1 Background

1.1 Predictive Shrinkage

Suppose we are given an $n \times p$ centered data matrix X and an $n \times 1$ vector of responses Y—the pair (X, Y) constitutes the training data. Assume the linear model $Y_i = \alpha + (\beta^*)^\mathsf{T} X_i + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Define the ordinary least squares (OLS) estimate

$$(\widehat{\alpha}, \widehat{\beta}) := \arg\min_{\alpha,\beta} \left\{ \|Y - \mathbf{1}\alpha - X\beta\|_2^2 \right\}, \tag{1}$$

$$\text{given by } \widehat{\alpha} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } \widehat{\beta} = \left(X^\mathsf{T} X\right)^{-1} X^\mathsf{T}(Y - \overline{Y}\mathbf{1}). \tag{2}$$

If we have test data $x \sim (0, \Sigma)$ and $y = \alpha + x^\mathsf{T}\beta^* + \varepsilon$, the OLS prediction is $\widehat{y} = \widehat{\alpha} + \widehat{\beta}^\mathsf{T} x$. The bias-variance decomposition of prediction mean square error (PMSE) is

$$\begin{aligned} \mathbb{E}\left[(y - \widehat{y})^2\right] &= \left(1 + \frac{1}{n}\right)\sigma^2 + \mathbb{E}\left[(\beta^* - \widehat{\beta})^\mathsf{T}\Sigma(\beta^* - \widehat{\beta})\right] &(3)\\ &= \left(1 + \frac{p+1}{n} + \frac{1}{n}\mathrm{tr}\left(\mathbb{E}\left[(\Sigma - S)\, S^{-1}\right]\right)\right)\sigma^2, &(4) \end{aligned}$$

where $S = n^{-1}X^\mathsf{T}X$. If the design of the training set is fixed (so S is constant) and we assume the test set follows the distribution of the training set in the sense that $\Sigma = S$, then the last term vanishes and so the PMSE is $\left(1 + \frac{1}{n} + \frac{p}{n}\right)\sigma^2$. Alternatively, if the rows X_i of X are all i.i.d. $\mathcal{N}(0, \Sigma)$, then nS is a Wishart matrix, and so $\mathbb{E}\left[(nS)^{-1}\right] = \frac{\Sigma^{-1}}{\nu}$ where $\nu = n - p - 1$, yielding a larger overall PMSE of $\left(1 + \frac{1}{n} + \frac{p}{\nu}\right)\sigma^2$.

Under the first set of assumptions, where $\Sigma = S$ exactly, we can write the last term in equation (3) as $\mathbb{E}\left[(\beta^* - \widehat{\beta})^\mathsf{T}\Sigma(\beta^* - \widehat{\beta})\right] = \mathbb{E}\left[\|\widehat{\xi} - \xi\|_2^2\right]$, where $\widehat{\xi} = \Sigma^{1/2}\widehat{\beta} \sim \mathcal{N}(\xi^*, (\sigma^2/n)I_p)$ and $\xi^* = \Sigma^{1/2}\beta^*$. This is a normal-means estimation problem, so when $p > 2$ we can achieve lower MSE $\mathbb{E}\left[\|\widetilde{\xi} - \xi^*\|_2^2\right] < \mathbb{E}\left[\|\widehat{\xi} - \xi^*\|_2^2\right]$ with the James-Stein estimate

$$\widetilde{\xi} = \left(1 - \frac{(p-2)(\widehat{\sigma}^2/n)\nu}{(\nu+2)\|\widehat{\xi}\|_2^2}\right)\widehat{\xi}, \tag{5}$$

yielding the shrunk regression coefficients $\widetilde{\beta} = \widehat{K}\widehat{\beta}$, where $\widehat{K} = \left(1 - \frac{(p-2)(\widehat{\sigma}^2/n)\nu}{(\nu+2)\widehat{\beta}^\mathsf{T} S\widehat{\beta}}\right)$. It follows that $\widetilde{y} = \widehat{\alpha} + \widehat{K}\widehat{\beta}$ has strictly better PMSE than OLS $\widehat{y}$. Since $\widehat{K} < 1$ we are left to conclude that the OLS predictions on held-out data were too large in magnitude. Pre-shrunk predictors of this form were first studied by Copas (1983), who also provided the Stein-shrinkage interpretation.

Another form of shrinkage is *ridge regression*

$$(\widehat{\alpha}, \widehat{\beta}(\lambda)) := \arg\min_{\alpha,\beta} \left\{ \|Y - \mathbf{1}\alpha - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\}, \tag{6}$$

$$\text{given by } \widehat{\alpha} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } \widehat{\beta}(\lambda) = \left(X^\mathsf{T} X + \lambda I_p\right)^{-1} X^\mathsf{T}(Y - \overline{Y}\mathbf{1}). \tag{7}$$

In particular, $\widehat{\beta}(0) = \widehat{\beta}$ is OLS and $\lim_{\lambda\uparrow\infty}\widehat{\beta}(\lambda) = 0$. Note $\widehat{\beta}(\lambda) = \left(X^\mathsf{T}X + \lambda I_p\right)^{-1}(X^\mathsf{T}X)\widehat{\beta}$ shrinks $\widehat{\beta}$ towards zero in a manner that accounts for the variability in the covariates X. For $\lambda > 0$ it is not the case that $\widehat{\beta}(\lambda) = K\widehat{\beta}$ for some K, which is to say the family of estimators $\widetilde{\beta}(K, \lambda) = K\widehat{\beta}(\lambda)$ is not overparametrized. We allow any $K > 0$, since for λ large, $\widehat{\beta}(\lambda)$ tends to overshrink. The estimator minimizes the following loss functions:

$$\begin{aligned} \widetilde{\beta}(K, \lambda) &:= \arg\min_{\beta} \left\{ \|K(Y - \mathbf{1}\overline{Y}) - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\} &(8)\\ &= \arg\min_{\beta} \left\{ \|Y - \mathbf{1}\overline{Y} - X\beta\|_2^2 + \frac{\lambda}{K}\|\beta\|_2^2 + (K^{-1} - 1)\|X\beta\|_2^2 \right\}. &(9) \end{aligned}$$

The first minimization performs ridge regression on the centered design X and shrunk (or inflated) centered responses $K(Y - \mathbf{1}\overline{Y})$. The second is like ridge with an additional penalty. If $K < 1$ then $(K^{-1} - 1)\|X\beta\|_2^2$ penalizes large $X\beta$, and if $K > 1$ it penalizes small $X\beta$.

When $\lambda > 0$, the ridge regression estimate $\widehat{\beta}(\lambda)$ is biased, but has the advantage of being well-defined even in the high dimensional case $n < p$. Whether we can benefit from the additional pre-factor K depends on how reliably we can detect whether we need to inflate or shrink. In the case where β^* is known to be sparse, we can ask the same set of questions about the magnitude of our predictions after Lasso, where the regularization term in (6) is replaced with $\lambda\|\beta\|_1$. We start with simulation studies to get a sense of in what cases we can significantly improve predictions by including K.