MPO 542 – Applied Data Analysis Exam   Name_______________________________________
April 15, 2014

Answer 4 of the 6 questions for credit. Each problem has 25 points.

Indicate here which ones to count: ____

Space for final answers if pages get crowded (label clearly please):
More scratch room available on last page.

1. Fill in the open parts of this table for variables q and T (1 point each).
You may invoke symbols from higher rows within lower rows.

name	bar & prime	summation	integrals over probability density	Matlab or pseudocode	math symbol
mean (1st moment)				mean(q)	xxx
deviations or perturbations			xxx	xxx	xxx
median	xxx	xxx		xxx	xxx
variance (2nd moment)					xxx
skewness (3rd moment)				mean(q'^3)	xxx

standard deviation		xxx		sqrt(var(q))
				std(q)	

covariance btw. T & q					xxx
correlation coefficient			xxx		
fraction of variance explainable by regression		xxx	xxx	xxx	r^2
root mean square of variable (may include its mean!)				RMS(q)	xxx

2. Joint distributions of temperature T (K) and water vapor mixing ratio q (g/kg) are often used in fractional cloudiness schemes for coarse-grid atmosphere models.

a. **Science setting (3)**: Think up and explain a plausible physical scenario for positive or negative correlation between T' and q' fluctuations within some 100km x 100km x 1km cube of air. (I can imagine phenomena causing either sign).

b. **Sketch** a joint probability density p(T, q), and its two marginals. Assume both marginals are Gaussian, and make clear your nonzero correlation from a. (of whichever sign). In what corner of the diagram is cloudy (saturated) air? (7)

c. **Sketch** the *conditional* distribution p(q | T= $\overline{T} + 1\sigma_T$), comparing its shape to the marginal distribution (even though their units are different). (5)

d. **Sketch** and label the axes (including numbers on the y axis) of one of the *cumulative marginal distributions. (5)*

e. **What are the units** of the joint distribution? Write the calculus derivative notation for p(T,q) which makes these units clear. (5)

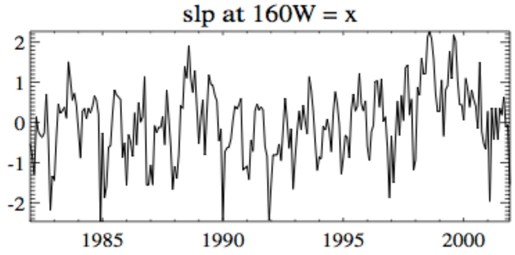
3. Variance as a quantitative number: trends vs. Fourier harmonics

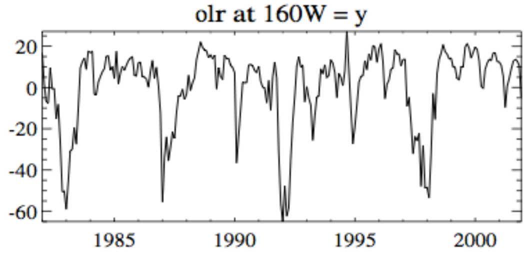
a. What is the variance of a linear trend with slope A in a finite time series? To answer, compute the mean square of y'=At on the t interval [-1,1]. Don't forget to divide by 2 (the length of the interval) when you evaluate the overbar in $\overline{y^2}$(because the bar is an average, not just an integral). (10)

b. Q: What is the variance of a sine or cosine curve with amplitude A? Hint for calculation: the answer is the same for sine and cosine, and $\sin^2 + \cos^2 = 1$, so average both sides of $\sin^2 + \cos^2 = 1$ over a complete cycle [-π,π] and solve. (10)

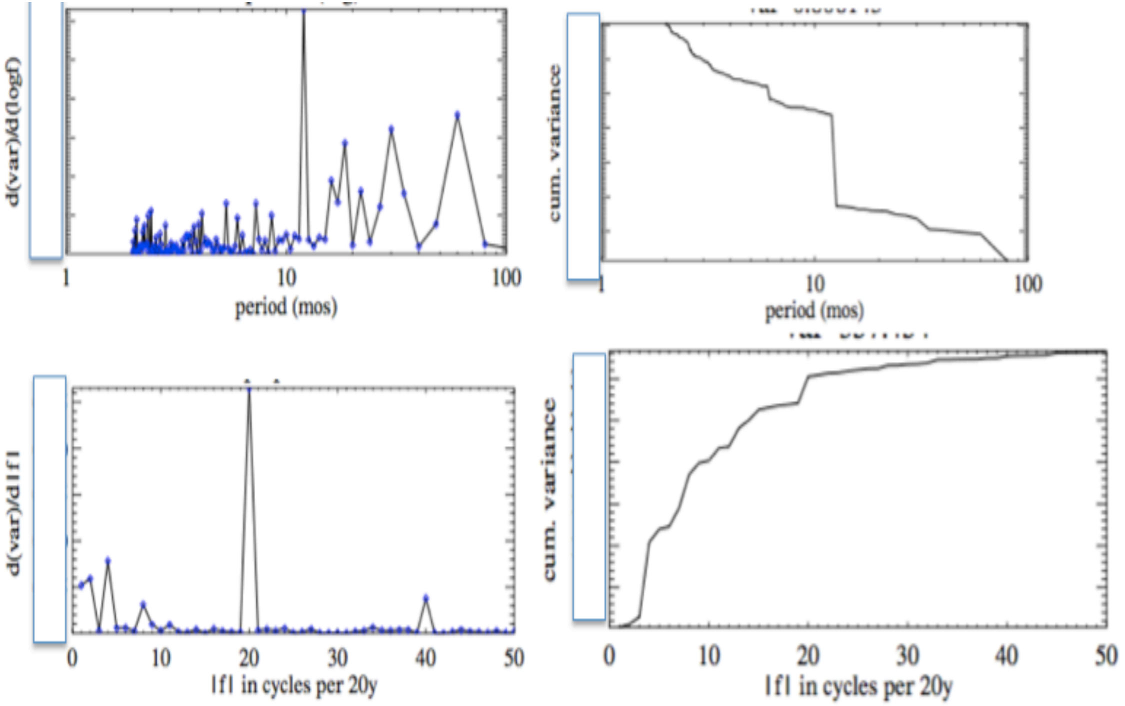
c. Sketch At and Asin(t) on the interval on the t interval [-π/2, π/2], using the fact that that sin(x)~x for small x. . Which has larger variance? Explain how your answer illustrates or is consistent with a. and b. (5)

4. SLP (mb) and OLR (Wm^{-2}) time series from our 20-year equatorial monthly time-longitude dataset are plotted below, for one longitude (160W).





a. Identify which series corresponds to each of these 4 spectrum plots. Estimate and write in the covered-up numbers on each plot's vertical axis. Indicate units on each plot's vertical axis or in the plot title area. (15)



b. Sketch the autocovariance functions for these two variables, and label the axes. Explain the differences you are indicating with the sketch, and how you derived those from features of the spectra and/or of the raw time series. Use the words "redder" and "periodicitiy" in your explanation. (10)

5. Interpreting multiple regression.
ENSO cycles in the central Pacific involve warm phases -- when suface T is anomalously warm, air converges more than usual, and rainfall is enhanced -- interspersed with cold phases when the opposite anomalies occur.

Student A decides to "explain" or "predict" precipitation anomalies associated with ENSO cycles as resulting from T anomalies, using a simple univariate regression:
P'(t) = R T'(t)   +resid1
and derives the usual least-squares regression coefficient, R = corr(T',P') σ_P/σ_T.

Student B prefers to "explain" or "predict" P as a combination of thermodynamics and dynamics, using multiple regression on T and wind divergence d:
P'(t) = $\boldsymbol{m_1}$ ***T'(t) +*** $\boldsymbol{m_2}$ ***d'(t)   +resid2.***

a. **Express** Student B's problem in a **d = Gm** form as used in the book and class. (5)

$$\begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \quad & \quad \\ \\ \\ \\ \end{bmatrix} \left[\qquad \right] + resic$$

b. The best estimate of **m** (minimizing the summed squared residual SSR) is $(\mathbf{G}^T\mathbf{G})^{-1}$ $\mathbf{G}^T\mathbf{d}$**. Use words to explain** the meaning or interpretation of the factors $(\mathbf{G}^T\mathbf{G})^{-1}$ and $\mathbf{G}^T\mathbf{d}$ in this case. Use "projection" or "covariance" to describe what $\mathbf{A}^T\mathbf{B}$ measures about two time series (column vectors) **A** and **B**. (5)

c. Based on the associations noted in the Background information, what is the challenge in solving for **m**? Why, and how would it manifest mathematically in your calculation? Sketch SSR contours in (m1, m2) space. How might Student B address the uncertainty or ambiguity from this challenge? (10)

d. The following equality is incomplete. Use the Chain Rule for P(T,d) to complete the equation. Indicate which derivatives in the equation correspond to R and m_1. (5)

$$\frac{dP}{dT} = \frac{\partial P}{\partial T}\Big|_d$$

6. Orthogonal decompositions are nice because we can postulate a relationship in physical space (d = term1 + term2 + …), and then evaluate its truth using observations of variability and the fact that $d^2 = \text{term1}^2 + \text{term2}^2 + \text{term3}^2$…with no cross-terms.

a. What is the mathematical condition of orthogonality? Use overbar notation. (5)

b. Show or explain why that condition is met for each of the following (5 each):

i. $q = \overline{q} + q'$

ii. $q'(t) = A_0 + A_1 \cos(w_1 t) + B_3 \sin(w_3 t) + \ldots$

iii. q' = mT' + resid, *<u>if and only if</u> m is a least-squares regression coefficient*

iv. $q'(x,t) = EOF_1(x)\, PC_1(t) + EOF_2(x)\, PC_2(t) + \ldots$, from a Maximum Covariance Analysis (MCA) using Singular Value Decomposition (SVD) or eigendecomposition.

Scratch space: