5.5 Ridge Regression

In linear regression, we have a real valued measurement Y of a signal $\mathbf{x}$ (potentially a vector) that we want to measure, potentially corrupted by noise. Linear regression is ubiquitous as a component of several algorithms, but a commonplace standalone (and important) example is fMRI signals. We will assume that the measurements are linear (*i.e.,* the signal $\mathbf{x}$ is transformed by linear operations, which is equivalent to multiplying by a matrix in general).

This is quite a vast topic in itself, and this module covers what is known as Ridge Regression. The formulation we adopt here is the "Bayesian" view. You will contrast this with the *frequentist* approach, and it is recommended you familiarize yourself with the frequentist (Maximum Likelihood) approach first.

We posit, as in the Maximum Likelihood case that the target Y and the measurements B of the unknown signal X be written in matrix form as

$$Y = BX + Z,$$

where specifically, Y is a vector of the n targets, B is the $n \times k$ measurement matrix, and Z is a vector of discrepancies. Here again we assume Z is a vector of n Gaussian random variables. But now we assume something more, that X is a multivariate Gaussian as well, independent of Z. Therefore, X, Y and Z are all jointly Gaussian. Specifically, we assume $X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $Z \sim \mathcal{N}(0, \nu^2 I)$ (and X and Z are independent). We assume X has a non-degenerate pdf, *i.e.,* Σ is invertible.

The best estimate of Y from X in the mean square sense, namely

$$\hat{X}(Y) = \arg\min_{f(Y)} \mathbb{E}||X - f(Y)||^2,$$

where the minimization is over any function of the observations Y, is given by

$$\hat{X}(Y) = \mathbb{E}[X|Y]$$

by standard arguments from EE342.

Of course, the conditional mean $\mathbb{E}[X|Y]$ has a special form for Gaussians, and we can reuse the insights from multivariate Gaussians.

5.6 Gaussians and ridge regression

Find the answers to the (why?) questions in the Gaussian module or by elementary manipulations.

Since X and Y are jointly Gaussian, we know therefore that $\mathbb{E}[X|Y]$ is a linear function of Y (why?). Furthermore since $\mathbb{E}X = \mathbb{E}Z = \mathbf{0}$ by assumption, we have $\mathbb{E}Y = \mathbf{0}$ as well (why?). Therefore,

$$\mathbb{E}[X|Y] = AY,$$

and the orthogonality principle yields,

$$A = cov(X, Y)cov(Y, Y)^{-1}$$

Now, show that

$$cov(X, Y) = \Sigma B^T$$

and

$$cov(Y, Y) = B\Sigma B^T + \nu^2 I.$$

Therefore,

$$\begin{aligned} A &= \frac{1}{\nu^2}\Sigma B^T(\frac{1}{\nu^2}B\Sigma B^T + I)^{-1} \\ &= \Sigma(I + \frac{1}{\nu^2}B^T B\Sigma)^{-1}\frac{1}{\nu^2}B^T \\ &= (B^T B + \Sigma^{-1}\nu^2)^{-1}B^T \end{aligned}$$

Can you prove the last and the second to last equalities? They are very useful, and form part of the series of equalities that go into the Matrix Inversion Lemma. The second equality states for any X and Y such that both XY and YX exist, and $I + XY$ is invertible, we will have that $I + YX$ is also invertible and

$$(I + YX)^{-1}Y = Y(1 + XY)^{-1}.$$

The last equality is simple manipulations using $(AB)^{-1} = B^{-1}A^{-1}$ and noting that Σ is invertible.

Therefore, our estimate of X in the Bayesian sense is

$$(B^T B + \nu^2\Sigma^{-1})^{-1}B^T Y, \tag{5}$$

which actually looks quite close to the OLS estimate of $(B^T B)^{-1}B^T Y$, differing only by the $\nu^2\Sigma^{-1}$ term within the inverse.

Regularization view To get a little more insight into this, let $\Sigma = \sigma^2 I$. The expression for the BAyesian estimate of X reduces to

$$(B^T B + \frac{\nu^2}{\sigma^2} I)^{-1} B^T Y.$$

Now consider solving the following problem

$$\arg\min_{\mathbf{x}} ||Y - B\mathbf{x}||^2 + \frac{\nu^2}{\sigma^2}||\mathbf{x}||^2. \tag{6}$$

We would take the gradient of the expression above with respect to $\mathbf{x}$, and this turns out to be

$$2B^T(B\mathbf{x} - Y) + \frac{\nu^2}{\sigma^2} 2\mathbf{x}.$$

Setting the gradient to $\mathbf{0}$ and rearranging we get

$$(B^T B + \frac{\nu^2}{\sigma^2} I)\mathbf{x} = B^T Y,$$

and solving for $\mathbf{x}$ gives us the Bayesian optimal solution. You can verify that the Hessian,

$$(B^T B + \frac{\nu^2}{\sigma^2} I)^T = (B^T B + \frac{\nu^2}{\sigma^2} I)$$

is positive definite if B has a trivial null space, as in the OLS case, thus the Bayesian optimal solution $(B^T B + \frac{\nu^2}{\sigma^2} I)^{-1} B^T Y$ is a minima. It is also the global minima because the objective being minimized is *convex* (over a convex domain) and we can have only one minimum.

Therefore, the Bayesian framework is equivalent to minimizing Equation˜(6). Notice the objective we are minimizing is the least squares loss ($||Y - B\mathbf{x}||^2$), but we add to it a term that depends on the length squared of $\mathbf{x}$, *i.e.,* $||\mathbf{x}||^2$.

If $\mathbf{x}_b$ is the Bayes optimal solution, we must have

$$||Y - B\mathbf{x}_b||^2 + \frac{\nu^2}{\sigma^2}||\mathbf{x}_b||^2 \leq ||Y - B\mathbf{x}_{OLS}||^2 + \frac{\nu^2}{\sigma^2}||\mathbf{x}_{OLS}||^2, \tag{7}$$

because, of course, $\mathbf{x}_b$ is the minima of . In fact, we usually have

$$||Y - B\mathbf{x}_b||^2 > ||Y - B\mathbf{x}_{OLS}||^2$$

but

$$||\mathbf{x}_b||^2 < ||\mathbf{x}_{OLS}||^2,$$

while satisfying Equation˜(7).

Therefore the Bayes optimal solution $\mathbf{x}_b$ shrinks the OLS estimate. Estimates like this, where we give preference to solutions that have some simplicity (here by simplicity we mean smaller Euclidean length), are called *regularized* estimates. Here the term $||\mathbf{x}||^2$ in the objective function of˜(6) is called the regularization penalty, and since this is the ℓ_2 norm of $\mathbf{x}$, this sort of regularization is called ℓ_2 regularization or Ridge regression.

Problem Show that the general Bayes optimal solution in Equation˜(5) is the solution of the optimization problem

$$\arg\min_{\mathbf{x}} ||Y - B\mathbf{x}||^2 + \frac{1}{\nu^2}\mathbf{x}^T\Sigma^{-1}\mathbf{x},$$

where Σ is the covariance matrix of X (hence symmetric and positive definite).

Problem Find answers to all the (why?) questions in the handout.