Propagating Errors: Scalar Products

Often the variable we want to study is not measured directly. Instead we calculate it from other measured quantities. For example, we measure the electrical conductivity of the ocean C, and from that compute the salinity S. With values of temperature T and salinity S, we compute density ρ. Given the errors in my raw data, how do I determine the error in the variables that I study? What are the uncertainties in my conclusions? *Error propagation* is the formal mechanism for tracking errors.

Here's an example. A simplified form of the equation of state of seawater dictates that:

$$\rho = \rho_o + \alpha(T - T_o) + \beta(S - S_o). \tag{1}$$

where T and S are measured variables, and α, β, ρ_o, T_o, and S_o are constants that we'll assume we know to high accuracy. We can simplify this a step further, by assuming that $S = S_o$ and does not vary, so tha $\rho = \rho_o + \alpha(T - T_o)$. Then given noisy measurements of T with uncertainty δ_T, what is the uncertainty in ρ?

You might guess that the errors in ρ should scale like $\alpha\delta_T$. We can show this formally by using the PDF. If T has a distribution $P_T(T)$, then, the mean density is:

$$\langle\rho\rangle = \int_{-\infty}^{\infty} (\rho_o + \alpha(T - T_o))\, P_T(T)\, dT = \rho_o + \alpha(\langle T\rangle - T_o). \tag{2}$$

and the variance is:

$$\sigma_\rho^2 = \mathrm{var}(\rho) = \int_{-\infty}^{\infty} \left[(\rho_o + \alpha(T - T_o)) - \langle\rho\rangle\right]^2\, P_T(T)\, dT = \int_{-\infty}^{\infty} \left[\alpha(T - \langle T\rangle)\right]^2\, P_T(T)\, dT = \alpha^2\sigma_T^2. \tag{3}$$

Since the standard deviation is the squareroot of the variance, $\sigma_\rho = \alpha\sigma_T$.

Propagating Errors: Addition

Now what happens if our computed quantity depends on more than one random variable? We will again look at the PDF in order to figure out what to expect for the variance. Consider $\rho = \rho_o + \alpha(T - T_o) + \beta(S - S_o)$, where now both T and S are measured, nosiy variables.

$$\begin{aligned}
\sigma_\rho^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[(\rho_o + \alpha(T - T_o) + \beta(S - S_o)) - \langle\rho\rangle\right]^2\, P_T(T)\, dT\, P_S(S)\, dS & (4)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\alpha(T - \langle T\rangle) + \beta(S - \langle S\rangle)\right]^2\, P_T(T)\, dT\, P_S(S)\, dS & (5)\\
&= \alpha^2\sigma_T^2 + \beta^2\sigma_S^2 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 2\alpha\beta(T - \langle T\rangle)(S - \langle S\rangle) P_T(T)\, dT\, P_S(S)\, dS & (6)\\
&= \alpha^2\sigma_T^2 + \beta^2\sigma_S^2 + 2\alpha\beta \int_{-\infty}^{\infty} (T - \langle T\rangle) P_T(T)\, dT \int_{-\infty}^{\infty} (S - \langle S\rangle) P_S(S)\, dS & (7)\\
&= \alpha^2\sigma_T^2 + \beta^2\sigma_S^2 & (8)
\end{aligned}$$

since the integrals of the form $\int_{-\infty}^{\infty}(T - \langle T\rangle)P_T(T)\, dT$ are zero. Thus the squared error is the sum of the squares of each of the summed components.

We can easily extend this to sum an arbitrary number of values. As we showed earlier when we considered the central limit theorem, if $y = 1/N \sum_{i=1}^{N} x_i$, then $\sigma_y = \sigma_x/sqrtN$.

Propagating Errors: General rules

What happens if we are computing a nonlinear quantity that might be a function of several different variables. For example, the full official UNESCO equation of state for density of sea water involves polynomial terms: $\rho = a_0 + a_1T + a_2T^2 + a_3T^3 + a_4T^4 + a_5T^5 + b_0S + b_1TS + b_2T^2S + b_3T^3S + b_4T^4S + c_0S^{3/2} + c_1TS^{3/2} + c_2T^2S^{3/2} + d_0S^2$. How do we deal with uncertainties in T and S here?

Let's consider a simple case: $y = x^2$, where I have a measurement x_o with an associated uncertainty δx_o, as indicated in Figure 1. What's the uncertainty in y? As Figure 1 indicates, for a given small deviation in x, the resulting deviation in y will depend on the local slope of y as a function of x. Since the derivative of y in terms of x tells us the slope, our uncertainty can be calculated as:

$$\delta_y = y(x_o + \delta_x) - y(x_o) = \frac{\partial y(x_o)}{\partial x}\delta_x. \tag{9}$$

(Formally, we can think of this as a Taylor expansion of y about x_o.) Thus our uncertainty is $\frac{\partial y(x_i)}{\partial x}\delta_x$.

What happens for cases such as the density of sea water, where our computed quantity depends on multiple variables? In this case we need to sum the errors for each possible variable. Thus if we want to calculate q as a function of $x, ..., z$, then

$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial x}\sigma_x\right)^2 + ... + \left(\frac{\partial q}{\partial z}\sigma_z\right)^2} \tag{10}$$

So as an example, if $q = xy$, then

$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial x}\sigma_x\right)^2 + \left(\frac{\partial q}{\partial y}\sigma_y\right)^2} = \sqrt{x^2\sigma_y^2 + y^2\sigma_x^2} \tag{11}$$

Error propagation depends on an assumption that the errors are small perturbations from the measured quantities and that errors of independent variables are uncorrelated. The method assumes that we can linearize our errors (by taking only a first derivative) about our measurements. If our measurement errors were large, then the errors estimated for our final results might also be large. We'll look at this more carefully a little later.

Examining Assumptions in Error Estimates

Correlated Errors

Our guidelines for propagating errors assume that errors in each measured variable are uncorrelated. What happens if the errors are correlated? For example, we might find that temperature and salinity measurements made from the same electronics package tend to suffer from similar problems. To deal with correlated errors, ideally we should derive a joint probability density function for all of the variables. We won't explore this in detail in this class.

Non-Gaussian Errors

We tend to assume that our uncertainties have a Gaussian distribution. If that assumption fails, then the uncertainties of our computed quantities may or may not be Gaussian.

Suppose we measure wind velocities, which tend to have a double exponential distribution. (Here we'll do tests using a random number generator to create a 128 by 1000 element matrix with a double exponential distribution.) Then we average the data in groups of 1, 2, 4, or 128 elements. What do we expect? Figure 2 shows PDFs for these 4 cases. First, since we're averaging, the variance should decrease like $\sqrt{N}$. This is the case. The original "data" has a variance of 1. The 2-point averages have a variance of $1/\sqrt{2}$, the 4-point averages have a varince of 1/2, and the 128-point averages have a variance of 0.0894, compared with $1/\sqrt{128} = 0.0884$.

Second, following the central limit theorem, we expect that as we average more and more data, the PDF should become progressively more Gaussian. This effect is particularly obvious if we look at the PDFs on a logarithmic scale as in Figure 3. A double exponential PDF has a distinct triangular shape on a log scale. Even the PDF of two data points averaged together is more rounded than the PDF of the raw data.

Finally, since the data are non-Gaussian, we cannot assume that they'll have all of the same statistical properties as Gaussian data. For example, in a Gaussian distribution, 68% of all observations are within $\pm 1\sigma$

of the mean. For the double exponential 88% are within $\pm 1\sigma$, 97% within $\pm 2\sigma$ and $> 99\%$ within $\pm 3\sigma$. Thus in theory the interpretation of the standard deviation should change as I average progressively more and more data. (At this point, in practice people typically ignore the fact that their data are non-Gaussian and assume that the Gaussian interpretation of a standard deviation applies.)

Nonlinear Effects

Our error propagation procedure is also built on an assumption that errors are essentially linear. How well does this work?

As an example, let's try to estimate the error in $z = \text{atan}(x/y)$, given that x has an uncertainty σ_x and y has an uncertainty σ_y. According to our rule, the uncertainty in z is

$$\sigma_z^2 = \left(\frac{\partial z}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial z}{\partial y}\right)^2 \sigma_y^2 = \frac{x^2 y^2}{(x^2+y^2)^2}\left(\frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2}\right) \tag{12}$$

Algebraically this is a lot to track, but more importantly, when we plot estimated errors in z for values of x and y between 0 and 1, the errors are ill-behaved when x or y is zero. Figure 4 shows that z varies rapidly near the origin, so small errors in x or y can result in big errors in z. Moreover, since z is an angle, the largest possible errors in z are $\pm\pi$.

We might expect our propagation formula to have trouble, so we can check our error propagation procedure using a Monte Carlo method. To do this, we will add small, normally distributed random perturbations to our estimates of x and y. For example, to add noise to x using Matlab we'd say:

```
delta=0.01;
x=(-1:0.005:1)'*ones(1,401);
x_perturbed=x+delta*randn(size(x));
```

(The function "randn" is a random number generator that produces normally distributed random numbers with mean zero and standard deviation 1; the function "rand" produces random numbers that are uniformly distributed between 0 and 1.) We repeat the procedure a bunch of times to get perhaps 100 or 1000 realizations, and for each realization we compute z. Then we can compute statistics on our ensemble of values of z.

When σ_x and σ_y are small, σ_z is estimated without difficulty, as shown in the left panel of Figure 5, which compares results from the Monte Carlo simulation with results from error propagation. But when x and y are small and their uncertainties are large, error propagation gives us an unreliable estimate of σ_z. The errors estimated through error propagation are also problematic near the for variations in y and in particular near the origin, where error propagation vastly over estimates the true error, as shown in Figure 6. Thus, when we have doubts about the results of error propagation (because our calculations are nonlinear and our errors are large, or because our errors are non-Gaussian, or because the algebra required to propagate errors is unwieldy), we can always try a Monte Carlo approach to estimate our uncertainties. Monte Carlo methods depend on assumptions, but sometimes the assumptions are easier to track using a Monte Carlo approach than a conventional statistical error propagation approach. Monte Carlo approaches are also computationally intensive.

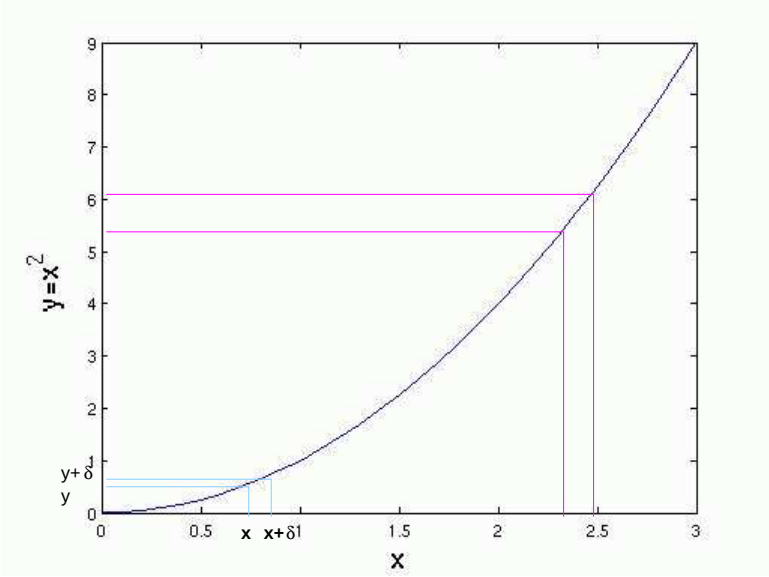

Figure 1: Example of uncertainty estimates for a parabola. Light blue straight vertical lines indicate position of x_o and $x_o + \delta_{x_o}$. Horizontal lines show corresponding difference in $y(x_o)$ and $y(x_o + \delta_{x_o})$. Magenta lines show that if y has a steeper slope, then the difference between $y(x_o)$ and $y(x_o + \delta_{x_o})$ is greater, implying greater uncertainty.

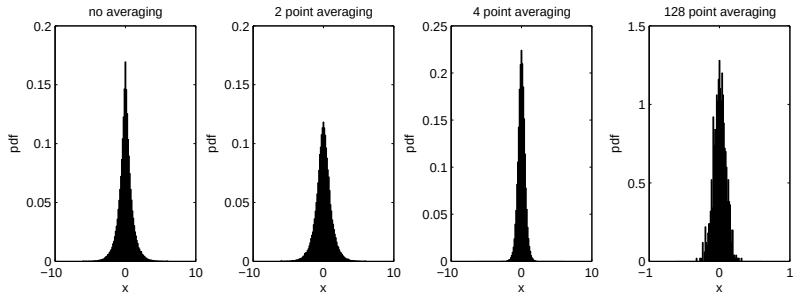

Figure 2: (left) Probability density function for 128,000 random elements with a double exponential distribution and variance 1, (second from left) PDF for same data averaged in pairs, (third from left) PDF for same data averaged in quadruples, (right) PDF for same data averaged in groups of 128 points.

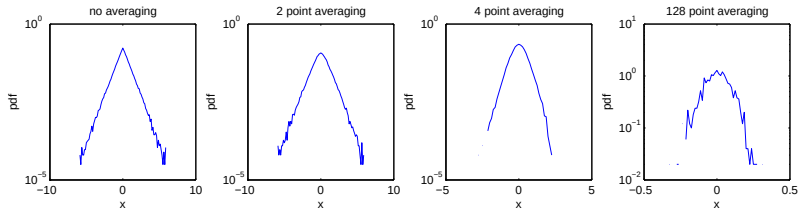

Figure 3: Same as Figure 2 but plotted with log scale for y-axis.

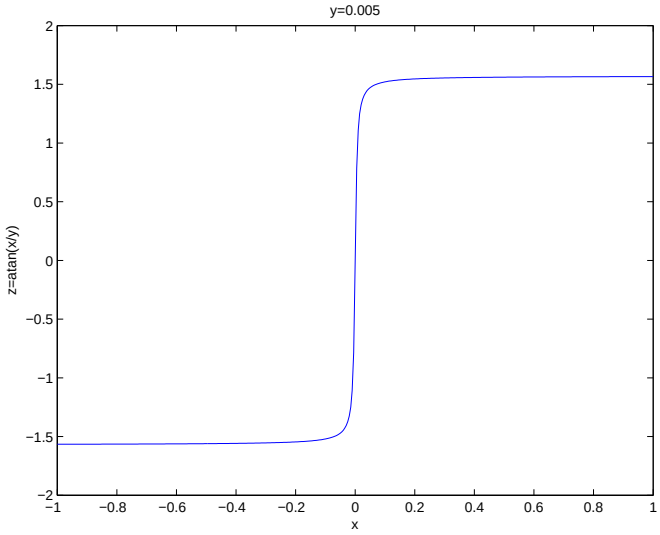

Figure 4: Plot of $z = \text{atan}(x/y)$, with $y = 0.005$.

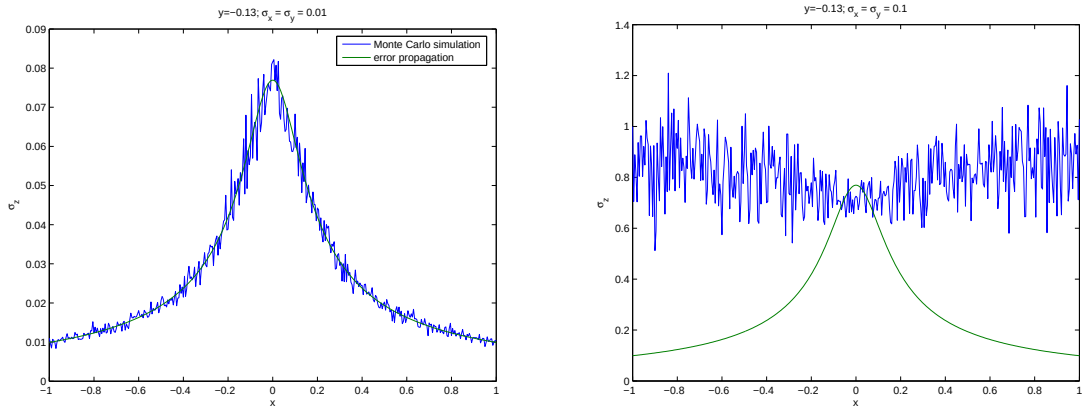

Figure 5: Uncertainties in $z = \text{atan}(x/y)$ as a function of x, with $y = -0.13$, for (left) $\sigma_x = \sigma_y = 0.01$ and (right) $\sigma_x = \sigma_y = 0.1$.

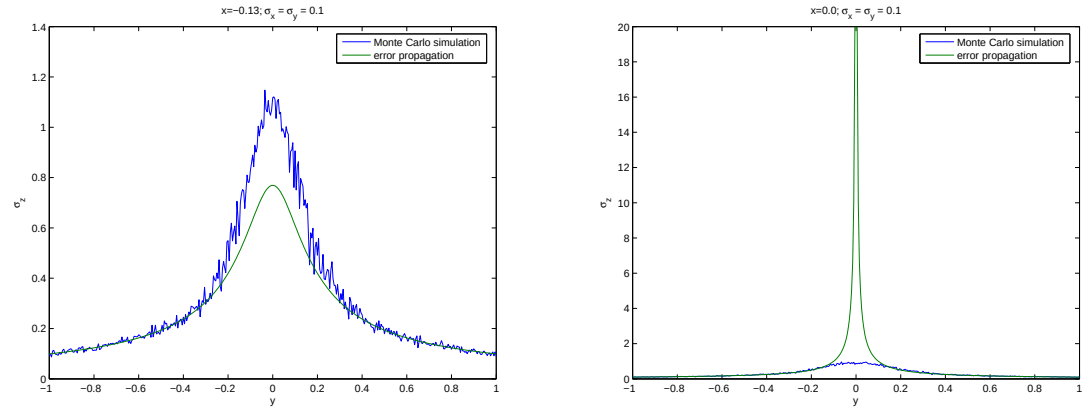

Figure 6: Uncertainties in $z = \text{atan}(x/y)$ as a function of y with $\sigma_x = \sigma_y = 0.1$ (left) at $x = -0.13$ and (right) at $x = 0.0$.