Problem Set 7: Solutions

1. The covariance of the variable x is of the form:

$$\langle x(t)x(t+\delta t)\rangle = \begin{cases} (1-|t|)\langle x(t)x(t)\rangle & \text{for } |t| \le 1 \text{ day} \\ 0 & \text{otherwise} \end{cases}$$

Assume time is measured in days. If data are sampled continuously for 365 days, how many independent samples do you have?

Solution:

$$N_E = \frac{365}{\int_{-1}^{1}(1-|t|)\langle x(t)x(t)\rangle dt} = \frac{365}{2\langle x^2\rangle \int_0^1 (1-t)dt} = \frac{365}{1} = 365$$

so there are 365 independent samples.

More formally, you can allow for the finite size of the data set by computing

$$N_E = \frac{365}{\int_{-1}^{1}\left(1-\frac{|t|}{365}\right)(1-|t|)\langle x(t)x(t)\rangle dt} = 365.33$$

This slightly increases the apparent number of independent samples.

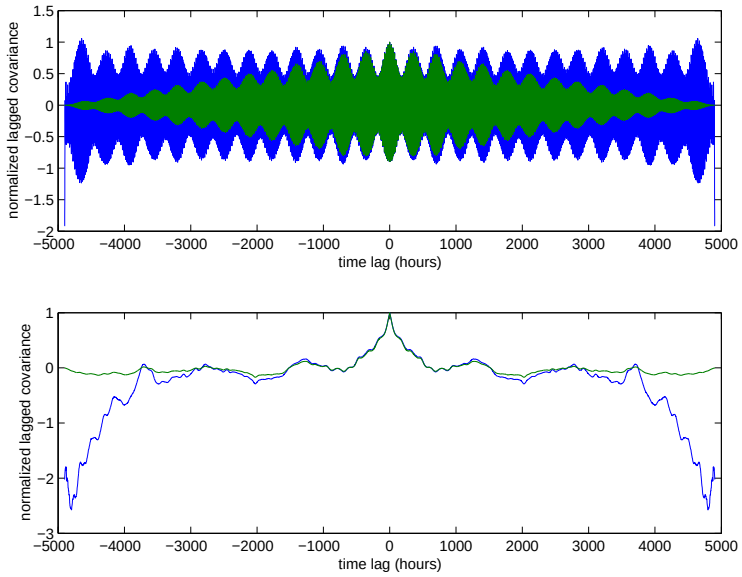
2. Download the Amsterdam Island bottom pressure recorder record (amsterdam_bpr.dat) from http://www-mae.ucsd.edu/∼sgille/sio221b/. In this data record, data are archived hourly. Column 6 contains bottom pressure, and column 8 contains detided bottom pressure.

a. Compute and plot the discrete autocovariance for columns 6 and 8. (You can do this using the "xcov" function in matlab, but be sure to read the help screen for this function, and consider whether you should use the biased or unbiased estimator.)

b. How many independent samples are in the two data records?

c. How does the time interval between independent samples compare with the time required for the autocovariance function to drop by $1/e$ from its maximum? What about the time lag required to reach the first zero crossing of the autocovariance function?

Solution:
a. Biased and unbiased covariances are plotted in the figure. The biased covariance (green) converges to zero for large lags. The unbiased covariance (blue) grows for large lags, because the covariance is computed from few observations. As it turns out, neither is fully satisfactory for computing the number of independent samples. The biased covariance converges to zero for large lags, which is good, but may be artificially small for large lags. In contrast, the unbiased covariance is unbiased, which is good, but may not provide a good estimate of the mean for large lags.



b. There are several possible approaches for finding a robust estimate of the number of degrees of freedom. One is to use the formalism that we derived in class for finite data sets:

$$N_E = \frac{N}{\sum_{-N}^{N} \left(1 - \frac{|n|}{N}\right) \rho(n)}$$

Using the time range between days 904 and 4903, this leads to N_E of 836 for the raw data (column 6) and 170 for the detided data (column 8).

An alternative strategy is to use the expression that we derived for infinitely large samples, but carry out calculations only for the core of the covariance (for lags less than N_l), where the mean is well-defined:

$$N_E = \frac{N}{\sum_{n=-N_l}^{N_l} \rho(n)}.$$

This converges to $N_E \approx 400$ for the raw data for $800 < N_l < 1200$ and to $N_E \approx 30$ for the detided data for $300 < N_l < 500$. These values are smaller than the previous results; since they don't depend on a biased estimator, they may prove more satisfactory.

c. Often researchers use the e-folding scale or the first zero crossing to determine the time interval between independent samples. A quick way to estimate an e-folding scale is to look for the point in time when the covariance drops to e^{-1}. Here, for column 6 data (with tides), the e-folding scale is about 4 hours and the first zero crossing occurs after 5 hours. Thus measurements taken 8 to 10 hours apart should be completely decorrelated, implying that N_E should be at least 500 to 600. The analysis in part b suggested a smaller number. For column 8 (detided data), the e-folding scale is 102 hours and the first zero crossing occurs after 229 hours. $N/229$ implies 17 independent samples, and $N/(102 \cdot 2)$ implies 20 independent samples. These numbers are more conservative than the results from part b,

but are of the same magnitude, suggesting that the e-folding scale or zero crossing can help provide a rough estimate of degrees of freedom.

3. (Notes on statistically optimal linear estimators), problem 5). A numerical model of tides involves the large-scale flow field U, but not the small-scale components u. In this model, the bottom drag $\tau = C_D(U+u)|U+u|$ must be parameterized in terms of U. Consider the approximation $\hat{\tau} = \gamma(U) \cdot U$ in one dimensional flow, with $\langle u \rangle = 0$.

a. Find equations determining the optimal function $\gamma(U)$ for the "beauty principle" of minimizing the mean-square error in tidal dissipation. You may wish to assume that there is a scale-separation between large-scale and small-scale motions, so that $\langle Uu \rangle = 0$.

b. Find γ if u is normally distributed with known variance. In other words, $\langle u^2 \rangle = \mu_2$.

c. Compare the mean-square dissipation error using the answer from (b) with that from the naive modeler's choice $\hat{\tau} = C_D U|U|$.

Solution:
a. First we define a cost function:

$$\begin{aligned} \mathcal{L} &= \langle (\hat{\tau}U - \tau U)^2 \rangle = \langle (C_D(U+u)|U+u|U - \gamma U^3)^2 \rangle \\ &= C_D^2 U^2 \langle (U^2 + 2uU + u^2)^2 \rangle + \langle \gamma^2 U^4 \rangle - 2C_D U^2 \langle \gamma U(U+u)|U+u| \rangle \end{aligned}$$

Minimizing this produces:

$$\frac{\partial \mathcal{L}}{\partial \gamma} = 2\gamma \langle U^4 \rangle - 2C_D \langle U^3(U+u)|U+u| \rangle = 0,$$

so

$$\gamma(U) = C_D \left(\frac{\langle U^2|U+u| \rangle + \langle Uu|U+u| \rangle}{\langle U^2 \rangle} \right).$$

b. If $\langle u^2 \rangle = \mu_2$, then we can rewrite the above expression. If $U + u > 0$, then

$$\gamma(U) = C_D \left(\frac{\langle U^3 \rangle + \langle U\mu_2 \rangle}{\langle U^2 \rangle} \right) = C_D U \left(1 + \frac{\mu_2}{U^2} \right)$$

If $U + u < 0$, then

$$\gamma(U) = C_D \left(\frac{-\langle U^3 \rangle - \langle U\mu_2 \rangle}{\langle U^2 \rangle} \right) = -C_D U \left(1 + \frac{\mu_2}{U^2} \right)$$

Combining these:

$$\gamma(U) = \begin{cases} C_D U \left(1 + \frac{\mu_2}{U^2} \right) & \text{for } U + u > 0 \\ -C_D U \left(1 + \frac{\mu_2}{U^2} \right) & \text{for } U + u < 0 \end{cases}$$

Typically, we might have estimates of μ_2 and U, but no knowledge of u at any point in time, so we'd approximate this as:

$$\gamma(U) = C_D|U|\left(1 + \frac{\mu_2}{U^2}\right).$$

c. Finally we compute the mean square error for the two solutions.

$$\begin{aligned}
\epsilon_\gamma^2 &= \left\langle \left(C_D(U+u)|U+u|U - C_D U^2 \left(1 + \frac{\mu_2}{U^2}\right)|U| \right)^2 \right\rangle \\
&= C_D^2 U^2 \left\langle (U+u)^4 - (U^4 + 2\mu_2 U^2 + \mu_2^2) \right\rangle \\
&= C_D^2 U^2 \left\langle 4\mu_2 U^2 + 2\mu_2^2 \right\rangle
\end{aligned}$$

in the case where $U + u$ and U are both greater than zero or both less than zero. (Results would change if u^2 is substituted for μ_2 in this calculation—using u^2 presupposes that we know something about the details of u.) In contrast, the naive modeler's choice produces an error:

$$\begin{aligned}
\epsilon_{\text{naive}}^2 &= \left\langle \left(C_D(U+u)|U+u|U - C_D U^2|U| \right)^2 \right\rangle \\
&= C_D^2 U^2 \left\langle (U+u)^4 + U^4 - 2(U+u)|U+u|U|U| \right\rangle \\
&= C_D^2 U^2 \left\langle u^4 + 4u^2 U^2 \right\rangle \\
&= C_D^2 U^2 \left\langle 3\mu_2 + 4\mu_2 U^2 \right\rangle
\end{aligned}$$

in the case where $U + u$ and U are of the same sign. This error is larger than the error in the optimal solution.

In the alternate case, when $U + u$ and U are of opposite sign, the error in the naive case $4U^4 + 8U^2\mu_2 + 3\mu_2^2$ will be smaller than the error $\epsilon_\gamma^2 = 4U^4 + 12U^2\mu_2 + 6\mu_2^2$.