Lecture 10: Constrained least squares

Recap

In Lecture 9, we looked at some specific examples of least-squares fitting, specifically focused on setting up inversion problems. In this lecture, we'll expand our repertoire by considering additional constraints, starting by looking at the linear regression case when our dependent variable (e.g. time or position) has uncertainties.

Linear regression with uncertain variables

In class we examined results using "ordinary least squares" compared with results based on the "standard major axis" approach (see Bellacicco et al, 2019). In a classic least-squares problem we define a model:

$$\mathbf{y} = \mathbf{G}\mathbf{m} + \mathbf{n} \tag{1}$$

where

$$\hat{\mathbf{y}} = \mathbf{G}\mathbf{m} \tag{2}$$

where $\mathbf{n}$ is the vector of the noise or misfit, which we aim to minimize. We minimize the cost function:

$$\epsilon = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{3}$$

to obtain the standard least-squares solution. For a linear regression that finds a constant and a slope, of the form $\hat{\mathbf{y}} = m_1 + m_2\mathbf{x}$, and the matrix $\mathbf{G}$ is:

$$\mathbf{G} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}. \tag{4}$$

The standard least-squares solution gives us

$$\mathbf{m} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{y}, \tag{5}$$

which gives

$$m_1 = \langle \mathbf{y} \rangle - m_2 \langle \mathbf{x} \rangle \tag{6}$$

$$m_2 = \frac{C_{xy}}{C_{xx}}, \tag{7}$$

where C_{xy} is the covariance of x and y, in the form $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$.

The standard major axis method assumes that both x and y have uncertainties so minimizes the area of a triangle between the data point y and the line defining $\hat{y}$. The produces:

$$m_1 = \langle \mathbf{y} \rangle - m_{2_{OLS}} \langle \mathbf{x} \rangle \tag{8}$$

$$m_2 = \sqrt{\frac{C_{yy}}{C_{xx}}} = \pm \frac{s_y}{s_x}, \tag{9}$$

where s_y is the estimated standard deviation of y and x_s is the estimated standard deviation in x. Check the appendix of Bellacicco et al (2019) for details.

Uncertainties in model parameters

All of the above discussion is a temporary digression. In general, when we compute a least squares fit, we probably want to know uncertainties for our fitted parameters m_i. If our data have a known covariance, we can define a weight matrix:

$$\mathbf{W} = \langle \mathbf{d}\mathbf{d}^T \rangle = \sigma^2 \mathbf{I}. \tag{10}$$

If we weight each line of our matrix equation by the uncertainty in the data, we have

$$\mathbf{d}\mathbf{W}^{-1/2} = \mathbf{W}^{-1/2}\mathbf{G}\mathbf{m} \tag{11}$$

As we noted before, this yields

$$\mathbf{m} = (\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}^{-1}\mathbf{d}. \tag{12}$$

We can also estimate the covariance of $\mathbf{m}$:

$$\begin{aligned}
\langle \mathbf{m}\mathbf{m}^T \rangle &= \langle (\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}^{-1}\mathbf{d}((\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}^{-1}\mathbf{d}\mathbf{d})^T \rangle & (13)\\
&= (\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}^{-1}\langle \mathbf{d}\mathbf{d}^T \rangle \mathbf{W}^{-1}\mathbf{G}(\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1^T} & (14)\\
&= (\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}^{-1}\mathbf{W}\mathbf{W}^{-1}\mathbf{G}(\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1^T} & (15)\\
&= (\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})(\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1^T} & (16)\\
&= (\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})(\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1^T} & (17)\\
&= (\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}. & (18)
\end{aligned}$$

This is conveniently just the matrix that we were inverting, and it tells us that the inverted matrix, weighted appropriately by the uncertainties, will provide the uncertainties in $\mathbf{m}$.

Let's test this out in the simplest possible case. Consider the case where our inversion is simply used to find the mean of $\mathbf{d}$. We define an $N \times 1$ matrix:

$$\mathbf{G} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{19}$$

and

$$\mathbf{W} = \sigma^2 \mathbf{I} \tag{20}$$

The standard least-squares solution gives us

$$\begin{aligned}
\mathbf{m} &= (\mathbf{G}^T\mathbf{W}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}^{-1}\mathbf{d} & (21)\\
&= \left(\begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix} \sigma^{-2}\mathbf{I} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix} \sigma^{-2}\mathbf{I}\mathbf{d} & (22)\\
&= \left(\sum_{i=1}^{N} \frac{1}{\sigma^2} \right)^{-1} \sum_{i=1}^{N} \frac{d_i}{\sigma^2} & (23)\\
&= \overline{\mathbf{d}}, & (24)
\end{aligned}$$

where the overbar indicates the mean. The uncertainty in this estimate comes from the inverted matrix:

$$\begin{aligned}
\langle \mathbf{m}\mathbf{m}^T \rangle &= \left(\sum_{i=1}^{N} \frac{1}{\sigma^2} \right)^{-1} && (25)\\
&= \left(\frac{N}{\sigma^2} \right) && (26)\\
&= \sigma^2 N. && (27)
\end{aligned}$$

This wonderfully shows us that the uncertainty in our estimate m_1 of the mean is the standard error of the mean, $\sigma/\sqrt{N}$.

Constrained least squares

Now, 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

$$\epsilon = (\mathbf{G}\mathbf{m} - \mathbf{d})^T(\mathbf{G}\mathbf{m} - \mathbf{d}) \tag{28}$$

we can also impose K constraints of the form

$$\mathbf{F}\mathbf{m} = \mathbf{h} \tag{29}$$

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 (29), substitute into (28), 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

$$\mathcal{L} = (\mathbf{G}\mathbf{m} - \mathbf{d})^T(\mathbf{G}\mathbf{m} - \mathbf{d}) + \boldsymbol{\lambda}^T(\mathbf{F}\mathbf{m} - \mathbf{h}). \tag{30}$$

In essence, we have added K unknowns $\boldsymbol{\lambda}$, but we have K additional equations (29). This is guaranteed to work since we are simply adding zero to our original measure of misfit. Differentiating and setting the result to zero,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{m}} = 2\mathbf{G}^T(\mathbf{G}\mathbf{m} - \mathbf{d}) + \mathbf{F}^T\boldsymbol{\lambda} = 0 \tag{31}$$

whhich gives the solution

$$\mathbf{m} = (\mathbf{G}^T\mathbf{G})^{-1}\left(\mathbf{G}^T\mathbf{d} - \frac{1}{2}\mathbf{F}^T\boldsymbol{\lambda}\right) \tag{32}$$

Now we can plug into the constraint (29),

$$\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\left(\mathbf{G}^T\mathbf{d} - \frac{1}{2}\mathbf{F}^T\boldsymbol{\lambda}\right) = \mathbf{h} \tag{33}$$

and solve for the vector of Lagrange multipliers

$$\frac{1}{2}\boldsymbol{\lambda} = \left[\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{F}^T\right]^{-1}\left[\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d} - \mathbf{h}\right]. \tag{34}$$

Substituting into (32), this gives us

$$\mathbf{m} = (\mathbf{G}^T\mathbf{G})^{-1}\left(\mathbf{G}^T\mathbf{d} - \mathbf{F}^T\left[\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{F}^T\right]^{-1}\left[\mathbf{F}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d} - \mathbf{h}\right]\right). \tag{35}$$

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

Underdetermined 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.)

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:

$$\mathbf{m}^T\mathbf{m} \tag{36}$$

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

$$\mathcal{L} = \mathbf{m}^T\mathbf{m} + \boldsymbol{\lambda}^T(\mathbf{Gm} - \mathbf{d}). \tag{37}$$

Following the procedure for a constrained least squares problem, we can find a solution by differentiating (37) with respect to $\mathbf{m}$, and setting the result to zero

$$\frac{\partial \mathcal{L}}{\partial \mathbf{m}} = 2\mathbf{m} + \mathbf{G}^T\boldsymbol{\lambda} = 0. \tag{38}$$

Solve for $\mathbf{m}$:

$$\mathbf{m} = -\frac{1}{2}\mathbf{G}^T\boldsymbol{\lambda} \tag{39}$$

Substitute (39) into $\mathbf{Gm} = \mathbf{d}$

$$-\frac{1}{2}\mathbf{G}\mathbf{G}^T\boldsymbol{\lambda} = \mathbf{d} \tag{40}$$

Solve for $\boldsymbol{\lambda}$:

$$-\frac{1}{2}\boldsymbol{\lambda} = (\mathbf{G}\mathbf{G}^{\mathbf{T}})^{-\mathbf{1}}\mathbf{d} \tag{41}$$

Substitute (41) into (39) to arrive at the solution

$$\mathbf{m} = \mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{d} \tag{42}$$

This solution requires that the matrix $\mathbf{G}\mathbf{G}^T$ 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

$$\mathcal{L} = (\mathbf{Gm} - \mathbf{d})^{\mathcal{T}}(\mathbf{Gm} - \mathbf{d}) + \lambda \mathbf{m}^{\mathcal{T}}\mathbf{m} \tag{43}$$

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.

$$\begin{aligned} \mathcal{L} &= 2\mathbf{G}^{T}(\mathbf{Gm} - \mathbf{d}) + 2\lambda\mathbf{m} = 0 && (44) \\ \mathbf{m} &= (\mathbf{G}^{T}\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^{T})\mathbf{d} && (45) \end{aligned}$$

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.

Bellacicco, Marco, Vincenzo Vellucci, Michele Scardi, Marie Barbieux, Salvatore Marullo, and Fabrizio D'Ortenzio. 2019. Quantifying the Impact of Linear Regression Model in Deriving Bio-Optical Relationships: The Implications on Ocean Carbon Estimations, *Sensors* 19, no. 13: 3032. https://doi.org/10.3390/s19133032