Lecture 18: Linear estimation theory applied to the ocean

Recap

In lecture 16, we explored linear estimation theory, which provides the under pinnings for "objective mapping" (also known as kriging), which is routinely used to map oceanographic variables. Now we're going to delve into some of the details, to look at the practicalities of implementing objective mapping.

First, to recap, we assume that we would like to estimate a variable y from our data vector $\mathbf{x}$. Our estimate of y will be $\hat{y}$

$$\hat{y} = \mathbf{a}^T \mathbf{x} \tag{1}$$

where $\mathbf{a}$ is the gain vector. We found that

$$\mathbf{a} = \langle \mathbf{x}\mathbf{x}^T \rangle^{-1} \langle \mathbf{x}y \rangle, \tag{2}$$

where $\mathbf{x}\mathbf{x}^T$ is the data–data covariance matrix, and $\mathbf{x}y$ is the data–model (or perhaps data–mapped quantity) covariance. The skill is

$$\frac{\langle \hat{y}^2 \rangle}{\langle y^2 \rangle} = \frac{\langle y\mathbf{x}^T \langle \mathbf{x}\mathbf{x}^T \rangle^{-1} \langle \mathbf{x}y \rangle}{\langle y^2 \rangle}, \tag{3}$$

and the MSE is

$$\langle \epsilon^2 \rangle = \langle (\hat{y} - y)^2 \rangle = \langle y^2 \rangle \left(1 - \frac{\langle (\mathbf{a}^T \langle \mathbf{x}y \rangle}{\langle y^2 \rangle} \right) = \langle y^2 \rangle \left(1 - \frac{\langle y\mathbf{x}^T \rangle \langle \mathbf{x}\mathbf{x}^T \rangle \langle \mathbf{x}y \rangle}{\langle y^2 \rangle} \right). \tag{4}$$

In this lecture, we want to explore further complexity in this problem.

Linear estimation theory: Adding noise to the problem

So what do we do about noise? When we initially went through the derivation for linear estimation, we neglected the fact that measurements are intrinsically noisy. Data are noisy for a multitude of reasons:

1. The instrument used to measure a variable has intrinsic errors— instrumental error.

2. Our data may be sparse and not representative of the processes that we want to study—that is, our temperature profile might be in the middle of an eddy, when we think we're trying to map the mean state of the ocean. This is referred to as **representation error**. We need to take into account this intrinsic variability in the system we are measuring.

3. Beyond the challenges of representation error, we might simply have missing physics in our model that will show up as noise—for example, surface wave effects impact the the drag coefficient C_D that links wind speed to wind stress, but they aren't taken into account in many formulations of C_D, which can lead to spread when we try to infer wind stress.

While the process implied in the model $\hat{y} = \alpha x$ may be the object of our study, other processes affecting the measured variables are certainly also occurring. Suppose that linear relationship truly exists between two variables in the form

$$\tilde{y} = \tilde{\alpha}\tilde{x} \tag{5}$$

where the $\sim$ indicates that this is true relationship between the variables. The reality is that our measurements are noisy so that we have access to the variables as follows

$$y = \tilde{y} + y_e = \tilde{\alpha}\tilde{x} + y_e \tag{6}$$
$$x = \tilde{x} + x_e \tag{7}$$

where the e subscript indicates noise. Statistics we calculate from our observations will include contributions from noise. So the gain calculated from this data would be

$$\alpha = \frac{\langle xy \rangle}{\langle x^2 \rangle} \tag{8}$$
$$= \frac{\langle (\tilde{x} + x_e)(\tilde{\alpha}\tilde{x} + y_e) \rangle}{\langle (\tilde{x} + x_e)^2 \rangle} \tag{9}$$
$$= \frac{\tilde{\alpha}\langle \tilde{x}^2 \rangle + \tilde{\alpha}\langle \tilde{x}x_e \rangle + \langle \tilde{x}y_e \rangle + \langle x_e y_e \rangle}{\langle \tilde{x}^2 \rangle + 2\rangle \tilde{x}x_e \rangle + \langle x_e^2 \rangle} \tag{10}$$

Assume that the noise on x and y is uncorrelated with the true variables (the signal) and with each other

$$\langle \tilde{x}x_e \rangle = \langle \tilde{x}y_e \rangle = \langle x_e y_e \rangle = 0 \tag{11}$$

so that

$$\alpha = \tilde{\alpha}\frac{\langle \tilde{x}^2 \rangle}{\langle \tilde{x}^2 \rangle + \langle x_e^2 \rangle} \tag{12}$$

In the limit that the noise variance $\langle x_e^2 \rangle$ is zero, the true gain is recovered. With larger noise, the estimated gain α will always be closer to zero than the true gain $\tilde{\alpha}$. As the noise variance approaches infinity, the gain approaches zero. The reason the optimum gain is reduced in the presence of noise is that amplifying noise degrades the MSE. As we are considering variables with zero mean from here on, the essential notion is that minimum MSE estimates tend to fade toward the mean in the presence of noise.

Since our estimate always will be closer to zero than the true value, this is a reminder that we always want to start from our best prior guess, so that our estimate will be as close as possible to what we already know.

Mapping with two points

Last week we looked at a simplified scenario in which we had two data points and mapped values between them. We initially looked at a problem that matched the code, with data at $t_1 = -1$ and $t_2 = +1$, but let's now make this a little more general by using $t_1 = -\delta$ and $t_2 = +\delta$.

In order to estimate values of $y = x(t)$, we'll need to know covariances. To get started we assume a Gaussian covariance of the form:

$$\langle x(t_1)x(t_2) \rangle = A \exp\left[\frac{-(t_1 - t_2)^2}{T^2}\right] \tag{13}$$

but you'll notice that the covariance has no intrinsic dependence on the time t_1 or t_2, but depends only on their separation $t_2 - t_1 = \Delta t = \tau$.

We can rewrite the covariance as

$$\rho(\tau) = A \exp\left[\frac{-\tau^2}{T^2}\right]. \tag{14}$$

For our two data points, we can write a 2×2 data–data covariance matrix:

$$\langle \mathbf{x}\mathbf{x}^T \rangle = A \begin{bmatrix} 1 & \exp\left[\frac{-4\delta^2}{T^2}\right] \\ \exp\left[\frac{-4\delta^2}{T^2}\right] & 1 \end{bmatrix}. \tag{15}$$

You'll notice that this matrix is square and symmetric, which is critical since we need it to be invertable.

The data–model covariance marix is:

$$\langle y\mathbf{x} \rangle = A \begin{bmatrix} \exp\left[\frac{-(t+\delta)^2}{T^2}\right] \\ \exp\left[\frac{-(t-\delta)^2}{T^2}\right] \end{bmatrix}. \tag{16}$$

As homework you were asked to look closely at this problem. Here I just want to highlight a few key issues. In the appendix, I'll look at the limit as δ and t are small, when this resembles interpolation.

In setting this up, we have made some key simplifications, that we should examine in detail.

Noise matters

The symmetric data–data covariance matrix has an Achilles' heel that we'll need to watch. Suppose that our decorrelation scale T is really long. That would mean that adjacent points are highly correlated. You can readily see that in such a case, the matrix would approach:

$$\langle \mathbf{x}\mathbf{x}^T \rangle = A \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \tag{17}$$

which is singular. What do we do about this?

The key is to recall that our data are noisy for all the reasons that we mentioned above, and in practice, mostly because of representation error. This means that even closely spaced measurements will not be as correlated as any given measurement is with itself. We need to build this measurement uncertainty into the matrix by adding noise. In essence, the covariance $\langle x_1 x_1 \rangle$ needs to exceed the covariance of $x_1 x_2$ by our estimate of the noise variance. We should represnet this noise by adding the noise variance along the diagonal.

$$\langle \mathbf{x}\mathbf{x}^T \rangle = A \begin{bmatrix} 1+\sigma^2 & \exp\left[\frac{-4\delta^2}{T^2}\right] \\ \exp\left[\frac{-4\delta^2}{T^2}\right] & 1+\sigma^2 \end{bmatrix}. \tag{18}$$

Provided our noise is reasonably sized, this will protect us from having a singular matrix, and also provide a good representation of the uncertainty in the system.

In essence, our data covariance should be:

$$\langle \mathbf{x}\mathbf{x}^T \rangle = \langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^T \rangle + \sigma^2 \mathbf{I}, \tag{19}$$

where σ is the prior uncertainty for the measurements. This assumes that the uncertainty is the same everywhere, but we are free to have σ depend on location.

Choosing a decorrelation function

Objective mapping problems are classically laid out using Gaussian covariance functions. However, there's no obligation to specify any given analytic form for the covariance. We could

assume a different analytic form (e.g. a double exponential), an empirical form based on observations, or if y was a complex function of x, we could build the functional relationships between y and x into to covariance—this is done to map dynamic topography from velocity information, for example.

One obvious choice is the double exponential:

$$\rho(\tau) = A \exp\left(-\frac{|\tau|}{T}\right) \tag{20}$$

Dependence on time or space

We set up the covariance to depend only on the separation between t_1 and t_2, but not on the actual values of t_1 and t_2. This is a computationally convenient decision that is appropriate much of the time, but it's not required. It is however, essential that your covariance matrix be symmetric. In other words, I need to require that

$$\langle x(t_1)x(t_2)\rangle = \langle x(t_2)x(t_1)\rangle \tag{21}$$

In the case above, we'd run into trouble if we used $t_1 - t_2$ with an exponential and without an absolute value sign, or if we varied the decorrelation scale T, but had it depend only on the first index. Consider the challenges in having a covariance matrix of the form:

$$\langle \mathbf{x}\mathbf{x}^T\rangle = A \begin{bmatrix} 1+\sigma^2 & \exp\left[\frac{-\tau}{T}\right] \\ \exp\left[\frac{+\tau}{T}\right] & 1+\sigma^2 \end{bmatrix} \tag{22}$$

or

$$\langle \mathbf{x}\mathbf{x}^T\rangle = A \begin{bmatrix} 1+\sigma^2 & \exp\left[\frac{-|\tau|}{T_1}\right] \\ \exp\left[\frac{-|\tau|}{T_2}\right] & 1+\sigma^2 \end{bmatrix}. \tag{23}$$

Neither of these matrices meets the fundamental requirement that the data–data covariance be symmetric, that is that the covariance between $x(t_1)$ and $x(t_2)$ has to be the same as the covariance between $x(t_2)$ and $x(t_1)$.

Mapping in multiple dimensions

So far we've been looking at linear estimation in one dimension only, essentially considering variations on interpolating in time to between $x(t_1)$ and $x(t_2)$ to find $x(t)$. However, our most interesting and challenging problems involve mapping in two-dimensional space, or three-dimensional space, or some combination of space and time. This requires a little thought for the covariance.

Isotropic covariance. An easy scenario is to map in two dimensional space mapping point measurements at the ocean surface onto a regular grid. An example of this might be high-frequency radar data, which are measured continuously from multiple radars along the California and Oregon coast. Since the measurements are continuous, we can take snapshot measurements, all from the same time, to produce mapped, gridded fields. That requires a covariance function in space. We can write:

$$\langle s(x_1,y_1)s(x_2,y_2)\rangle = \rho(r) = A \exp\left[\frac{-r^2}{L^2}\right], \tag{24}$$

where $r = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ is the spatial separation, and L is a length scale. This assumes an isotropic decorrelation—that distance is measured in equivalent units (e.g. km) in both x and y, and that the lengthscales are the same.

Anisotropic covariance. greater level of complexity is introduced if we assume that the covariance. This could occur because features are elongated in the zonal direction so that L_x is longer than L_y, or we consider depth as well as horizontal distance, or we consider time as well as space, so we need to account for variables with entirely different units. In this case, we can write a more refined covariance function:

$$\langle s(x_1,y_1,z_1,t_1)s(x_2,y_2,z_2,t_2)\rangle = \rho(r) = A\exp\left(-\left[\frac{\Delta x^2}{L_x^2}+\frac{\Delta y^2}{L_y^2}+\frac{\Delta z^2}{L_z^2}+\frac{\tau^2}{T^2}\right]\right), \tag{25}$$

and we could even use different functional forms for different variables.

Linear operators

Finally, we should consider a situation in which our mapped quantity is not equivalent to x, but instead depends on some linear operator. We initially considered the estimate of a continuous function of time $\hat{y}(t)$ given discrete data $\mathbf{x}$:

$$\hat{y}(t) = \mathbf{x}^T\langle\mathbf{x}\mathbf{x}^T\rangle^{-1}\langle\mathbf{x}y(t)\rangle. \tag{26}$$

Now apply a linear operator L to the estimate to get the result

$$L\left[\hat{y}(t)\right] = \mathbf{x}^T\langle\mathbf{x}\mathbf{x}^T\rangle^{-1}\langle\mathbf{x}L\left[y(t)\right]\rangle. \tag{27}$$

The linear operator works only on the part of the estimate that is a continuous function. Thus the result of the linear operation on the estimate is the same as the estimate of the linear operation.

For a specific example, consider the estimate of the time derivative discrete time data. In this case, the estimate would be

$$\frac{d\hat{x}(t)}{dt} = \mathbf{x}^T\langle\mathbf{x}\mathbf{x}^T\rangle^{-1}\left\langle\mathbf{x}\frac{dx(t)}{dt}\right\rangle, \tag{28}$$

and the skill would be

$$\frac{\left[\frac{d\hat{x}(t)}{dt}\right]^2}{\left[\frac{dx(t)}{dt}\right]^2} = \frac{\left\langle\frac{dx(t)}{dt}\mathbf{x}^T\right\rangle\langle\mathbf{x}\mathbf{x}^T\rangle^{-1}\left\langle\mathbf{x}\frac{dx(t)}{dt}\right\rangle}{\left[\frac{dx(t)}{dt}\right]^2}. \tag{29}$$

To be even more specific, suppose the autocovariance to be Gaussian as in (14), so the data covariance matrix is given by (15) if there are two data. To calculate the estimate of the derivative and its skill (50-51), we need two more statistics:

$$\text{data–model covariance:} \qquad \left\langle\mathbf{x}\frac{dx(t)}{dt}\right\rangle \tag{30}$$

$$\text{model–model covariance:} \qquad \left\langle\left[\frac{dx(t)}{dt}\right]^2\right\rangle \tag{31}$$

To calculate the covariance of the data with the time derivative, we need the time derivative of the autocovariance (14):

$$\left\langle x(t_i)\frac{dx(t)}{dt}\right\rangle = \frac{d}{dt}\langle x(t_i)x(t)\rangle = -\frac{2A(t-t_i)}{T^2}\exp\left[-\frac{(t-t_i)^2}{T^2}\right]. \tag{32}$$

where the subscript i refers to the data. The variance of the time derivative is

$$\left\langle \left[\frac{dx(t)}{dt}\right]^2 \right\rangle = \left[\frac{\partial^2}{\partial t_1 \partial t_2}\langle x(t_1)x(t_2)\rangle\right] = \frac{2A}{T^2}\exp\left(-\frac{(t_2-t_1)^2}{T^2}\right) - \frac{4A(t_1-t_2)^2}{T^4}. \tag{33}$$

At $t_1 = t_2$ (e.g. along the diagonal), this becomes

$$\left\langle \left[\frac{dx(t)}{dt}\right]^2 \right\rangle_{t_1=t_2} = \frac{2A}{T^2}. \tag{34}$$

This type of framework allows us to map derivatives in a single step (e.g. to map geostrophic velocities from measured sea surface heights, without having to map and then discretize.

The sample code "intgauss.m" provides examples for mapping time derivatives of x as well as x itself. In the case of perfect data, the first difference estimate of time derivative is approached as the time scale T grows. The skill also improves with increasing time scale.

Noise requires an additional term on the diagonal of the data covariance matrix, but all other required covariances remain the same. The estimate of derivative half-way between the data are generally smaller than for perfect data. The skill is strikingly different for noisy data. A maximum in skill is reached for $T = 2\delta$, the separation of the data. There is an apparent advantage to having data spacing that matches the intrinsic time scale of the observed variable. The skill then decreases for increasing T. In the presence of noise, closely separated data (where close is defined relative to T) are not useful for calculating derivatives. That is, any difference between the data is more likely to be caused by noise than a real change in the observed variable.

Appendix: Interpolation in the limit of small data separation

In the two point case, we might ask when linear estimation begins to look line linear interpolation. Linear estimation has all the machinery of a covariance matrix. The result of straight line interpolation for small data separation is quite general regardless of the functional form of the autocovariance, as we now prove. Our estimate of a continuous function can be written as

$$\hat{x}(t) = a_1 x(-\delta) + a_2 x(\delta) \tag{35}$$

where a_1 and a_2 are components of the vector $\mathbf{a}$ in (1). Assuming stationary statistics, the autocorrelation is

$$\rho(\tau) = \frac{\langle x(\tau + t_0)x(t_0)\rangle}{\langle x^2\rangle}. \tag{36}$$

Using (36) in (2), and inserting into (35), the solution is

$$\hat{x}(t) = \frac{\rho(t+\delta) - \rho(2\delta)\rho(t-\delta)}{1-\rho^2(2\delta)}x(-\delta) + \frac{\rho(t-\delta) - \rho(2\delta)\rho(t+\delta)}{1-\rho^2(2\delta)}x(\delta) \tag{37}$$

To make progress, we need to know how the autocovariance behaves at small lag. Start by making the Taylor series expansion of $x(t + t_0)$:

$$x(\tau + t_0) = x(t_0) + \dot{x}(t_0)\tau + \frac{1}{2}\ddot{x}(t_0)\tau^2 + ... \tag{38}$$

where the dots indicate time derivatives. Using (38) the autocovariance is

$$\langle x(\tau + t_0)x(t_0)\rangle = \langle x^2\rangle + \langle x\dot{x}\rangle\tau + \frac{1}{2}\langle x\ddot{x}(t_0)\rangle\tau^2 + ... \tag{39}$$

The assumption of stationarity allows simplification of the second two terms on the righthand side of (42). Consider the covariance of a variable with its time derivative

$$\langle x\dot{x}\rangle = \left\langle x(t_0)\frac{\partial x(t_0)}{\partial t_0}\right\rangle = \frac{1}{2}\frac{\partial}{\partial t_0}\langle x^2(t_0)\rangle = 0. \tag{40}$$

The last equality is a result of stationarity, that the variance is independent of time. Consider the covariance of a variable with its second time derivative

$$\langle x\ddot{x}\rangle = \frac{\partial}{\partial t_0}\langle x(t_0)\dot{x}(t_0)\rangle - \langle \dot{x}(t_0)\dot{x}(t_0)\rangle = -\langle \dot{x}^2\rangle \tag{41}$$

where stationarity is used to get rid of one of the terms. Using (40-41), the covariance (42) becomes

$$\langle x(\tau + t_0)x(t_0)\rangle = \langle x^2\rangle - \frac{1}{2}\langle \dot{x}^2\rangle\tau^2 + ... \tag{42}$$

For small t, the higher order terms are negligible, and the autocorrelation is

$$\rho(\tau) = 1 - \frac{1}{2}\frac{\langle \dot{x}^2\rangle}{\langle x^2\rangle}\tau^2. \tag{43}$$

Substituting (43) into (37) and some simplifying produces

$$\hat{x}(t) = \frac{\delta - t}{2\delta}x(-\delta) + \frac{\delta + t}{2\delta}x(\delta) \tag{44}$$

$$= \frac{x(\delta) + x(-\delta)}{2} + \frac{x(\delta) - x(-\delta)}{2\delta}t \tag{45}$$

which is old-fashioned straight line interpolation.