Comps Question based on MPO581 – Applied Data Analysis

Suppose you are given two time series: a temperature anomaly T'(t) in K, and a wind anomaly u'(t) in m/s. Each has been sampled at the same 1000 sampling times, and the mean has been removed from each data set (hence the word "anomaly" and the primes T' and u').

These 2 data sets can be written in terms of the part that is correlated, called "signal" S, plus the uncorrelated "noise" remainder n that is different for each of the two series. Mathematically, we can write this decomposition as:

$T'(t) = a_T\, S(t) + n_T(t)$
$u'(t) = a_u\, S(t) + n_u(t)$

where S(t) is a standardized (mean 0, variance 1), dimensionless series and the amplitude coefficients a_T and a_u have been defined to minimize the root mean square (RMS) of n_T and n_u. That is, they are *least-squares regression* coefficients. The noise variables are thus uncorrelated with S, although they could be correlated with each other: In a high dimensional space, two vectors that are both perpendicular to a given vector are not necessarily perpendicular to each other.

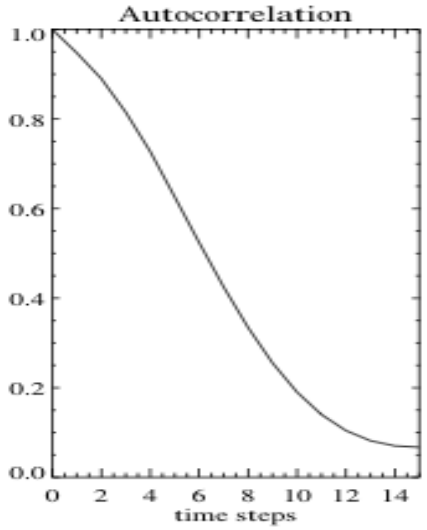
QUESTIONS:

a) Given the variances below, what is the (absolute value of) the correlation coefficient cor(T',u')? Explain your work.

var(T') = $\overline{T'T'} = \frac{1}{1000}\sum_{i=1}^{1000} T_i'^2 = 5\ \mathrm{K}^2$

var(u') = 9 $(\mathrm{m/s})^2$
var(n_T) = 3 K^2
var(n_u) = 5 $(\mathrm{m/s})^2$

b) Autocorrelation as a function of time lag is shown here, for both for T' and u' series →



About how many "degrees of freedom" (df, think of them as "independent chunks of information") are in each time series?

Hint: the characteristic size of the independent "bumps" or "events" on the time series is taken to be twice the e-folding lag, or about 15 time steps, so divide the total number of time steps by that.

c) Based on your answers to a) and b), is the correlation coefficient from (a) statistically significant at 95% confidence? That is, is its absolute value larger than the indicated values in the .05 column in this standard table based on the Student t-test?

Level of Significance (p) for Two-Tailed Test	.10	.05	.02	.01
df				
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
5	.669	.754	.833	.874
10	.497	.576	.658	.708
25	.323	.381	.445	.487
50	.231	.273	.322	.354
100	.164	.195	.230	.254

d) If additional random noise were added to the u'(t) series (for example by switching to a cheaper, noisier, anemometer), the variance of u will go up while corr(T',u') decreases. What is the effect on the *linear regression of u on T* (i.e., on coefficient R_{uT} in this equation):
$u' = R_{uT} T'$ + (least-squares minimized residual)

Sketch a scatter plot of u' on they y axis vs. T' on the x axis with a few points and a fitted line, indicating what (summed squared) distance is being minimized. Now sketch points in the presence additional "vertical" (u') random noise. Explain the symmetry argument: can such noise pull asymmetrically on the two ends of the line and thus affect its slope?

e) What is the effect on R_{uT} if additional random noise is added to the T' data instead of the u' data? Hint: consider the sketch above again, and imagine adding huge "horizontal" T' noise: what happens to the slope of the line?