In-sample prediction and Mallow's C_p

Consider a linear regression model with p predictors (let's ignore the intercept at this moment).

- Index all possible variable subsets by a p-dimensional binary vector (totally 2^p subsets or models):
$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^t \in \{0,1\}^p.$$
Especially, $\boldsymbol{\gamma} = (1,1,\ldots,1)$ denotes the biggest (full) model that includes all the predictors, and $\boldsymbol{\gamma} = (0,0,\cdots,0)$ denotes the smallest (null) model that does not include any predictors.

- For a variable set $\boldsymbol{\gamma}$, define $p_{\boldsymbol{\gamma}} = \sum_j \gamma_j$ to denote the number of variables included in this set, and use $\mathbf{X}_{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ to denote the corresponding $n \times p_{\boldsymbol{\gamma}}$ design matrix and $p_{\boldsymbol{\gamma}}$-dim LS regression parameter, respectively.

In-sample prediction

The so-called *in-sample prediction error* measures prediction errors at the n sample points $\mathbf{x}_i$'s. For a model $\boldsymbol{\gamma}$, the error is defined to be

$$R(\boldsymbol{\gamma}) = \mathbb{E}\|\mathbf{y}^* - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2, \tag{1}$$

where

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}, \quad \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y},$$

and $\mathbf{y}^*$ is a set of imaginary, new data points observed at $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ which are independent of the training data $\mathbf{y}$.

The expectation in (1) is taken with respect to the true distribution over $\mathbf{y}$ and $\mathbf{y}^*$. Here is our assumption on the true data generating process:

$$\mathbf{y}_{n\times 1},\ \mathbf{y}^*_{n\times 1} \text{ i.i.d.} \sim \mathsf{N}_n\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n\right). \tag{2}$$

Or equivalently, assume

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\mu} + \mathbf{e}, \\ \mathbf{y}^* &= \boldsymbol{\mu} + \mathbf{e}^* \\ \mathbf{e}_{n\times 1},\ \mathbf{e}^*_{n\times 1} \text{ i.i.d.} &\sim \mathsf{N}_n\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right). \end{aligned}$$

Next we decompose the prediction error into three components.

$$\begin{aligned} R(\boldsymbol{\gamma}) &= \mathbb{E}\|\mathbf{y}^* - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\ &= \mathbb{E}\|\mathbf{y}^* - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\ &= \mathbb{E}\|(\mathbf{y}^* - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\ &= \mathbb{E}\|\mathbf{y}^* - \boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \mathbb{E}\|\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\ &= I + II + III. \end{aligned}$$

Note that all the cross-product terms are equal to zero

$$\begin{aligned} \mathbb{E}(\mathbf{y}^* - \boldsymbol{\mu})^t(\boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}) &= (\boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^t\ \mathbb{E}(\mathbf{y}^* - \boldsymbol{\mu}) = (\boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^t\mathbf{0} = 0. \\ \mathbb{E}(\boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^t(\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) &= (\boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^t\ \mathbb{E}(\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = 0 \\ \mathbb{E}\Big[(\mathbf{y}^* - \boldsymbol{\mu})^t(\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})\Big] &= \Big[\mathbb{E}(\mathbf{y}^* - \boldsymbol{\mu})\Big]^t\Big[\mathbb{E}(\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})\Big] = 0 \end{aligned}$$

We decompose $R(\boldsymbol{\gamma})$ as $I + II + III$.

- The 1st term: the unavoidable error that you would encounter even if you know the true parameter $\boldsymbol{\beta}$:
$$I = \|\mathbf{y}^* - \boldsymbol{\mu}\|^2 = \mathbb{E}\|\mathbf{e}^*\|^2 = n\sigma^2.$$
- The 2nd term: Let $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ be the $p_{\boldsymbol{\gamma}}$-dim vector that achieves the following minimal
$$II = \min_{\boldsymbol{\alpha}\in\mathbb{R}^{p_{\boldsymbol{\gamma}}}} \|\boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\alpha}\|^2 = \|\boldsymbol{\mu} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2.$$
That is, $\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \mathbf{H}_{\boldsymbol{\gamma}}\boldsymbol{\mu}$ is the projection of the true mean vector $\boldsymbol{\mu}$ onto the subspace spanned by columns of $\mathbf{X}_{\boldsymbol{\gamma}}$. The bias will be zero, if $\mu = \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ (this would happen if $\boldsymbol{\gamma}$ contains all the true predictors). The bias will not be zero if the model $\boldsymbol{\gamma}$ misses any true predictors.
- The 3rd term: the variance of model $\boldsymbol{\gamma}$ (due to estimating $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$). Note that $\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \mathbf{H}_{\boldsymbol{\gamma}}\boldsymbol{\mu}$ and $\mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y}$. Then
$$\begin{aligned} III &= \mathbb{E}\|\mathbf{H}_{\boldsymbol{\gamma}}\boldsymbol{\mu} - \mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y}\|^2 \\ &= \mathbb{E}\|\mathbf{H}_{\boldsymbol{\gamma}}\big(\mathbf{y} - \boldsymbol{\mu}\big)\|^2 = \sigma^2\mathrm{tr}(\mathbf{H}_{\boldsymbol{\gamma}}) = \sigma^2 p_{\boldsymbol{\gamma}}. \end{aligned}$$

Mallow's C_p

In practice, we do not know the true model, i.e., we cannot calculate the 2nd term (the bias). We try to get that information from the RSS from model $\boldsymbol{\gamma}$. Next let's look at the

expected RSS_γ, and do a similar decomposition.

$$\begin{aligned}\mathbb{E}[\text{RSS}_\gamma] &= \mathbb{E}_{\mathbf{y}}\|\mathbf{y}-\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2 \\ &= \mathbb{E}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\boldsymbol{\beta}-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma+\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma-\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2 \\ &= \mathbb{E}\|\mathbf{y}-\boldsymbol{\mu}\|^2+\|\boldsymbol{\mu}-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma\|^2+\mathbb{E}\|\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma-\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2 \\ &\quad -2\mathbb{E}(\mathbf{y}-\boldsymbol{\mu})^T(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma-\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)\end{aligned}$$

The first 3 terms are the same as the ones in $R(\gamma)$. But now we have a cross-product term which does not appear in our derivation for prediction, since there we evaluate the error at a set of new test data $\mathbf{y}^*$ that is independent of the training data $\mathbf{y}$, therefore the cross-product term is zero (they are uncorrelated). But for RSS, the data $\mathbf{y}$ is used both for evaluation and for learning (it's used twice), so the cross-product term will not disappear.

The cross-product term (the last term) in $\mathbb{E}[\text{RSS}_\gamma]$ is equal to

$$\mathbb{E}(\mathbf{y}-\boldsymbol{\mu})^T(\mathbf{H}_\gamma\mathbf{y}-\mathbf{H}_\gamma\mathbf{X}\boldsymbol{\beta}) = \sigma^2\text{tr}(\mathbf{H}_\gamma) = \sigma^2 p_\gamma$$

So

$$R(\gamma) \approx \text{RSS} + 2p_\gamma\sigma^2,$$

which gives rise to Mallow's C_p: $\text{RSS}_\gamma + 2p_\gamma\hat{\sigma}^2$, where we replace σ^2 by an estimate from the full model.

Solution of LASSO

The LASSO solution is define to be

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}|\right).$$

The orthogonal case

Suppose $\mathbf{X}_{n\times p}$ is ON, i.e., $\mathbf{X}^T\mathbf{X}=\mathbf{I}_p$. Then

$$\begin{aligned}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 &= \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2+\|\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}\|^2 \qquad (3)\end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ denotes the LS estimate and the cross-product term is 0,

$$2(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) = 2\mathbf{r}^T(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) = 0,$$

since the n-dim vector in red (which is a linear combination of columns of $\mathbf{X}$, no matter what value $\boldsymbol{\beta}$ takes) is orthogonal to the residual vector $\mathbf{r}$. Also note that the 1st term in (3) is not a function of $\boldsymbol{\beta}$, so we can redefine the objective function to be

$$\begin{aligned}\Omega(\boldsymbol{\beta}) &= \|\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda|\boldsymbol{\beta}| = \|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2+\lambda|\boldsymbol{\beta}| \\ &= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+\lambda|\boldsymbol{\beta}| = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+\lambda|\boldsymbol{\beta}|\end{aligned}$$

since $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$. Now the objective function can be expressed as a sum over the p dimensions

$$\Omega(\boldsymbol{\beta}) = \sum_{j=1}^{p} \Big[\beta_j - \hat{\beta}_j)^2 + \lambda|\beta_j|\Big].$$

So to optimize $\Omega(\boldsymbol{\beta})$, we can find the optimal β_j for each of $j = 1, \ldots, p$ separately.

A one-dimensional generic problem

Now we can work with the following generic problem: minimize $f(x)$ where

$$f(x) = (x-a)^2 + \lambda|x|, \quad \lambda > 0. \tag{4}$$

Without loss of generality, we assume $a > 0$ (the derivation for $a \le 0$ is similar). The function $f(x)$ is continuous, but $f'(x)$ is not defined at $x = 0$,

$$f'(x) = \begin{cases} 2(x-a) + \lambda = 2\big(x + \lambda/2 - a\big), & x > 0 \\ 2(x-a) - \lambda = 2\big(x - \lambda/2 - a\big), & x < 0 \end{cases}$$

- $a \in [0, \lambda/2]$. When $x > 0$, $f'(x) > 0$, i.e., $f(x)$ is an increasing function and (due to continuity) reaches its minimum at $x = 0$; $f'(x) < 0$ when $x < 0$, i.e., $f(x)$ is a decreasing function and reaches its minimum at $x = 0$. So

$$0 = \arg\min_{x\in\mathbb{R}} \Big[(x-a)^2 + \lambda|x|\Big], \text{ when } a \in [0, \lambda/2].$$

- $a \in (\lambda/2, \infty]$. When $x > 0$, $f'(x)$ is positive when x is large and is negative when x is small, i.e., when $x > 0$, $f(x)$ first decreases and then increases, and reaches its minimum at the point $x^* = a - \lambda/2$ with $f'(x^*) = 0$. When $x < 0$, $f'(x)$ is always negative, i.e., the value of $f(x)$ when $x < 0$ is always bigger than $f(0)$, and we've known that $f(0) > f(x^*)$. So

$$a - \lambda/2 = \arg\min_{x\in\mathbb{R}} \Big[(x-a)^2 + \lambda|x|\Big], \text{ when } a \in (\lambda/2, \infty).$$

- You can carry out the analysis for $a \in [-\lambda/2, 0)$ and $a \in (-\infty, -\lambda/2)$. Finally we have

$$x^* = \arg\min_{x} \Big[(x-a)^2 + \lambda|x|\Big] = \text{sgn}(a)\Big[|a| - \frac{\lambda}{2}\Big]_+, \tag{5}$$

  where $sgn(z)$ is the sign function that returns 1 if $z > 0$, 0 if $z = 0$, and -1 if $z < 0$.

  For example, suppose $\lambda = 2$. If $a = 1.5$, we have $x^* = 0.5$; if $a = 0.5$, we have $x^* = 0$; if $a = -1.5$, we have $x^* = -1$; if $a = -0.5$, we have $x^* = 0$.

The KKT condition

If $f'(x)$ exists, we can characterize the minimizer of $f(x)$ by $f'(x) = 0$. Due to the term $|x|$, The derivative of the f function defined at (4) is not well-defined at $x = 0$. However, since (4) is a convex function, we can characterize the minimizer via the subgradient. The minimizer of (4) must satisfy

$$\begin{cases} 2(x-a) + \lambda \text{sng}(x) = 0, & \text{if } x \neq 0; \\ 2(x-a) + \lambda s = 0, & \text{if } x = 0, \end{cases}$$

where $s \in [-1, 1]$. You can plug in the solution given at (5) to check that the equalities above hold true.

For the general design matrix $\mathbf{X}$, the Lasso solution $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ that minimizes $\big(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda|\boldsymbol{\beta}|\big)$ satisfies the following KKT conditions:

$$2X_j^t\Big(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{lasso}}\Big) = \lambda \cdot s_j,$$

where

$$s_j \begin{cases} = \text{sgn}(\hat{\beta}_j^{\text{lasso}}), & \text{if } \hat{\beta}_j^{\text{lasso}} \neq 0 \\ \in [-1, 1], & \text{if } \hat{\beta}_j^{\text{lasso}} = 0. \end{cases}$$

SCAD and other penalty functions

The penalty function on the p-dim coefficient vector $\boldsymbol{\beta}$ often takes an additive and symmetric (i.e., the penalty depends on the absolute value of each coefficient) form, $\sum_{j=1}^p p(|\beta_j|)$.

Consider a simple setting in which the design matrix $\mathbf{X}$ is orthonormal. Then we can express the penalized least squares as the following

$$\begin{aligned} & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p p(|\beta_j|) \\ = \; & \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \sum_{j=1}^p (\hat{\beta}_j - \beta_j)^2 + \lambda \sum_{j=1}^p p(|\beta_j|) \end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ denotes the LS estimate of $\boldsymbol{\beta}$ and $\hat{\beta}_j$ denotes the j-th element of $\hat{\boldsymbol{\beta}}$. Apparently minimizing the objective function above (wrt $\boldsymbol{\beta}$) is equivalent to minimizing it component wise (wrt β_j for $j = 1 : p$). This leads us to consider the following generic form of penalized least squares

$$(z - \theta)^2 + \lambda p(|\theta|), \tag{6}$$

where z is a realization of a random variable Z whose distribution depends on the 1-dim parameter value θ, e.g., $Z \sim N(\theta, 1)$.

A natural question: what kind of "desirable" properties do we want $p(\cdot)$ to have? Fan and Li (2001)[1] listed the following three properties.

- *Unbiaseness*: when the observed data $|z|$ is large, the resulting estimator $\hat{\theta}$ should be nearly unbiased, which corresponds to requiring $p'(|\theta|) = 0$ when $|\theta|$ is large.
- *Sparsity*: when $|z|$ is small, the resulting estimator $\hat{\theta}$ should be shrunk to 0, which corresponds to requiring
$$\min_{|\theta|} \Big[|\theta| + \lambda p'(|\theta|) \Big] > 0.$$
- *Continuity*: the resulting estimator $\hat{\theta}(z)$ is continuous in data z, which corresponds to requiring
$$\arg\min_{|\theta|} \Big[|\theta| + \lambda p'(|\theta|) \Big] = 0.$$

It is easy to check that the model selection estimator (i.e., the hard-thresholding rule) isn't continuous, the ridge estimator isn't sparse, and the Lasso estimator (the soft-thresholding rule) is biased, but the SCAD estimator (see eq 3.82 in the textbook) proposed by Fan and Li (2001) possesses all three properties.

To know more on SCAD and related research, you can take a look of this review paper by Fan and Lv, "A selective overview of variable selection in high dimensional feature space," Statstica Sinica 20 (2010), 101-148.

[1] Fan and Li (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties", *Journal of the American Statistical Association*, Vol. 96:1348–1360.