Lecture 10: Weighted and constrained least squares

Recap

In Lecture 9, we looked closely at weighted least squares problems and uncertainties, noting that the weight matrix $\mathbf{W}$ is a covariance matrix representing the data–data covariance. Weighting is easiest computationally if $\mathbf{W}$ is a diagonal matrix, implying no covariance between data, but a diagonal form is not formally required.

We'll now examine additional ways to constrain least squares problems.

Constrained least squares

To start, let's think back to the OMP problem. For that, we included a mass constraint ($\alpha+\beta+\gamma=1$) in our least-squares fit matrix $\mathbf{G}$. What if we wanted to make that an absolute requirement? We have another strategy we an try. In addition to minimizing the original cost function

$$\epsilon = (\mathbf{Gm}-\mathbf{d})^T(\mathbf{Gm}-\mathbf{d}) \tag{1}$$

we can also impose K constraints of the form

$$\mathbf{Fm}=\mathbf{h} \tag{2}$$

that have to satisfied exactly. For this we need to have $K<M$. One straightforward, but difficult way to deal with this is to solve for K of the model parameters using (2), substitute into (1), and solve for the remaining $M-K$ elements of $\mathbf{m}$.

A much easier way to do the problem is to use the method of Lagrange multipliers. The idea is to introduce a K-vector of unknowns $\boldsymbol{\lambda}$, and minimize the function

$$\mathcal{L} = (\mathbf{Gm}-\mathbf{d})^T(\mathbf{Gm}-\mathbf{d}) + \boldsymbol{\lambda}^T(\mathbf{Fm}-\mathbf{h}). \tag{3}$$

In essence, we have added K unknowns $\boldsymbol{\lambda}$, but we have K additional equations (2). This is guaranteed to work since we are simply adding zero to our original measure of misfit. Differentiating and setting the result to zero,

$$\frac{\partial\mathcal{L}}{\partial\mathbf{m}} = 2\mathbf{G}^T(\mathbf{Gm}-\mathbf{d}) + \mathbf{F}^T\boldsymbol{\lambda} = 0 \tag{4}$$

whhich gives the solution

$$\mathbf{m} = (\mathbf{G}^T\mathbf{G})^{-1}\left(\mathbf{G}^T\mathbf{d} - \frac{1}{2}\mathbf{F}^T\boldsymbol{\lambda}\right) \tag{5}$$

Now we can plug into the constraint (2),

$$\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\left(\mathbf{G}^T\mathbf{d} - \frac{1}{2}\mathbf{F}^T\boldsymbol{\lambda}\right) = \mathbf{h} \tag{6}$$

and solve for the vector of Lagrange multipliers

$$\frac{1}{2}\boldsymbol{\lambda} = \left[\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{F}^T\right]^{-1}\left[\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d} - \mathbf{h}\right]. \tag{7}$$

Substituting into (5), this gives us

$$\mathbf{m} = (\mathbf{G}^T\mathbf{G})^{-1}\left(\mathbf{G}^T\mathbf{d} - \mathbf{F}^T\left[\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{F}^T\right]^{-1}\left[\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d} - \mathbf{h}\right]\right). \tag{8}$$

For the record, this problem is called least squares with linear equality constraints.

Underdetermined systems: The geostrophic reference velocity problem

Now, let's consider the usual problem $\mathbf{Gm} = \mathbf{d}$, but think about what happens when we want to find more unknowns than we have data. Least squares problems are set up for $N > M$. However, it's easy to imagine having more model paramters than data so that $N < M$. (Intrinsically, to take this to an extreme, in any situation when we have noisy data and model misfit, if we want to solve for the noise, we might imagine having as many unknowns (the misfit) as we have data.)

Consider the time-honored oceanographic problem of determining a reference velocity for a geostrophic velocity calculation. Suppose hydrographic stations were made in the form of sections that enclose a volume of water (Figure 1). Define a coordinate system where x is horizontal distance along the section enclosing the control volume, and z is vertical. The section can then be unfolded and visualized in the x–z plane (Figure **??**). Using the thermal wind equation, the geostrophic velocity is

$$v(x,z) = v_0(x) - \frac{g}{f\rho_0}\int_{z_0}^{z}\frac{\partial\rho}{\partial x}(x,z')\,dz' \tag{9}$$

Here v_0 is the reference velocity normal to the section at depth z_0, ρ is density, g is gravitational acceleration, f is the Coriolis parameter, and ρ_0 is reference density. In a classic oceanographic scenario, we measure density as a function of x and z, but we do not know the reference veocity v_0, so our challenge will be to find the best possible estimate of v_0. We also want to impose a set of constraints on the system:

1. The system is in geostrophic balance.
2. Mass is conserved in the closed domain.
3. Water properties are conserved in the closed domain. For example, there is no net heat gain within the box.
4. The system is in steady state.
5. The level of no motion (or alternatively "known motion") is relatively siple, and reference velocities do not oscillate wildly between adjacent stations.

When this question first arose, decades ago, oceanographers questioned whether all of these constraints could be met simultaneously. Work guided by Carl Wunsch and his students and collaborators formulated a framework for finding an optimal solution based on these constraints.

We start by discretize for horizontal location m, and vertical location n. Velocity is then

$$v_{nm} = v_{0m} + v'_{nm} \tag{10}$$

where v_{0m} is the unknown depth-independent reference velocity, and v'_{nm} is the known depth-dependent relative velocity from thermal wind. Conservation of a property C_{nm} in a vertical range Δz_{nm} implies

$$\sum_{m=1}^{M} C_{nm}(v_{0m} + v'_{nm})\Delta z_{nm}\Delta x_m = 0 \tag{11}$$

We suppose there are N such conservation statements for different properties and layers. Then the components of the data vector, model parameter vector, and data kernel matrix are

$$d_n = \sum_{m=1}^{M} C_{nm} v'_{nm} \Delta z_{nm} \Delta x_m \tag{12}$$

$$m_m = v_{0m} \tag{13}$$

$$G_{nm} = C_{nm} \Delta z_{nm} \Delta x_{nm} \tag{14}$$

We can write conservation statements for mass ($C_{nm} = 1$), for temperature or heat, for salt, for O_2, for nutrients and for potential vorticity. We can also require conservation in an arbitrarily large number of potential density layers. Thus you might suppose that you could set this problem up to be formally overdetermined so that you could solve for the reference velocities. Unfortunately, this is easier said than done, for several reasons. First, if you force the problem to have more equations than unknowns, the rows of G might turn out to be highly correlated, perhaps because the different variables are highly correlated (e.g. variations in temperature mirror variations in salinity), or perhaps because adjacent density layers are highly correlated. This will lead to an ill-conditioned matrix that will not invert. You could run into further problems because you might conclude that properties are not completely conserved within density layers and that you need to account for an unknown background diffusivity (or perhaps a vertical advection term) that connects adjacent layers.

In essence, there are usually more stations in a section (and more unknown reference velocities v_o) than sensible conservation statements. We'll start by treating this as an **underdetermined** problem. It might be argued that all oceanographic inverse problems are underdetermined, as we never have enough data to determine the complete state of the ocean.

Underdetermined problem: Solution approaches What do we do? Typically, one chooses to minimize some norm of the solution. A particularly simple choice is to minimize the L_2 norm of $\mathbf{m}$. We then have the constrained least squares problem to minimize:

$$\mathbf{m}^T\mathbf{m} \tag{15}$$

subject to the constraint $\mathbf{Gm} = \mathbf{d}$. Using the method of Lagrange multipliers, we minimize

$$\mathcal{L} = \mathbf{m}^T\mathbf{m} + \boldsymbol{\lambda}^T(\mathbf{Gm} - \mathbf{d}). \tag{16}$$

Following the procedure for a constrained least squares problem, we can find a solution by differentiating (16) with respect to $\mathbf{m}$, and setting the result to zero

$$\frac{\partial \mathcal{L}}{\partial \mathbf{m}} = 2\mathbf{m} + \mathbf{G}^T\boldsymbol{\lambda} = 0. \tag{17}$$

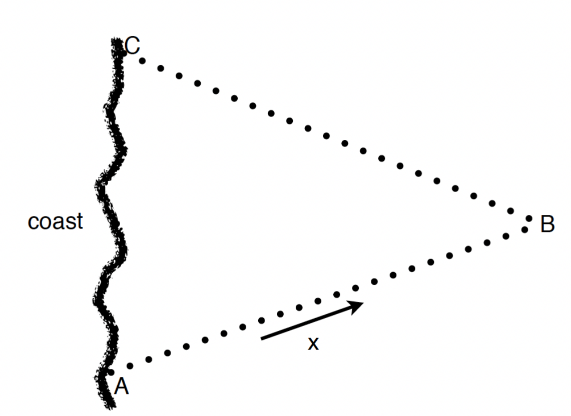

Figure 1: Hydrographic sections extending from the coast and enclosing a volume of water.

Solve for $\mathbf{m}$:

$$\mathbf{m} = -\frac{1}{2}\mathbf{G}^T\boldsymbol{\lambda} \tag{18}$$

Substitute (18) into $\mathbf{Gm} = \mathbf{d}$

$$-\frac{1}{2}\mathbf{G}\mathbf{G}^T\boldsymbol{\lambda} = \mathbf{d} \tag{19}$$

Solve for $\boldsymbol{\lambda}$:

$$-\frac{1}{2}\boldsymbol{\lambda} = (\mathbf{GG^T})^{\mathbf{-1}}\mathbf{d} \tag{20}$$

Substitute (20) into (18) to arrive at the solution

$$\mathbf{m} = \mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{d} \tag{21}$$

This solution requires that the matrix $\mathbf{GG}^T$ be invertible, which is assured if the constraints $\mathbf{Gm} = \mathbf{d}$ are consistent and unique. So the "underdetermined" problem can be thought of as just another constrained least squares problem.

Simultaneous minimization of misfit and model size

The simultaneous minimization of misfit and model size is referred to by a number of different names: **Levenburg-Marquardt stabilization**, **damped least squares**, or **ridge regression**. The name used depends on the field in which the idea was developed, but they all boil down to the notion that it is sometimes a good idea to minimize both misfit and model size at the same time. Consider a formally overdetermined problem, $N > M$, where there are nearly as many model parameters as data. In this case the misfit may be small, which is superficially desirable, but model parameters may be unrealistically large. In the formally underdetermined problem, the exact equality in $\mathbf{Gm} = \mathbf{d}$ may cause a similar problem of large model parameters, and allowing some misfit may be desirable. A way to deal with either of these problems is to minimize a combination of misfit and model size, which in it simplest form may be accomplished by minimizing

$$\mathcal{L} = (\mathbf{Gm} - \mathbf{d})^{\mathcal{T}}(\mathbf{Gm} - \mathbf{d}) + \lambda\mathbf{m}^{\mathcal{T}}\mathbf{m} \tag{22}$$

where λ is an adjustable parameter that varies the relative importance of minimizing the misfit and the model size. We find the solution in the usual way.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{m}} &= 2\mathbf{G}^T(\mathbf{Gm} - \mathbf{d}) + 2\lambda\mathbf{m} = 0 && (23) \\ \mathbf{m} &= (\mathbf{G}^T\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^T\mathbf{d} && (24) \end{aligned}$$

Here $\mathbf{I}$ is the identity matrix (ones along the diagonal, zeros elsewhere). Note that as λ approaches zero, the solution is overdetermined least squares (see above), and as λ approaches infinity, the solution is zero. In practice, one varies λ to achieve a compromise between misfit and model size.

Weighted systems: Accounting for model and data covariances

Adding some weighting to the measures of misfit and model size is often desirable. The essential idea is to weight some of the data, or some of the model parameters, more heavily than others. That is, we may have prior knowledge about the accuracy of the data, or the value of the model parameters.

Suppose some of the data are known more accurately than others. Then an appropriate measure of misfit might be

$$\epsilon = (\mathbf{Gm} - \mathbf{d})^T\mathbf{W}_e(\mathbf{Gm} - \mathbf{d}) \tag{25}$$

where $\mathbf{W}_e$ is a weight matrix with diagonal elements σ_i^{-2}, the inverse variance of each datum.

In general, errors in the data may be correlated, and $\mathbf{W}_e$ might not be diagonal. A reasonable choice for $\mathbf{W}_e$ might then be the inverse of the data–data covariance matrix. Writing the data as a mean plus a fluctuation

$$\mathbf{d} = \langle\mathbf{d}\rangle + \mathbf{d}', \tag{26}$$

the data–data covariance matrix is $\langle\mathbf{d}'\mathbf{d}'^2\rangle^{-1}$, and $\mathbf{W}_e$ would be

$$\mathbf{W}_e = \langle\mathbf{d}'\mathbf{d}'^T\rangle. \tag{27}$$

Now let's turn our attention to the model parameters. Suppose that minimizing the size of the model using $\mathbf{m}^T\mathbf{m}$ is not desirable. A general measure of the model size may be written as

$$\gamma = (\mathbf{m} - \mathbf{m}_0)^T\mathbf{W}_m(\mathbf{m} - \mathbf{m}_0), \tag{28}$$

where $\mathbf{m}_0$ expresses prior knowledge of the solution, and $\mathbf{W}_m$ allows weighting. The matrix $\mathbf{W}_m$ represents the covariance of the model solution and could be constructed to constrain the size of $\mathbf{m}$ or alternatively to minimize some other quantity, such as the curvature of $\mathbf{m}$, for example.

The general problem of simultaneously minimizing misfit (25) and model size (28) involves minimizing the cost function

$$\begin{aligned} \mathcal{L} &= \epsilon + \lambda\gamma && (29) \\ &= (\mathbf{Gm} - \mathbf{d})^T\mathbf{W}_e(\mathbf{Gm} - \mathbf{d}) + \lambda(\mathbf{m} - \mathbf{m}_0)^T\mathbf{W}_m(\mathbf{m} - \mathbf{m}_0). && (30) \end{aligned}$$

We find the solution for this in the usual way, by minimizing $\partial\mathcal{L}/\partial\mathbf{m}$. This is most easily done by defining:

$$\mathbf{m}' = \mathbf{m} - \mathbf{m}_0 \tag{31}$$

so that

$$\mathcal{L} = (\mathbf{Gm}' + \mathbf{Gm}_0 - \mathbf{d})^T \mathbf{W}_e (\mathbf{Gm}' + \mathbf{Gm}_0 - \mathbf{d}) + \lambda \mathbf{m}'^T \mathbf{W}_m \mathbf{m}'. \tag{32}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{m}'} &= 2\mathbf{G}^T\mathbf{W}_e(\mathbf{Gm}' + \mathbf{Gm}_0 - \mathbf{d}) + 2\lambda \mathbf{W}_m \mathbf{m}' && (33)\\ &= 2(\mathbf{G}^T\mathbf{W}_e\mathbf{G} + \lambda\mathbf{W}_m)\mathbf{m}' - 2\mathbf{G}^T\mathbf{W}_e(\mathbf{d} - \mathbf{Gm}_0) && (34)\\ &= 0. && (35)\end{aligned}$$

This implies that

$$\begin{aligned}\mathbf{m}' &= (\mathbf{G}^T\mathbf{W}_e\mathbf{G} + \lambda\mathbf{W}_m)^{-1}\mathbf{G}^T\mathbf{W}_e(\mathbf{d} - \mathbf{Gm}_0) && (36)\\ \mathbf{m} &= \mathbf{m}_0 + (\mathbf{G}^T\mathbf{W}_e\mathbf{G} + \lambda\mathbf{W}_m)^{-1}\mathbf{G}^T\mathbf{W}_e(\mathbf{d} - \mathbf{Gm}_0). && (37)\end{aligned}$$

You can think of $\mathbf{m}_0$ as a prior guess that is perturbed or updated through the weighted least-squares fitting process. This weighted solution can provide an iterative update to a previous solution, and you can think of it as a (slightly simplified) representation of what a data assimilation procedure does when it takes in data to update a state estimate.

Since $\mathbf{W}_e$ and $\mathbf{W}_m$ are inverses of covariance matrices, sometimes it's easier to work with the covariance matrices. In some publications, $\mathbf{W}_e^{-1} = \mathbf{R}$, which is the data-data covariance, representing the noise in the data. And $\lambda\mathbf{W}_m^{-1} = \mathbf{P}$ is the model-model covariance. The ratio between $\mathbf{P}$ and $\mathbf{R}$ provides a measure of signal to noise. In this terminology,

$$\mathbf{m} = \mathbf{m}_0 + (\mathbf{G}^T\mathbf{R}^{-1}\mathbf{G} + \mathbf{P}^{-1})^{-1}\mathbf{G}^T\mathbf{R}^{-1}(\mathbf{d} - \mathbf{Gm}_0). \tag{38}$$

In these solutions, as $\lambda \to 0$, the covariance of the solution is allowed to be large, and no model solutions are imposed, so that

$$\mathbf{m} = (\mathbf{G}^T\mathbf{W}_e\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}_e(\mathbf{d}), \tag{39}$$

which is the weighted least squares solutions.

Alternatively, as $\lambda \to \infty$, we find:

$$\mathbf{m} = \mathbf{m}_0. \tag{40}$$

In summary, it is possible to minimize any combination of criteria, and satisfy any number of constraints. The real challenge is conceptual rather than technical. The real value of these calculations is based on the science expressed in the minimization criteria and the constraints. In class, we can talk all we want about techniques, but coming up with a sensible model will depend on the scientific problem at hand.

Examples

In class, we considered several specific examples.

1. *For a constrained problem, the model parameters* $\mathbf{m}$ *(or* $\mathbf{x}$*) need to be as close as possible to a prior guess that is not zero. What is our cost function?*
   If we want a cost function that minimizes the model parameters, then we need to include a term that forces $(\mathbf{m} - \mathbf{m}_0)^T(\mathbf{m} - \mathbf{m}_0)$ to be small:

$$\begin{aligned}\mathcal{L} &= \epsilon + \lambda\gamma && (41)\\ &= (\mathbf{Gm} - \mathbf{d})^T\mathbf{W}_e(\mathbf{Gm} - \mathbf{d}) + \lambda(\mathbf{m} - \mathbf{m}_0)^T\mathbf{W}_m(\mathbf{m} - \mathbf{m}_0). && (42)\end{aligned}$$

2. *In our final solution, we fit an annual cycle and a diurnal cycle, but we expect them to have different amplitudes, so different covariances. How do we represent that?*
   If we are fitting for model parameters $\mathbf{m}$ to represent an annual cycle (i.e. coefficients of $\cos(2\pi t/(365.25 \text{ days})$ and $\sin(2\pi t/(365.25 \text{ days}))$ and a diurnal cycle (i.e. $\cos(2\pi t/(1 \text{ day})$ and $\sin(2\pi t/(1 \text{ day}))$, then we'll want to include different a priori covariances for these parameters in the weight matrix $\mathbf{W}_m$.

Kachelein, L., B. D. Cornuelle, S. T. Gille, and M. R. Mazloff, 2022. Harmonic analysis of nonstationary tides with red noise using the red_tide package, *J. Atmos. Ocean. Tech.*, **39**, 1031-1051, doi:10.1175/JTECH-D-21-0034.1.