From least squares to shrinkage

Dror Baron

This supplement adds some information about shrinkage, ridge regression, and the least absolute shrinkage and selection operator (LASSO) formulation.

Least squares: Recall that least squares (LS) seeks the β vector that minimizes the squared Euclidean error,

$$\widehat{\beta}^{\mathrm{LS}} = \arg\min_{\beta}\{\|y - X\beta\|^2\}. \tag{1}$$

This minimization was justified by the underlying assumption that the modeling error, $y - X\beta$, can be modeled as Gaussian i.i.d. Moreover, the Gaussianity of the errors is plausible in many problem settings and applications, because the central limit theorem shows that a sum of independent variables (each of them generated by some unmodeled term) tends to a Gaussian distribution.

Disadvantages of LS: While elegant and offering a simple closed form solution, the LS formulation (1) is not plausible in many situations. For example, many communication channels can be modeled as linear,

$$y = x * \beta + z,$$

where x is a finite impulse response (FIR) filter corresponding to the channel structure, β is an unknown input transmitted into the channel, $h * \beta$ denotes convolution, z is additive noise, and y is the noisy output of the channel. Because convolution can be expressed as a matrix vector product, where the matrix is Toeplitz, we can rewrite the channel output in the linear form we have been focusing on,

$$y = \mathrm{Toeplitz}(x)\beta + z.$$

In many communication applications, the objective is to estimate the unknown channel input β from y and knowledge of x, which yields the matrix form $X = \mathrm{Toeplitz}(x)$. That said, the least squares solution is often sub-optimal. To see why it is sub-optimal, observe that the channel input β is often comprised of discrete valued symbols, and when LS multiplies the real valued y by the pseudo inverse matrix X^+, we very likely have a real valued $\widehat{\beta}^{\mathrm{LS}}$, which contradicts the discrete valued nature of β. Below are some approaches that improve the estimation procedure by incorporating assumptions about the structure of β.

Ridge regression: Suppose that the structure of β is not known. (For example, it might be discrete valued, but we are unaware of this.) All else being equal, we would probably

prefer smaller β. The ridge regression formulation balances between a desire to minimize the residual, $\|y - X\beta\|^2$, while preferring solutions that have small ℓ_2 norm. That is,

$$\widehat{\beta}^{\text{ridge}} = \arg\min_{\beta}\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\}, \tag{2}$$

where λ is a Lagrangian parameter that governs the trade-off between the residual and size of β. Similar to LS, which has a convenient closed form solution via the pseudo inverse matrix X^+, the ridge regression formulation (2) can be solved using a closed form expression,

$$\widehat{\beta}^{\text{ridge}} = (XX^T + \lambda I)^{-1}X^T y,$$

where I is an identity matrix.

Least absolute shrinkage and selection operator (LASSO) formulation: While ridge regression (2) encourages the Euclidean norm of β to be small, the Euclidean norm does not encourage individual entries of β to be zero. In fact, small entries barely contribute to the Euclidean norm, and so it seems plausible to expect many small non-zero entries in $\widehat{\beta}^{\text{ridge}}$. However, in many applications we want β to be sparse, meaning that many of its entries are zero or small. From the machine learning angle, one advantage of sparsity relates to subset selection; the human end-user of our model will often prefer to deal with a modest number of features. Another advantage in signal acquisitions applications, which will be discussed in greater detail in the sparse signal processing part of the course, is that many natural signals are sparse with respect to appropriately chosen bases or tight frames such as Fourier and wavelets.

How then can we encourage sparsity in β? One way to do so is to attempt to minimize the residual, $\|y - X\beta\|^2$, while preferring solutions that have small ℓ_1 norm. In contrast to the Euclidean ℓ_2 norm minimized as part of ridge regression, the ℓ_1 norm penalizes small nonzero entries of β more severely, which often yields solutions that have numerous zeros. We now define the LASSO formulation as follows,

$$\widehat{\beta}^{\text{LASSO}} = \arg\min_{\beta}\{\|y - X\beta\|^2 + \lambda\|\beta\|_1\}, \tag{3}$$

where once again λ is a Lagrangian parameter that governs the trade-off between the residual and the size of β, which is now quantified using the sparsity-promoting ℓ_1 norm.

In contrast to LS and ridge regression, which have convenient linear closed form solutions, LASSO requires a more complicated optimization procedure. It can be shown that LASSO involves convex optimization, and lends itself to various algorithmic solvers. One such algorithmic approach is gradient projection for sparse reconstruction (GPSR) by Nowak et al. In the sparse signal processing part of the course, we will study the approximate message passing (AMP) algorithmic framework by Donoho et al., which provides a faster solution for LASSO in certain circumstances, for example when individual entries of the X matrix are i.i.d.

Interpretation of LASSO as shrinkage: Both ridge regression and LASSO tend to shrink β toward zero. All things being equal, we prefer smaller β. Let us now investigate the effect of the Lagrangian λ within the LASSO formulation (3). When $\lambda = 0$, the ℓ_1 norm receives

no weight, and LASSO returns the same result as LS. As mentioned earlier, y is expected to be continuous valued, and so $\widehat{\beta}^{LS}$ will likely be continuous valued, meaning that none of its entries will be zero. As we increase λ, the ℓ_1 norm will receive more weight. The optimal solution $\widehat{\beta}^{LASSO}(\lambda)$ is a function of λ, and as we increase λ there will be more zero entries. Note that the vector $\widehat{\beta}^{LASSO}(\lambda)$ is continuous in λ, but once λ is large enough to drive one of the entries to zero, that entry will remain zero.

Later in the course, we will see that other types of knowledge about the structure of β can be incorporated within various algorithms.