Final Exam: MAE 127

Tuesday, June 7, 2005

1. Consider data set X in Figure 1.

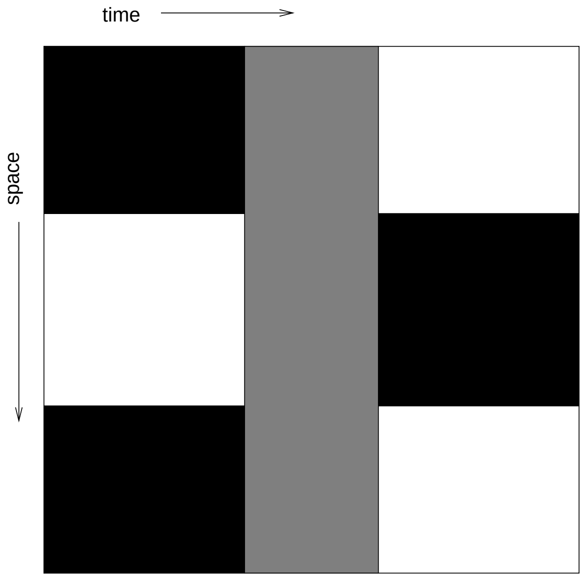

Figure 1: Sample data distribution. Black areas can be considered to have a numerical value of +1, gray areas 0, and white areas -1.

a. Sketch an estimate of the temporal first mode empirical orthogonal function.

b. Sketch an estimate of the spatial first mode empirical orthogonal function.

c. What fraction of the variance in the idealized data would you expect these modes to represent and why?

a-b. Figure 2 shows the first mode in time (top) and space (bottom). Here the scale is arbitrary, since we don't know how many data points are available, but the positions of positive and negative values are determined by the spatial and temporal structure of the original data.

c. This first mode EOF explains 100% of the variance, since there is no obvious noise and no pattern of variability that cannot be represented by this pattern.

2. Data plotted in Figure 3 show monthly average air temperatures measured at Scripps Pier. A least squares fit to the data indicates: $T = T_o + T_1 \cos(2\pi(m-1)/12) + T_2 \sin(2\pi(m-1)/12)$, where m is the integer number of the month, and $T_o = 17.88$, $T_1 = -3.11$, $T_2 = -1.04$. A Fourier transform of the 12 data points divided by 12 produces the following values:

```
fft(T)/12 =
  17.8794
  -1.5545 + 0.5206i
  -0.2738 - 0.3016i
```

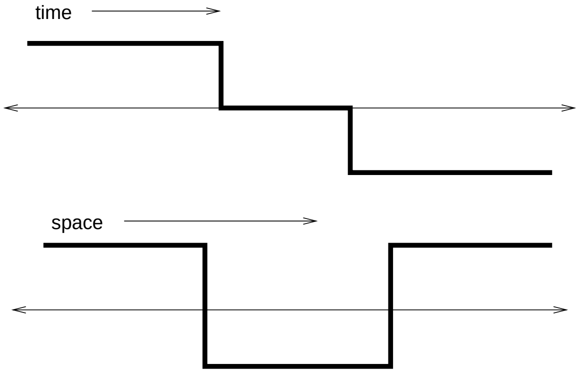

Figure 2: Schematic of first mode EOF in time (top) and space (bottom) for data sketched in Figure 1.

```
  0.1826 + 0.1348i
  0.0194 + 0.0377i
 -0.0644 - 0.0498i
  0.4321
 -0.0644 + 0.0498i
  0.0194 - 0.0377i
  0.1826 - 0.1348i
 -0.2738 + 0.3016i
 -1.5545 - 0.5206i
```

(a) What is the mean temperature for this record? How can you tell?

(b) Are the results of the Fourier transform consistent with the least squares fit? Explain.

a. The mean temperature is 17.88° C. You can tell this either from T_o in the least-squares fit or from the frequency zero result of the Fourier transform.

b. Results from the Fouerier transform are consistent with the least-squares fit. The least-squares fit for coefficients T_1 and T_2 computes the amplitude of the annual cycle. This is also the lowest frequency resolved by the Fourier transform. The results are equivalent. T_1 is equal to 2 times the real part of the fft of T divided by 12. T_2 is equal to 2 times the imaginary part of the fft of T divided by 12. This is exactly as we expect based on our understanding of the Fourier transform and the Matlab definitions of Fourier transforms.

3. Figure 4 shows two (artificial) time series, each containing 100 points spaced at intervals of 1 hour. Record A is random normally distributed white noise plus a cosine with frequency $\nu = 2\pi/20$. Record B is normally distributed noise plus a sine with frequency $\nu = 2\pi/20$. The records were produced with the following commands:

```
A=randn(100,1)+cos(2*pi*(0:99)'/20);
B=randn(100,1)+sin(2*pi*(0:99)'/20);
```

a. Roughly estimate the correlation of these two records?

b. What should the spectrum of record A look like?

c. What are the lowest frequency and highest frequencies resolved in the spectrum of A?

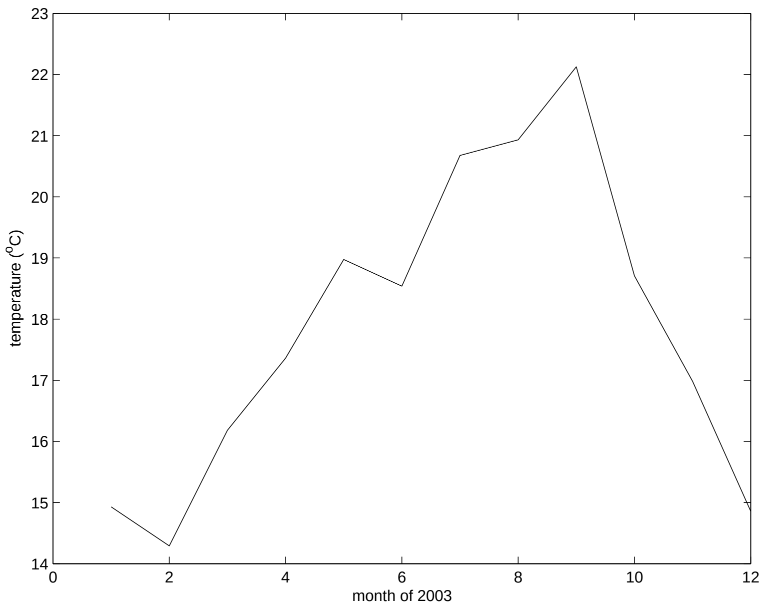

Figure 3: Monthly average air temperature from Scripps Pier for 2003.

d. At what frequencies do you expect statistically significant coherence (if any)?

e. What is the phase of the coherence?

a. The records are white noise plus a cosine and white noise plus a sine. White noise is uncorrelated, and cosine and sine are uncorrelated, so the correlation should be zero.

b. The spectrum of A should be roughly flat at all frequencies, except the frequency equivalent to 5 cycles per 100 data points (or 1 cycle per 20 data points = $2\pi/20$*), as illustrated in Figure 5.*

c. The lowest frequency resolved is 1 cycle per 100 data points ($2\pi/100$*), and the highest frequency, the Nyquist frequency, is 50 cycles per 100 data points (*$2\pi 100/100$*).*

d. Statistically significant coherence will occur at $\nu = 2\pi/20$ *and at no other frequencies (except by chance).*

e. The phase of the coherence will be such that A appears to lead B by $\pi/2$ *radians or 90° at frequency* $\nu = 2\pi/20$*. You can tell this because at frequency* ν*, the Fourier transform of A will be of the form* $1 + 0i$ *and the Fourier transform of B will be of the form* $0 + 1i$*. Thus product of them* C_k *will be* $-i$*, so the phase will be* $\text{atan}(1/0) = \pi/2$*.*

4. A set of observed quantities, a, b, and c, all have non-Gaussian probability density functions of the form:

$$P(x)\,dx = \begin{cases} \frac{1}{2}dx & \text{for } 0 \le x \le 2 \\ 0. & \text{otherwise} \end{cases}$$

(Assume that a, b, and c are statistically independent.)

a. What is the expected value of $d = a + 2b + c$?

b. What is the standard deviation of a (or b or c)?

c. What is the the standard deviation of d? (Hint: You can estimate this either from the PDF or by using error propagation.)

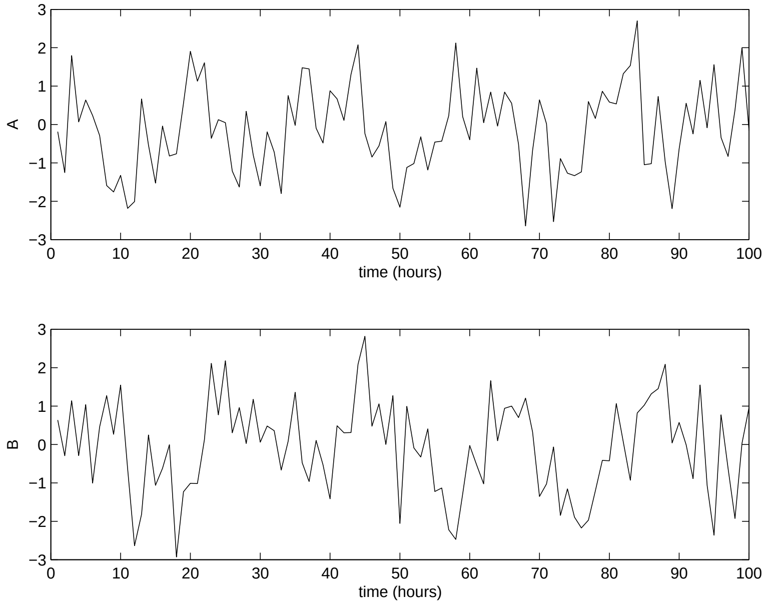

Figure 4: Artificial time series A (top) and B (bottom).

a. The expected value of d is $\langle d\rangle = \langle a+2b+c\rangle = \langle a\rangle + 2\langle b\rangle + \langle c\rangle$. *The expected value of a, b, and c is 1, which you can obtain either by inspection, or by integrating* $\int_{-\infty}^{\infty} xP(x)\,dx = 1/2\int_0^2 x\,dx = x^2/4|_0^2 = 1$. *Thus* $\langle d\rangle = 4$.

b. The standard deviation σ_a *of a is*

$$\sigma_a^2 = \int_{-\infty}^{\infty}(x-\langle x\rangle)^2 P(x)\,dx = \frac{1}{2}\int_0^2 (x-1)^2\,dx = \frac{1}{2}\int_{-1}^{1} x^2\,dx = \frac{x^3}{6}|_{-1}^{1} = \frac{1}{3}$$

Thus $\sigma_a = 1/\sqrt{3}$.

c. The standard deviation of d is $\sigma_d^2 = \sigma_a^2 + 4\sigma_b^2 + \sigma_c^2 = 6/3 = 2$. *Thus* $\sigma_d = \sqrt{2}$.

Alternatively, you could compute this using the PDF:

$$\sigma_d^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(a+2b+c-\langle d\rangle)^2\,da\,db\,dc = \frac{1}{8}\int_0^2\int_0^2\int_0^2 (a+2b+c-4)^2\,da\,db\,dc$$

This produces the same result after some algebra.

5. A simple limit of least-squares fitting involves fitting one unknown. Assume you have the data:

$$T = [-3 \quad -1 \quad 2], \tag{1}$$

which have been collected at times

$$t = [-1 \quad 0 \quad 1]. \tag{2}$$

The uncertainty in each value of T is 2.

a. You'd like to find a best fit to the data of the form $T = \alpha t$. How do you set up the problem?

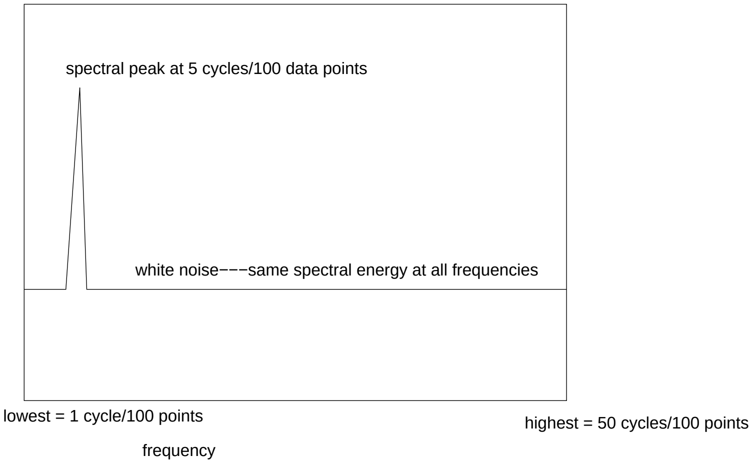

Figure 5: Schematic of spectrum of A (assuming sufficient averaging to reduce noise).

b. What is the best estimate of α?

c. Using the uncertainty for T, estimate the uncertainty in α.

d. What is χ^2 for this data?

e. What would you predict χ^2 to be? Are the results consistent?

a. To solve $T = \alpha t$, first define a matrix $A = t^T$. Then solve $\alpha = (A^T A)^{-1} A^T T^T$.

b. The results of this are $A^T A = 1 + 0 + 1 = 2$. Thus $(A^T A)^{-1} = 1/2$. The quantity $A^T T^T = (-1)(-3) + (0)(-1) + (1)(2) = 5$. Therefore $\alpha = 5/2$.

c. The variance comes from the diagonal of $(A^T A)^{-1}$ multiplied by σ_T^2. In this case $\sigma_\alpha = \sigma_T \sqrt{(A^T A)^{-1}} = sqrt2$. There was a missing square root in the crib sheet, so your results were graded leniently.

d. The weighted misfit $\chi^2 = \sum_{i=1}^{3} (\alpha t_i - T_i)^2 / \sigma_T^2 = ((1/2)^2 + (-1)^2 + (1/2)^2)/2^2 = 3/8$.

e. With 1 fitting parameter and 3 equations, we expect χ^2 to be about $N - M = 3 - 1 = 2$. The value of 3/8 that we found here is plausibly consistent, though perhaps a bit small.

6. The California Cooperative Oceanic Fisheries Investigations (CalCOFI) have surveyed water off coastal California for more than 55 years, starting in March 1949. In this time, CalCOFI has tracked large-scale temperature change in the ocean and also changes in zooplankton biomass and sardine spawning. Since 1985, the core of the CalCOFI data set has consisted of measurements collected 4 times a year at 66 locations (or “stations”), shown as circles in Figure 6. Prior to 1985, CalCOFI cruises covered a larger geographic span, shown schematically by the line in Figure 6. This might make you wonder if variability for the entire larger domain can be fully represented by the 66 measurement stations that are used now.

Explain how you might use one or more methods discussed in this course to evaluate the current CalCOFI sampling. For simplicity, assume that you concentrate your analysis on temperature data from 10 m depth, collected 4 times per year at each measurement station. Also assume that the larger sampling grid used prior to 1985 has about 90 stations.

There is no single correct answer to this question. You are encouraged to use equations as well as words to help explain your proposed strategy.

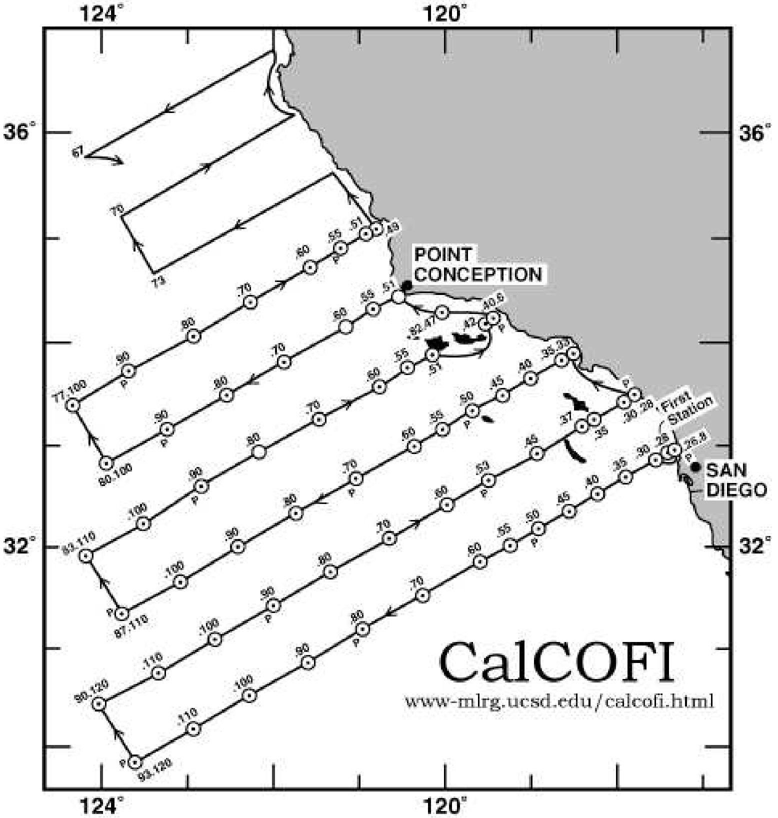

Figure 6: CalCOFI sampling. Circules show current 66 stations. Line indicates schematically the extended CalCOFI sampling used prior to 1985.

Here are three possible strategies for evaluating the large array of CalCOFI data compared with the small array.

a. For the time period between 1949 and 1985, use correlation analysis to compare data from the 66 continuously sampled stations with data from the ∼24 stations that are no longer regularly sampled. You might imagine doing this for all possible combinations of the 24 short-term stations compared with the 66 long stations to produce a total of 24×66 *correlation coefficients. That's a lot to digest, but if you found that the correlations were consistently statistically significant, this would be enough to give you confidence in analyses based on the 66-station record.*

b. To do something slightly more sophisticated, you might imagine least-squares fitting the measurements in each of the 24 short-term stations against the full array of observations from the 66 stations. Thus you could ask how best to represent T_{67} *as a weighted sum of the data from the 66 longer term stations. You might then do 24 separate least-squares fits for the 24 short-term records compared with the first parts of all of the 66 long-term records. Then you'd want to check how much of the overall variance in the short-term stations was captured by your fits. Is it statistically significant?*

c. Finally, you could compute EOF describing the leading modes of variability in the 66 long-term stations. You might also compute EOFs describing the leading variability in the 24 short-term stations or for the total collection of 90 or so stations. Finally you could ask *how well the temporal modes for the first few EOFs were correlated for these cases. If they're significantly correlated, then you might conclude that by measuring the 66 stations you could capture most of the relevant variability in the other 24 stations as well.*

Crib sheet: MAE 127

Moments computed from data

$$\begin{aligned}\overline{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i &(3)\\ \mu_n &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^n &(4)\end{aligned}$$

Moments computed from PDFs

$$\begin{aligned}\langle x\rangle &= \int_{-\infty}^{\infty} x\,P(x)\,dx &(5)\\ \mu_n &= \int_{-\infty}^{\infty}(x-\langle x\rangle)^n\,P(x)\,dx &(6)\end{aligned}$$

Error Propagation

$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial x}\sigma_x\right)^2 + ... + \left(\frac{\partial q}{\partial z}\sigma_z\right)^2} \qquad (7)$$

Correlation

$$\begin{aligned}r &= \frac{\overline{(x-\overline{x})(y-\overline{y})}}{\sqrt{\overline{(x-\overline{x})^2}\,\overline{(y-\overline{y})^2}}} &(8)\\ Pr &= \mathrm{erfc}\left(\frac{|r|\sqrt{N}}{\sqrt{2}}\right) &(9)\\ r_{sig} &= \mathrm{erf}^{-1}(s)\sqrt{\frac{2}{N}} &(10)\end{aligned}$$

Autocorrelation

$$C(\Delta t) = \frac{\overline{(x(t)-\overline{x})(x(t+\Delta t)-\overline{x})}}{\overline{(x-\overline{x})^2}}, \qquad (11)$$

Least-Squares Fitting

$$\begin{aligned}Ax &= b &(12)\\ x &= (A^TA)^{-1}A^Tb &(13)\\ \chi^2 &= \sum_{i=1}^{N}\frac{\left[\left(\sum_{j=1}^{M} A_{ij}x_j\right)-b_i\right]^2}{\sigma_{b_i}^2} &(14)\\ \sigma_x^2 &= \mathrm{diag}((A_w^TA_w)^{-1}) &(15)\end{aligned}$$

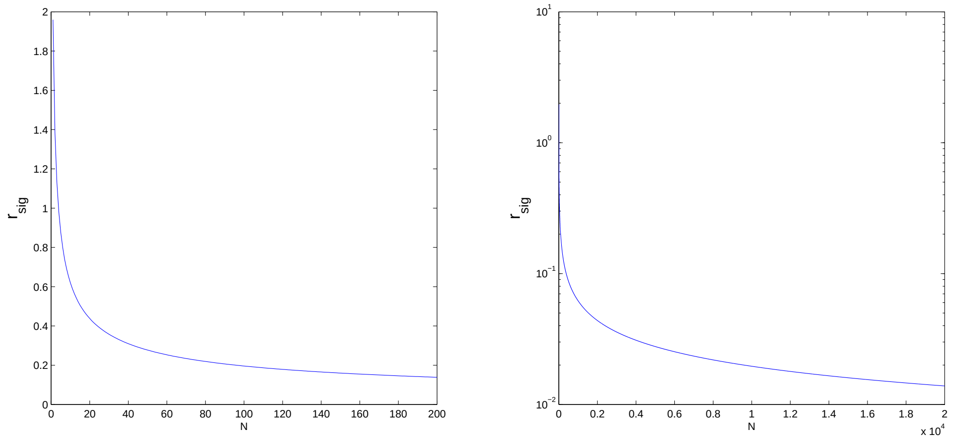

Figure 7: (left) The 95% significance level, r_{sig} plotted as a function of N for small N. (right) Same thing on semi-log scale for a larger range of N.

Fourier Transforms and Spectra

$$
\begin{aligned}
X_k &= \sum_{n=1}^{N} x_n \exp(-i2\pi(k-1)(n-1)/N) &(16)\\
x_n &= \frac{1}{N}\sum_{k=1}^{N} X_k \exp(i2\pi(k-1)(n-1)/N) &(17)\\
S_k &= \frac{\sum_{i=1}^{M} |X_{k,i}|^2}{NM} &(18)\\
\nu &= 2M &(19)\\
\nu/\chi^2_{\nu,\alpha/2} &< S_k/\hat{S}_k < \nu/\chi^2_{\nu,1-\alpha/2} &(20)
\end{aligned}
$$

with appropriate scalings of S_k by 2.

Coherence

$$
\begin{aligned}
C_k &= \frac{\sum_{i=1}^{M} X_{k,i}Y^*_{k,i}}{NM} &(21)\\
\text{coherence} &= \frac{|C_k|}{\sqrt{S_X S_Y}} &(22)\\
\delta_c &= \sqrt{1-\alpha^{1/M-1)}} &(23)\\
\tan(\phi) &= \frac{-\Im(C_k)}{\Re(C_k)} &(24)
\end{aligned}
$$

Empirical Orthogonal Functions

$$[U,S,V] = \text{svd}(A); \tag{25}$$