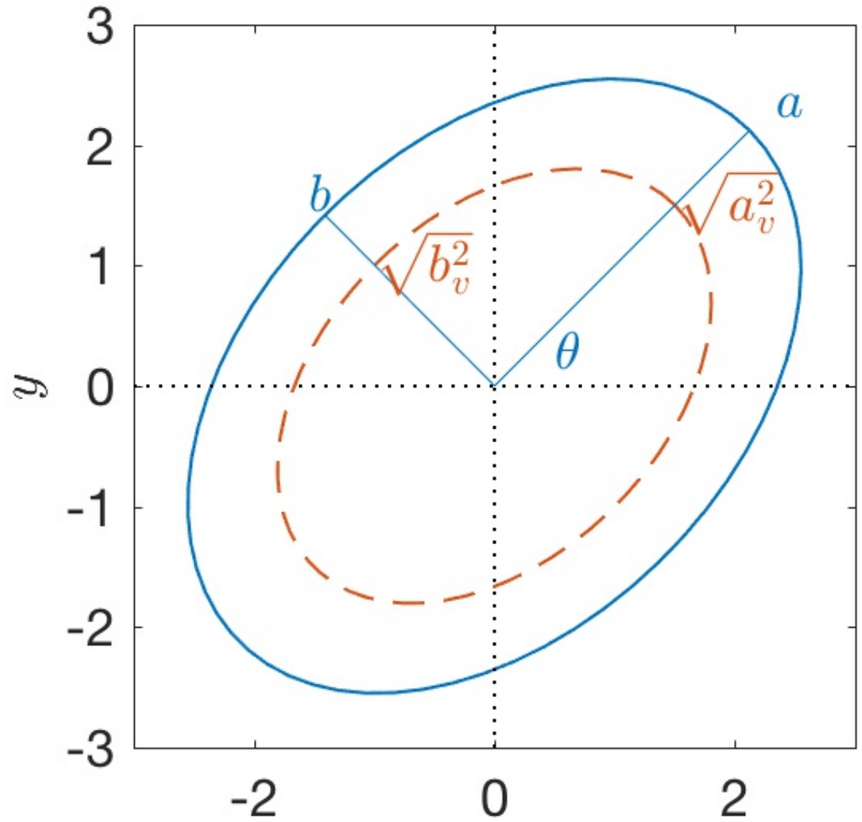
If $|\lambda| = 1$ the ellipse is flat and if $|\lambda| = 0$ the ellipse is a circle. λ is a relative of the more commonly known eccentricity parameter of the ellipse which is

$$e = \sqrt{1 - \frac{b_v^2}{a_v^2}}.$$

See Lilly and Gascard (2006) and Lilly J, Olhede S. (2010).

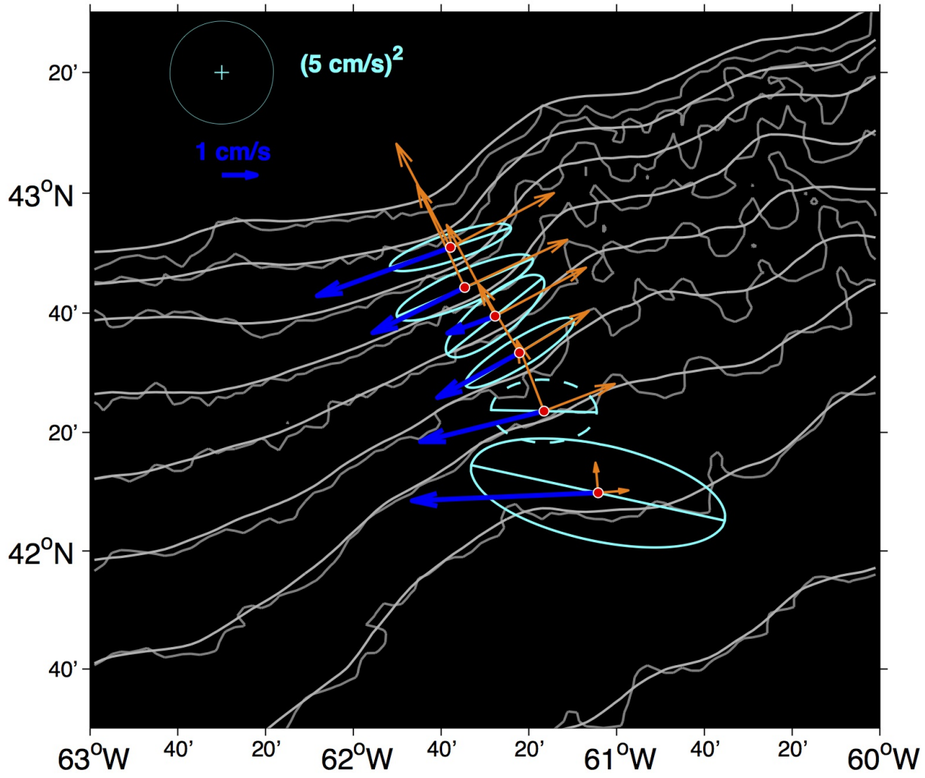
Standard deviation or variance ellipses

If your complex signal is $z(t) = e^{i\theta}\left[a\cos(t) + i\sin(t)\right]$ which represents an ellipse, then the major axis and minor semi-axis of the standard deviation ellipse are $a_v = \sqrt{a^2/2} < a$ and $b_v = \sqrt{b^2/2} < b$.



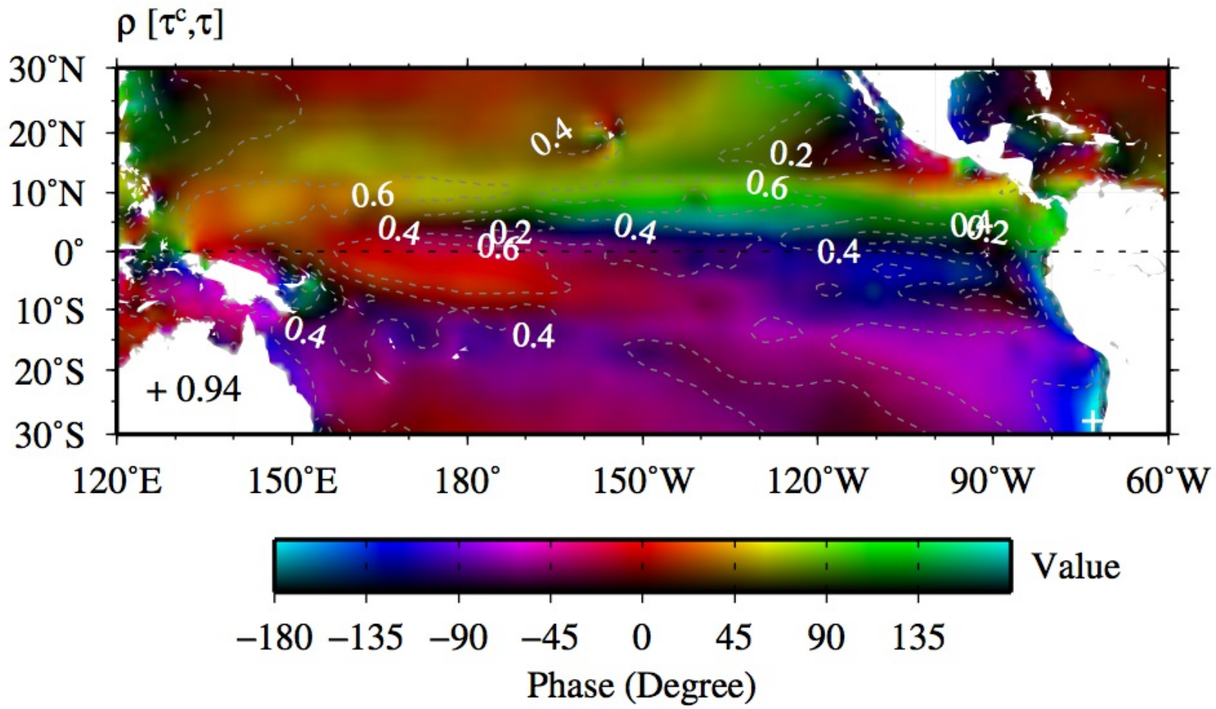
Standard deviation or variance ellipses

Example: Mean currents (blue) and standard deviation ellipses (light blue) of near bottom currents from the RAPID WAVE experiment. Orange axes show topography gradient. See Hughes et al. 2013



Standard deviation or variance ellipses

Example: Complementary auto correlation of 10-m winds at all locations from ECMWF reanalyses.



Standard deviation or variance ellipses

The standard ellipse is a statistical description of the variance of a bivariate variable and does not mean that any underlying variability is actually elliptical in time. Standard deviation ellipses can be computed from a number of individual pairs of Cartesian components of a bivariate quantity without these points actually forming consecutive time series.

Standard deviation or variance ellipses

The standard ellipse is a statistical description of the variance of a bivariate variable and does not mean that any underlying variability is actually elliptical in time. Standard deviation ellipses can be computed from a number of individual pairs of Cartesian components of a bivariate quantity without these points actually forming consecutive time series.

Example: Variance ellipses from drifters from Lumpkin and Johnson 2013

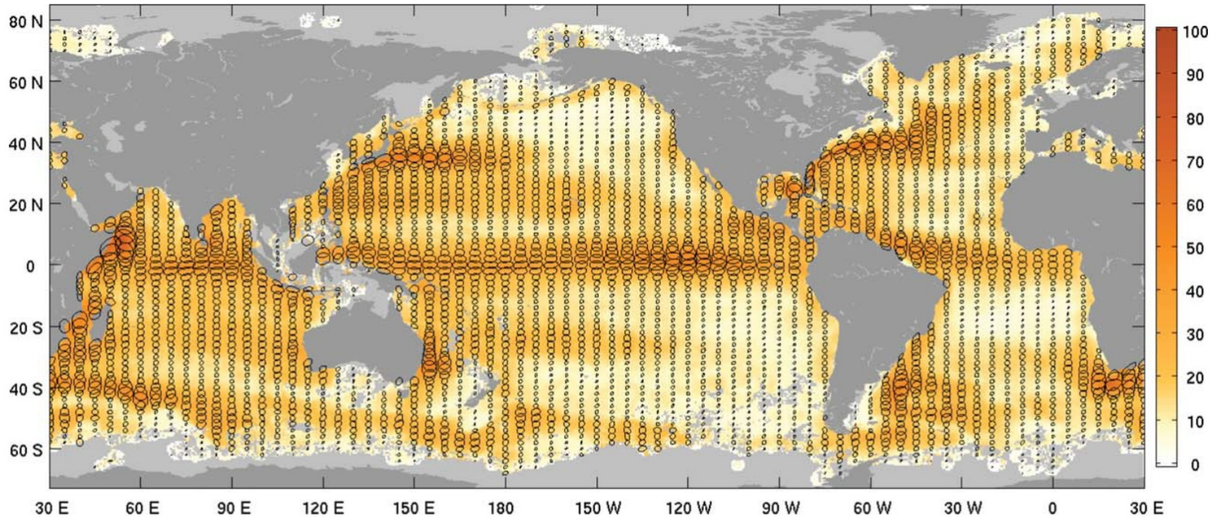


Figure 3. Variance ellipses centered every 5° longitude by 2° latitude (black lines) plotted over the square root of the magnitude of variance (colors in cm s^{-1}). Light gray areas have less than 0.8 drifter days per square degree.

Practical session

Please download data at the following link:

Please download the Matlab code at the following link:

Make sure you have installed and tested the free jLab Matlab toolbox from Jonathan Lilly at www.jmlilly.net/jmlsoft.html