Lecture 3:

Reading: Bendat and Piersol, Ch. 3.1-3.4, Ch. 4

Friday: field trip, meet at pier

Recap

In lecture 2, we talked about some probability density functions, the fact that we often assume Gaussianity, and that often geophysical variables aren't really Gaussian. We also noted that if we know the mean and standard deviation for a set of variables, then we determine the mean and standard deviation for a summed variable.

We noted that one of the clever aspects of the pdf is that we can use it to determine an expected value:

$$E(x(k)) = \int_{-\infty}^{\infty} xp(x)\,dx = \mu_x. \tag{1}$$

We can also use this for x^2 or for $(x-\mu_x)^2$.

$$E((x(k)-\mu_k)^2) = \int_{-\infty}^{\infty} (x-\mu_x)^2 p(x)\,dx = \sigma_x^2. \tag{2}$$

We can actually keep going to determine higher *moments of the pdf*. Moments of the pdf are traditionally identified by μ_n, where n represents the order of the moment:

$$\mu_n = \int_{-\infty/}^{\infty} (x-\mu_x)^n p(x)\,dx. \tag{3}$$

The third moment, normalized by the standard deviation cubed is called skewness, and the fourth moment is kurtosis.

$$\text{skewness} = \frac{\mu_3}{\sigma^3} \tag{4}$$

$$\text{kurtosis} = \frac{\mu_4}{\sigma^4} \tag{5}$$

Skewness measures the lobsidedness of the pdf; for a symmetric distribution, such as a Gaussian, skewness should be zero. Kurtosis provides a measure of the peakiness of the pdf; for a Gaussian, the kurtosis is 3. Because the kurtosis of a Gaussian is so well defined, people sometimes plot *excess kurtosis*, which is kurtosis minus three.

Error propagation

The lecture 2 notes finished with a discussion of the standard deviation of summed variables: If $x(k) = \sum_{i=1}^{N} a_i x_i(k)$, then the mean of x is

$$\mu_x = E(x(k)) = E\left[\sum_{i=1}^{N} a_i x_i(k)\right] = \left[\sum_{i=1}^{N} a_i E(x_i(k))\right] = \sum_{i=1}^{N} a_i \mu_i. \tag{6}$$

and

$$\sigma_x^2 = E\left[(x(k)-\mu_x)^2\right] = E\left[\sum_{i=1}^{N} a_i (x_i(k)-\mu_i)\right]^2 = \sum_{i=1}^{N} a_i^2 \sigma_i^2. \tag{7}$$

And we noted that the standard error of the mean is $\sigma/\sqrt{N}$.

The *standard error of the variance* is $\sigma^2\sqrt{2/(N-1)}$.

The results for the standard error of the mean led to one of the basic rules that we use to determine uncertainties for summed quantities. This is usually referred to as error propagation. If we sum a variety of measures together, then the overall uncertainty will be determined by the square root of the sum of the squares:

$$\delta_y = \sqrt{\sum_{i=1}^{N} a_i^2 \delta_i^2}, \tag{8}$$

where here we're using δ_i to represent the a priori uncertainties.

We left off with a quick example with more complicated computed quantites. What if we have to multiply quantities together? Then we simply linearize about the value of interest. So if $y = x^2$, and we have an estimate of the uncertainty in x, δ_x, then we know that locally, near x_o, we can expand in a Taylor series:

$$y(x_o + \Delta x) = y(x_o) + \frac{dy}{dx}\Delta x. \tag{9}$$

This means that I can use my rules for addition to estimate the uncertainty in y:

$$\delta_y(x_o) = \left|\frac{dy(x_o)}{dx}\right| \delta_x = 2x_o\delta_x \tag{10}$$

and you can extend from here. If $y = a_1x + a_2x^2 + a_3x^3$, what is δ_y? When will this estimate of uncertainty break down?

Let's consider the specific cases of turbulent heat fluxes. The sensible heat flux is:

$$Q_s = c_h(T_w - T_a)W, \tag{11}$$

where c_h is a constant, T_w is surface water temperature, T_a is air temperature (e.g. at 2 m elevation), and W is wind speed. (Of course there are some complications: c_h is not really a constant. We can approximate it as: $c_h = \rho_a C_{p,a} C_h$, where $\rho_a \approx 1.2$ kg m^{-3} is density of air, the constant pressure specific heat of air $C_{p,a} \approx 1$ kJ kg^{-1} °C, and $C_h \approx 10^{-3}$.) A typical value for Q_h is -5 W m^{-2}.

The latent heat flux is:

$$Q_L = c_e(q_w - q_a)W, \tag{12}$$

where c_e is a constant, q_w is specific humidity at the water surface, and q_a is specific humidity in air. More completely, we can represent c_e as:

$$c_e = \rho_a L_v C_e, \tag{13}$$

where $L_v = 2264.76$ kJ kg^{-1} is the latent heat of vaporization, and $C_e \approx 1.5 \times 10^{-3}$. A typical value for Q_e is about -20 W m^{-2}.

So suppose we measure T_w, T_a, and W with some uncertainties? What is the uncertainty in Q_s? To compute this, we simply follow our rules:

$$\begin{aligned} \delta(Q_s)^2 &= \left[\frac{\partial Q_s}{\partial T_w}\right]^2 \delta(T_w)^2 + \left[\frac{\partial Q_s}{\partial T_a}\right]^2 \delta(T_a)^2 + \left[\frac{\partial Q_s}{\partial W}\right]^2 \delta(W)^2 & (14) \\ &= c_h^2 W^2 \delta(T_w)^2 + (-1)^2 c_h^2 W^2 \delta(T_a)^2 + c_h^2 (T_w - T_a)^2 \delta(W)^2 & (15) \end{aligned}$$

We could further refine this to take into account the uncertainties in c_h, which might depend on ρ_a, and the other coefficients.

Likewise, the uncertainty in the latent heat flux can be estimated through error propagation, and we can decide how much to build the uncertainties in L_v and C_e into our estimate.

This formulation for error propagation works like a charm. But it's built on a few assumptions, and it behooves us to keep these in mind. Namely, we assume that our perturbations are small enough that it's OK to linearize. And we assume that errors are uncorrelated, so that we can treat each term (T_w, T_a, and W, for example) completely separately. What do we do if these assumptions break down?

The central limit theorem

One of the reasons we like Gaussian distributions is because of the central limit theorem. This says that when we sum variables together, the sum will tend to toward being Gaussian, even if the individual variables are not. And this is plausible, since lots of variables we study are derived quantities and therefore (sort of) Gaussian. Bendat and Piersol discussed summed variables under the heading "central limit theorem", but their discussion doesn't provide a clear demonstration of the central limit theorem, and I'm going to leave the formal derivation for 221B.

So let's test this empirically: If we start with data drawn from a uniform distribution, and sum together multiple values, how quickly do our results converge to Gaussian?

```
b=rand(100000,100)-.5; % define a matrix with 100 sets of random values,
                    % each with 100000 elements
cb=cumsum(b,2); % compute the summation of multiple random variables
% now compute the pdf
clear m1 m2
for i=1:100
 [m1(i,:),m2(i,:)]=hist(cb(:,i),-12:.1:12);
end
%
% plot the first five values
plot(m2(1,:),m1(1:5,:)/100000/.1,'LineWidth',2)
axis([-5 5 0 1])
ylabel('probability density','FontSize',14)
xlabel('random variable','FontSize',14)
legend('N=1','N=2','N=3','N=4','N=5')
```

The results of this calculation (shown in Figure 1 provide visual evidence for fairly rapid convergence for the uniform distribution.

Non-Gaussian distributions

As we noted before, unsummed geophysical variables are often non-Gaussian. We've talked about uniform distributions and double exponentials. Here are some particularly important special cases.

We noted last time that the Rayleigh distribution is a good representation for wind speed, which is necessarily positive. It is defined from the square root sum of two independent Gaussian components squared, $y = \sqrt{x_1^2 + x_2^2}$.

$$p(y) = \frac{y}{\sigma^2} \exp\left[-\frac{y^2}{2\sigma^2}\right]. \tag{16}$$

The more generic form of the Rayleigh distribution is the Weibull distribution (for positive x only):

$$p(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right] \tag{17}$$

If $k = 2$ and $\lambda = \sqrt{2}\sigma$, this is the Rayleigh distribution. If $k = 1$, it represents a one-sided exponential distribution.

The Rayleigh distribution that brings us to the χ^2 distribution. Suppose we define a variable:

$$\chi^2_n = z_1^2 + z_2^2 + z_3^2 + ... + z_n^2. \tag{18}$$

Then χ^2_n is a random chi-square variable with n degrees of freedom (and n is simply the number of independent elements that we sum.) Then we can define a functional form for this:

$$p(\chi^2_n) = \frac{1}{2^{n/2}\Gamma(n/2)} \exp\left(\frac{-\chi^2}{2}\right)(\chi^2)^{(n/2)-1}, \tag{19}$$

where $\Gamma(n/2)$ is the gamma function (and this is a function that you normally access through a look-up table or a function programmed into Matlab, for example). Lots of variables end up looking like χ^2, so we'll use this a lot to assess uncertainties, and for this we'll need the cumulative distribution function.

Cumulative distribution functions

The *cumulative distribution function* $C(x)$ is the probability of observing a value less than x. It can be computed by integrating the pdf.

$$C(x) = \int_{-\infty}^{x} p(x')dx'. \tag{20}$$

$C(x)$ is 0 when x approaches minus infinity, indicating that there's a negligibly small chance of having an infinitely small value of x, and it is 1 when x goes to plus infinity, which says that there is a 100% chance of observing some value. The midpoint, where $C(x) = 0.5$ is the median.

For a Gaussian, the cdf is defined to be an error function. For a chi-squared function, it's defined as

$$C(x) = \frac{1}{\Gamma(n/2)}\gamma(n/2, x/2), \tag{21}$$

where γ is the lower incomplete Gamma function (and like the Gamma function $\Gamma(n/2)$, it is accessed through a look-up table. What is the cdf of a uniform distribution?

Are two pdfs different?

So now let's return to the heart of our problem. How do we tell if two pdfs differ? We've already noted that two data sets can look wildly different but still have the same mean and variance, so clearly we need something more than just the mean and variance. We can go back to our Gaussian overlaid on empirical pdf and eyeball the difference to say that they're close enough, or not plausibly similar. We can evaluate whether the mean and standard deviation differ. All of this is good, but it doesn't exploit the full range of information in the pdf. We need a metric to measure how different two distributions are.

Here are a couple of strategies. One notion is to ask about the largest separation between 2 pdfs. We compute two cdfs—in this case one empirical and one theoretical, but we can also

do this with two empirical cdfs. We find the maximum separation between the distributions, the Komogorov-Smirnov statistic:

$$D_n = \sup_n |C_n(x) - C(x)| \tag{22}$$

and then we can predict the probability that a data set with n elements should differ from the ideal distribution by D_n. Matlab has a "kstest" function (or "kstest2") that sorts through the parameters for this. However, we have to be careful with this, because usually our data are correlated, and we don't have as many degrees of freedom as we think. The easiest solution is to decimate the data set so that the number of elements reflects the number of degrees of freedom.

A second strategy is to bin the data and ask whether the number of data in the bin is consistent with what we'd expect, using a χ^2 statistics. In this case for comparisons with a theoretical pdf,

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i}, \tag{23}$$

where N_i is the observed number of events in bin i, and n_i is the theoretical or expected number of events in bin i. For comparisons between two distributions,

$$\chi^2 = \sum_i \frac{(N_i - M_i)^2}{N_i + M_i}, \tag{24}$$

where N_i and M_i are each observed numbers of events for bin i. The values of χ^2 are evaluated using the χ^2 probability function $Q(\chi^2|\nu)$, which is an incomplete gamma function, where ν is the number of bins (or the number of bins minus one, depending on normalization). In Matlab this is

```
gammainc(chi_squared/2,nu/2)
```

Fitting a function to data: least-squares fitting

Now, we've laid a lot of ground work. Let's think about our time series. If we look at SST records, for example, how can we determine whether temperatures are increasing or decreasing over time. Let's suppose we're looking for a linear trend. Then

$$T = T_o + bt, \tag{25}$$

where T represents our measured temperature data, T_o is a constant (unknown), t is time, and b is the time rate of change. We have lots of observations, so we really should represent this using vectors (which we'll indicate with bold face):

$$\mathbf{T} = T_o + b\mathbf{t}. \tag{26}$$

We'll want to find the best estimates of the scalars T_o and b to match our data. Formally, provided that we have more than two measurements, this is an over-determined system. Of course, we're talking about real data, so we should acknowledge that we have noise, and our equations won't be perfect fits. We could write:

$$\mathbf{T} = T_o + b\mathbf{t} + \mathbf{n}, \tag{27}$$

where $\mathbf{n}$ represents noise and is unknown. Now the system is formally underdetermined. But we won't lose hope. We'll just move forward under the assumption that the noise is small.

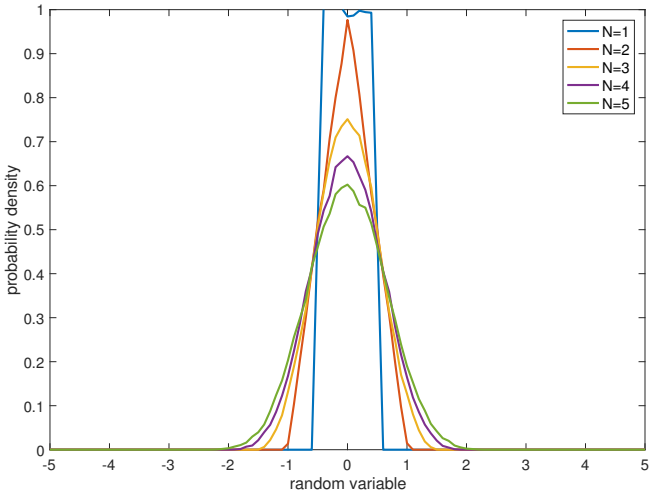

Figure 1: Probability density function for summed data drawn from a uniform distribution. If $N = 1$, so only one data value is used, the distribution is uniform. If $N = 2$, is is a triangle distribution. As N increases, the distribution rapidly evolves to more closely resemble a normal distribution.