Lecture 11: Weighted and constrained least squares in the real world

Recap

In Lecture 10, we went through an inventory of least-squares methods, including talking about determining uncertainties of fitted model parameters $\mathbf{m}$, considering constrained least squares problems (adding a constraing of the form $\mathbf{Fm} = \mathbf{H}$, underdetermined problems (when $N < M$, and the solution takes the form $\mathbf{m} = \mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{d}$), and simultaneous minimization of misfit and model size (with a solution of the form $\mathbf{m} = (\mathbf{G}^T\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^T\mathbf{d}$).

Now, we'll consider a final case, in which we formally set the covariance of both $\mathbf{d}$ and $\mathbf{m}$, and then we'll examine a few examples.

To get started in class, we looked at a weighted and constrained tidal solution from a paper by Kachelein et al (2022), in which imposing prior knowledge on the spectral structure of bottom pressure data modestly influences the inferred tidal amplitudes and their uncertainties. All of this required that we understand what to do with the covariance matrices for $\mathbf{d}$ and $\mathbf{m}$.

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

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

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

$$\mathbf{d} = \langle\mathbf{d}\rangle + \mathbf{d}', \tag{2}$$

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

$$\mathbf{W}_e = \langle\mathbf{d}'\mathbf{d}'^T\rangle. \tag{3}$$

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

$$\gamma = (\mathbf{m} - \mathbf{m}_0)^T\mathbf{W}_m(\mathbf{m} - \mathbf{m}_0), \tag{4}$$

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 (1) and model size (4) involves minimizing the cost function

$$\begin{aligned}\mathcal{L} &= \epsilon + \lambda\gamma & (5)\\ &= (\mathbf{Gm} - \mathbf{d})^T\mathbf{W}_e(\mathbf{Gm} - \mathbf{d}) + \lambda(\mathbf{m} - \mathbf{m}_0)^T\mathbf{W}_m(\mathbf{m} - \mathbf{m}_0). & (6)\end{aligned}$$

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:

$$\mathbf{m}' = \mathbf{m} - \mathbf{m}_0 \tag{7}$$

so that

$$\mathcal{L} = (\mathbf{G}\mathbf{m}' + \mathbf{G}\mathbf{m}_{,} - \mathbf{d})^{\mathcal{T}}\mathbf{W}_{\rceil}(\mathbf{G}\mathbf{m}' + \mathbf{G}\mathbf{m}_{,} - \mathbf{d}) + \lambda\mathbf{m}'^{\mathcal{T}}\mathbf{W}_{\Updownarrow}\mathbf{m}'. \tag{8}$$

and

$$\frac{\partial\mathcal{L}}{\partial\mathbf{m}'} = 2\mathbf{G}^T\mathbf{W}_e(\mathbf{G}\mathbf{m}' + \mathbf{G}\mathbf{m}_0 - \mathbf{d}) + 2\lambda\mathbf{W}_m\mathbf{m} \tag{9}$$

$$= 2(\mathbf{G}^T\mathbf{W}_e\mathbf{G} + \lambda\mathbf{W}_m)\mathbf{m}' - 2\mathbf{G}^T\mathbf{W}_e(\mathbf{d} - \mathbf{G}\mathbf{m}_0) \tag{10}$$

$$= 0. \tag{11}$$

This implies that

$$\mathbf{m}' = (2(\mathbf{G}^T\mathbf{W}_e\mathbf{G} + \lambda\mathbf{W}_m))^{-1}\mathbf{G}^T\mathbf{W}_e(\mathbf{d} - \mathbf{G}\mathbf{m}_0) \tag{12}$$

$$\mathbf{m} = \mathbf{m}_0 + (\mathbf{G}^T\mathbf{W}_e\mathbf{G} + \lambda\mathbf{W}_m)^{-1}\mathbf{G}^T\mathbf{W}_e(\mathbf{d} - \mathbf{G}\mathbf{m}_0). \tag{13}$$

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,

$$\mathbf{m} = \mathbf{m}_0 + (2(\mathbf{G}^T\mathbf{R}^{-1}\mathbf{G} + \mathbf{P}^{-1}))^{-1}\mathbf{G}^T\mathbf{R}^{-1}(\mathbf{d} - \mathbf{G}\mathbf{m}_0). \tag{14}$$

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

$$\mathbf{m} = (\mathbf{G}^T\mathbf{W}_e\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}_e(\mathbf{d}), \tag{15}$$

which is the weighted least squares solutions.

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

$$\mathbf{m} = \mathbf{m}_0. \tag{16}$$

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 $\mathbf{m}^T\mathbf{m}$ to be small:

$$\mathcal{L} = \epsilon + \lambda\gamma \tag{17}$$

$$= (\mathbf{G}\mathbf{m} - \mathbf{d})^T\mathbf{W}_e(\mathbf{G}\mathbf{m} - \mathbf{d}) + \lambda\mathbf{m}^T\mathbf{W}_m\mathbf{m}. \tag{18}$$

2. *In an annual record, data* ($\mathbf{y}$) *collected in summer are more accurate than data collected in winter. How do we represent that?*
   To account for varying accuracy in our data, we'll want to adjust our weight matrix $\mathbf{W}_e$ to have differing values of σ_i^{-2} on the diagonals. Smaller σ_i in summer imply larger weights. Note that overweighting summer values could mess up our estimates of the annual mean and annual cycle, so we would need to scrutinize our results fairly carefully.

3. *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.