Fitting a Linear Model with Priors

Samuel D. McDermott

February 14, 2024

Let's imagine you have some data y_j which you want to explain as a function of some independent parameters x_{aj}, where $j \in 1, 2 \dots N_d$ and $a \in 1, 2 \dots N_\ell$ for N_d data points and N_ℓ model parameters. Let's fit a model of the observations that assumes the data are linearly determined by the parameters, $f_j = \theta_a x_{aj}$. I will solve this closely following `astro-ph/0310577`, but adding a Gaussian prior on the θ_a. This essentially reproduces results in the Wikipedia entry on ridge regression with Tikhonov regularization but with somewhat more physical/Bayesian intuition and adding a derivation of the error bars on the estimators.

For variables with prior values ϑ_a, a general covariance matrix $\mathcal{C}_{jk}$ that describes the known and/or modeled correlations between observations and/or parameters, and a model covariance matrix $\mathcal{P}_{ab}$ that describes correlations between the variables θ_a, the loss function is

$$\mathcal{L} = \sum_{j=1}^{N_d} \sum_{i=1}^{N_d} \left(y_j - \sum_{a=1}^{N_\ell} \theta_a x_{aj} \right) \mathcal{C}_{ij}^{-1} \left(y_i - \sum_{b=1}^{N_\ell} \theta_b x_{bi} \right) + \sum_{a=1}^{N_\ell} \sum_{b=1}^{N_\ell} (\theta_a - \vartheta_a) \mathcal{P}_{ab}^{-1} (\theta_b - \vartheta_b), \tag{1}$$

where I choose *not* to use Einstein summation notation in the interest of keeping the variable counting unambiguous. I will henceforth assume that $\mathcal{P}_{ab} = 2\sigma_a^2 \delta_{ab}$ is diagonal (though not proportional to the identity). This can be motivated by the observation that a "very nondiagonal" $\mathcal{P}$ suggests that you chose bad variables θ, since they are strongly correlated, so you should choose independent variables θ such that $\mathcal{P}$ is diagonal.

Now, the minimum of the loss is where $d\mathcal{L}/d\theta_a = 0$ for all a. We can find this by solving the following equation for the optimal vector of variables, denoted $\hat{\theta}$:

$$\begin{aligned} 0 = \left. \frac{d\mathcal{L}}{d\theta_a} \right|_{\hat{\theta}_a} &= -\sum_{j=1}^{N_d} \sum_{i=1}^{N_d} x_{aj} \mathcal{C}_{ij}^{-1} y_i - \sum_{j=1}^{N_d} \sum_{i=1}^{N_d} \sum_{b=1}^{N_\ell} y_j \mathcal{C}_{ij}^{-1} \delta_{ab} x_{bi} + \sum_{j=1}^{N_d} \sum_{i=1}^{N_d} \sum_{b=1}^{N_\ell} x_{aj} \mathcal{C}_{ij}^{-1} \hat{\theta}_b x_{bi} \\ &\quad + \sum_{j=1}^{N_d} \sum_{i=1}^{N_d} \sum_{c=1}^{N_\ell} \sum_{b=1}^{N_\ell} \hat{\theta}_c x_{cj} \mathcal{C}_{ij}^{-1} \delta_{ab} x_{bi} + \frac{\hat{\theta}_a - \vartheta_a}{\sigma_a^2} \\ &= -\sum_{j=1}^{N_d} \sum_{i=1}^{N_d} x_{aj} \mathcal{C}_{ij}^{-1} y_i - \sum_{j=1}^{N_d} \sum_{i=1}^{N_d} y_j \mathcal{C}_{ij}^{-1} x_{ai} + \sum_{j=1}^{N_d} \sum_{i=1}^{N_d} \sum_{b=1}^{N_\ell} x_{aj} \mathcal{C}_{ij}^{-1} \hat{\theta}_b x_{bi} \\ &\quad + \sum_{j=1}^{N_d} \sum_{i=1}^{N_d} \sum_{c=1}^{N_\ell} \hat{\theta}_c x_{cj} \mathcal{C}_{ij}^{-1} x_{ai} + \frac{\hat{\theta}_a - \vartheta_a}{\sigma_a^2}, \end{aligned} \tag{2}$$

where in the second step we summed over the delta functions but are keeping the indices otherwise unchanged, to minimize ambiguity (e.g., the final symbol would be unclear otherwise).

Renaming one set of indices on $\hat{\theta}$ and some on x and y, and solving Eq. (2) for $\hat{\theta}_a$, gives

$$\sum_{j=1}^{N_d}\sum_{i=1}^{N_d}\sum_{b=1}^{N_\ell} x_{aj}\mathcal{C}_{ij}^{-1}\hat{\theta}_b x_{bi} + \hat{\theta}_a/2\sigma_a^2 = \sum_{j=1}^{N_d}\sum_{i=1}^{N_d} x_{aj}\mathcal{C}_{ij}^{-1}y_i + \vartheta_a/2\sigma_a^2. \quad (3)$$

This is a vector equation whose solutions appear to differ if $b = a$ or $b \neq a$. For simplicity, we define two new symbols:

$$d_a = \sum_{j=1}^{N_d}\sum_{i=1}^{N_d} y_j\mathcal{C}_{ij}^{-1}x_{ai}, \qquad b_{ab} = \sum_{j=1}^{N_d}\sum_{i=1}^{N_d} x_{bi}\mathcal{C}_{ij}^{-1}x_{aj}, \quad (4)$$

and we introduce another Kronecker delta such that Eq. (3) becomes

$$\sum_{b=1}^{N_\ell}\hat{\theta}_b\left(b_{ab} + \delta_{ab}/2\sigma_a^2\right) = d_a + \vartheta_a/2\sigma_a^2. \quad (5)$$

This is a family of N_ℓ different equations. These can be solved for each individual variable by multiplying by the inverse of the matrix in parentheses, giving the following equations for $\hat{\theta}_a$:

$$\hat{\theta}_a = \sum_{b=1}^{N_\ell}\left(b_{ab} + \delta_{ab}/2\sigma_a^2\right)^{-1}\left(d_b + \vartheta_b/2\sigma_b^2\right). \quad (6)$$

Notice that all of the indices evident in this equation are over the number of variables. The number of degrees of freedom N_d have all been summed over in Eq. (4). In the limit $\sigma_a \to 0$ (an infinitely strong prior), this becomes $\lim_{\sigma_a\to 0}\hat{\theta}_a = \vartheta_a$, and the limit $\sigma_a \to \infty$ (an infinitely weak prior), this recovers the result of `astro-ph/0310577`.

We can also ask about the covariance matrix for the solutions for the parameters $\hat{\theta}_a$, denoted $\mathrm{cov}(\hat{\theta}_a, \hat{\theta}_b)$. For notational convenience, I will define $\tilde{b}_{ab} \equiv b_{ab} + \delta_{ab}/2\sigma_a^2$ and $\tilde{d}_b \equiv d_b + \vartheta_b/2\sigma_b^2$. Because there will be no ambiguity in these manipulations, I will use Einstein summation convention:

$$\begin{aligned}
\mathrm{cov}(\hat{\theta}_a, \hat{\theta}_b) &= \mathrm{cov}(\tilde{b}_{ac}^{-1}\tilde{d}_c, \tilde{b}_{bf}^{-1}\tilde{d}_f) \\
&= \tilde{b}_{ac}^{-1}\tilde{b}_{bf}^{-1}\mathrm{cov}(\tilde{d}_c, \tilde{d}_f) \\
&= \tilde{b}_{ac}^{-1}\tilde{b}_{bf}^{-1}\mathrm{cov}(y_j\mathcal{C}_{ij}^{-1}x_{ci} + \vartheta_c/2\sigma_c^2, y_k\mathcal{C}_{kl}^{-1}x_{fl} + \vartheta_f/2\sigma_f^2) \\
&= \tilde{b}_{ac}^{-1}\tilde{b}_{bf}^{-1}\mathrm{cov}(y_j\mathcal{C}_{ij}^{-1}x_{ci}, y_k\mathcal{C}_{kl}^{-1}x_{fl}) \\
&= \tilde{b}_{ac}^{-1}\tilde{b}_{bf}^{-1}\mathcal{C}_{ij}^{-1}x_{ci}\mathcal{C}_{kl}^{-1}x_{fl}\mathrm{cov}(y_j, y_k) \\
&= \tilde{b}_{ac}^{-1}\tilde{b}_{bf}^{-1}\mathcal{C}_{ij}^{-1}x_{ci}\mathcal{C}_{kl}^{-1}x_{fl}\mathcal{C}_{jk} \\
&= \tilde{b}_{ac}^{-1}\tilde{b}_{bf}^{-1}\mathcal{C}_{il}^{-1}x_{ci}x_{fl} \\
&= \tilde{b}_{ac}^{-1}b_{cf}\tilde{b}_{bf}^{-1}.
\end{aligned} \quad (7)$$

In the first four steps, I either moved constant multiplicative factors outside of the cov function or annulled constant additive factors that were inside the cov function. For off-diagonal elements (or when $\sigma_a \to \infty$), this is the canonical result in `astro-ph/0310577` that $\mathrm{cov}(\hat{\theta}_a, \hat{\theta}_b) = b_{ab}^{-1}$, but the autocovariance can be smaller than this if the parameters have strong priors.