Shrinkage methods

김지수 (Jisu KIM)

통계적 기계학습(Statistical Machine Learning), 2024 1st semester

The lecture note is a minor modification of the lecture notes from Prof. Yongdai Kim's "Statistical Machine Learning", and Prof Larry Wasserman and Ryan Tibshirani's "Statistical Machine Learning". Also, see Section 3 from [11].

1 Review

1.1 Basic Model for Supervised Learning

- Input(입력) / Covariate(설명 변수) : $x \in \mathbb{R}^p$, so $x = (x_1, \ldots, x_p)$.
- Output(출력) / Response(반응 변수): $y \in \mathcal{Y}$. If y is categorical, then supervised learning is "classification", and if y is continuous, then supervised learning is "regression".
- Model(모형) :

  $$y \approx f(x).$$

  If we include the error ϵ to the model, then it can be also written as

  $$y = \phi(f(x), \epsilon).$$

  For many cases, we assume additive noise, so

  $$y = f(x) + \epsilon.$$

- Assumption(가정): f belongs to a family of functions $\mathcal{F}$. This is the assumption of a model: a model can be still used when the corresponding assumption is not satisfied in your data.
- Loss function(손실 함수): $\ell(y, a)$. A loss function measures the difference between estimated and true values for an instance of data.
- Training data(학습 자료): $\mathcal{T} = \{(y_i, x_i), i = 1, \ldots, n\}$, where (y_i, x_i) is a sample from a probability distribution P_i. For many cases we assume i.i.d., or x_i's are fixed and y_i's are i.i.d..
- Goal(목적): we want to find f that minimizes the expected prediction error,

  $$f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)\sim P} \left[\ell(Y, f(X))\right].$$

- Prediction model(예측 모형): f^0 is unknown, so we estimate f^0 by $\hat{f}$ using data. For many cases we minimizes on the empirical prediction error, that is taking the expectation on the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(Y_i,X_i)}$.

  $$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P_n} \left[\ell(Y, f(X))\right] = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)).$$

- Prediction(예측): if $\hat{f}$ is a predicted function, and x is a new input, then we predict unknown y by $\hat{f}(x)$.

1.2 Linear Regression

From the additive noise model

$$y = f(x) + \epsilon,\ f \in \mathcal{F},$$

Linear Regression Model (선형회귀모형) is that

$$\mathcal{F} = \left\{ \beta_0 + \sum_{j=1}^{p} \beta_j x_j : \beta_j \in \mathbb{R} \right\}.$$

For estimating β, we use least squares: suppose the training data is $\{(y_i, x_{ij}) : 1 \le i \le n, 1 \le j \le p\}$. We use square loss

$$\ell(y, a) = (y - a)^2,$$

then the eimpirical loss becomes the residual sum of square (RSS) as

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2. \end{aligned}$$

Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ be the nimimizor of RSS, then the predicted function is

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j.$$

2 Introduction

When the dimension of input is large (e.g. larger than the sample size), there are lots of problems in applying simple methods (e.g. least square method). Twontorious problems in high dimensional problems are

- Multicollinearity : Some input variables are highly correlated. For example, when the dimension of input is large than the sample size, the least square estimator is not unique.
- Overfitting: A model with too many input variables may be sub-optimal when the true model is sparse (a response variable depends only on a small number of input variables).

Possible remedies are:

- Variable selection: Best Subset Selection (최적부분집합선택)
- Shrinkage methods: Ridge Regression (능선회귀), Lasso (라쏘), SCAD
- Dimension reduction techniques: Principal component regression, Partial least square (not covered in the class. See the text book).

We consider the best subset also as a shrinkage method, and covers Best subset selection, Ridge regression, Lasso.

3 Regularization

How do we deal with such issues? The short answer is *regularization*. In our present setting, we would modify the least squares estimator in one of two forms:

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \ \text{ s.t. } \ \beta \in C \qquad \text{(Constrained form)}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + P(\beta) \qquad \text{(Penalized form)}$$

where C is some (typically convex) set, and $P(\cdot)$ is some (typically convex) penalty function. At its core, regularization provides us with a way of navigating the bias-variance tradeoff: we (hopefully greatly) reduce the variance at the expense of introducing some bias.

3.1 Three norms: ℓ_0, ℓ_1, ℓ_2

In terms of regularization, we typically choose the constraint set C to be a sublevel set of a norm (or seminorm), and equivalently, the penalty function $P(\cdot)$ to be a multiple of a norm (or seminorm).

Let's consider three canonical choices: the ℓ_0, ℓ_1, and ℓ_2 norms:

$$\|\beta\|_0 = \sum_{j=1}^{p} 1\{\beta_j \neq 0\}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^{p} \beta_j^2\right)^{1/2}.$$

(Truthfully, calling it "the ℓ_0 norm" is a misnomer, since it is not a norm: it does not satisfy positive homogeneity, i.e., $\|a\beta\|_0 \neq a\|\beta\|_0$ whenever $a \neq 0, 1$.)

In constrained form, this gives rise to the problems:

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ s.t. } \|\beta\|_0 \leq k \qquad \text{(Best subset selection)} \tag{1}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ s.t. } \|\beta\|_1 \leq s \qquad \text{(Lasso regression)} \tag{2}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ s.t. } \|\beta\|_2^2 \leq s \qquad \text{(Ridge regession)} \tag{3}$$

where $k, s \geq 0$ are tuning parameters. Note that it makes sense to restrict k to be an integer; in best subset selection, we are quite literally finding the best subset of variables of size k, in terms of the achieved training error

Though it is likely the case that these ideas were around earlier in other contexts, in statistics we typically subset selection to [1, 12], ridge regression to [13], and the lasso to [15, 6]

In penalized form, the use of ℓ_0, ℓ_1, ℓ_2 norms gives rise to the problems:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0 \qquad \text{(Best subset selection)} \tag{4}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad \text{(Lasso regression)} \tag{5}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \qquad \text{(Ridge regression)} \tag{6}$$

with $\lambda \geq 0$ the tuning parameter. In fact, problems (2), (5) are equivalent. By this, we mean that for any $s \geq 0$ and solution $\hat{\beta}$ in (2), there is a value of $\lambda \geq 0$ such that $\hat{\beta}$ also solves (5), and vice versa. The same equivalence holds for (3), (6). (The factors of $1/2$ multiplying the squared loss above are inconsequential, and just for convenience)

It means, roughly speaking, that computing solutions of (2) over a sequence of t values and performing cross-validation (to select an estimate) should be basically the same as computing solutions of (5) over some sequence of λ values and performing cross-validation (to select an estimate). Strictly speaking, this isn't quite true, because the precise correspondence between equivalent s, λ depends on the data X, y

Notably, problems (1), (4) are *not equivalent*. For every value of $\lambda \geq 0$ and solution $\hat{\beta}$ in (4), there is a value of $k \geq 0$ such that also solves (1), but the converse is not true.

3.2 A Toy Example

It is helpful to first consider a toy example. Let $Y_1, \ldots, Y_n$ be a random sample from $N_p(\mu, \sigma^2 I)$.

Note that the MLE of μ is $\bar{Y}$. When $p \leq 2$, $\bar{Y}$ is the best estimator. However, when $p \geq 3$, surprisingly, $\bar{Y}$ is sub-optimal.

A better estimator can be constructed by shrinking $\bar{Y}$ toward 0 as follows. In fact, the James-Stein estimator given as

$$\delta^{JS} = \left(1 - \frac{1}{\sum_{i=1}^{p} Y_i^2}\right)\bar{Y}$$

is better than $\bar{Y}$. Note that $|\delta^{JS}| \leq |\bar{Y}|$. It is known that efficiency gain of J-S over the MLE is substantial when p is large.

Let's consider the three different estimators we get using the following three different loss functions:

$$\frac{1}{2}(Y - \mu)^2 + \lambda||\mu||_0, \quad \frac{1}{2}(Y - \mu)^2 + \lambda|\mu|, \quad \frac{1}{2}(Y - \mu)^2 + \lambda\mu^2.$$

You should verify that the solutions are

$$\hat{\mu} = H(Y; \sqrt{2\lambda}), \quad \hat{\mu} = S(Y; \lambda), \quad \hat{\mu} = \frac{Y}{1+2\lambda}$$

where $H(y; a) = yI(|y| > a)$ is the hard-thresholding operator, and

$$S(y; a) = \begin{cases} y - a & \text{if } y > a \\ 0 & \text{if } -a \le y \le a \\ y + a & \text{if } y < a. \end{cases}$$

Hard thresholding creates a "zone of sparsity" but it is discontinuous. Soft thresholding also creates a "zone of sparsity" but it is scontinuous. The L_2 loss creates a nice smooth estimator but it is never sparse. (You can verify the solution to the L_1 problem using sub-differentials if you know convex analysis, or by doing three cases separately: $\mu > 0$, $\mu = 0$, $\mu < 0$.)

3.3 Sparsity

The best subset selection and the lasso estimators have a special, useful property: their solutions are *sparse*, i.e., at a solution $\hat{\beta}$ we will have $\hat{\beta}_j = 0$ for many components $j \in \{1, \ldots, p\}$. In problem (1), this is obviously true, where $k \ge 0$ controls the sparsity level. In problem (2), it is less obviously true, but we get a higher degree of sparsity the smaller the value of $s \ge 0$. In the penalized forms, (4), (5), we get more sparsity the larger the value of $\lambda \ge 0$

This is not true of ridge regression, i.e., the solution of (3) or (6) generically has all nonzero components, no matter the value of t or λ. Note that sparsity is desirable, for two reasons: (i) it corresponds to performing variable selection in the constructed linear model, and (ii) it provides a level of interpretability (beyond sheer accuracy)

That the ℓ_0 norm induces sparsity is obvious. But, why does the ℓ_1 norm induce sparsity and not the ℓ_2 norm? There are different ways to look at it; let's stick with intuition from the constrained problem forms (2), (5). Figure 1 shows the "classic" picture, contrasting the way the contours of the squared error loss hit the two constraint sets, the ℓ_1 and ℓ_2 balls. As the ℓ_1 ball has sharp corners (aligned with the coordinate axes), we get sparse solutions

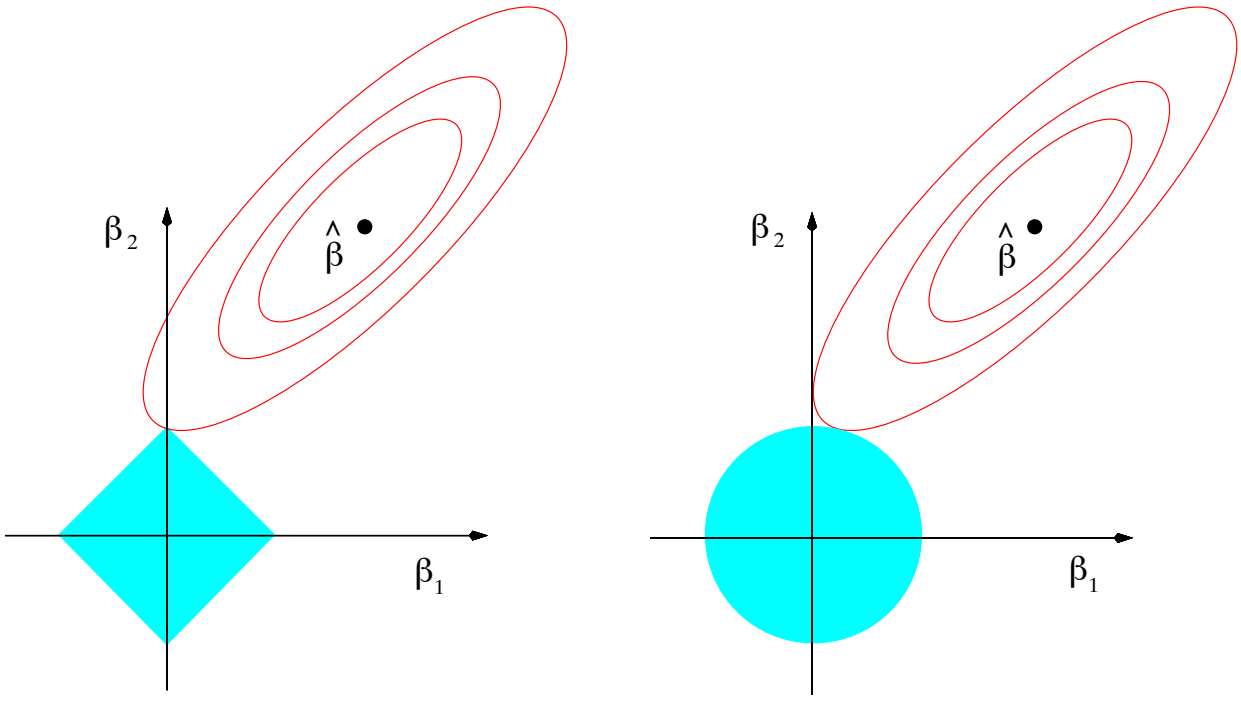

Figure 1: *The "classic" illustration comparing lasso and ridge constraints. From Figure 3.11 of* [11].

Intuition can also be drawn from the orthogonal case. When X is orthogonal, it is not hard to show that the

solutions of the penalized problems (4), (5), (6) are

$$\hat{\beta}^{\text{subset}} = H_{\sqrt{2\lambda}}(X^T y), \quad \hat{\beta}^{\text{lasso}} = S_\lambda(X^T y), \quad \hat{\beta}^{\text{ridge}} = \frac{X^T y}{1+2\lambda}$$

respectively, where $H_s(\cdot), S_s(\cdot)$ are the componentwise hard- and soft-thresholding functions at the level s. We see several revealing properties: subset selection and lasso solutions exhibit sparsity when the componentwise least squares coefficients (inner products $X^T y$) are small enough; the lasso solution exihibits shrinkage, in that large enough least squares coefficients are shrunken towards zero by λ; the ridge regression solution is never sparse and compared to the lasso, preferentially shrinkage the larger least squares coefficients even more

3.4 Convexity

The lasso and ridge regression problems (2), (3) have another very important property: they are convex optimization problems. Best subset selection (1) is not, in fact it is very far from being convex. Consider using the norm $||\beta||_p$ as a penalty. Sparsity requires $p \leq 1$ and convexity requires $p \geq 1$. The only norm that gives sparsity and convexity is $p = 1$. The appendix has a brief review of convexity.

4 Variable Selection

- For given $k \leq p$, choose k many input variables, with which the residual mean square error is minimized among all models having k many input variables. Denote this model M_k.
- Select the optimal model among $M_0, \ldots, M_p$. (we will see this later in Model selection.)
- The complexity of the model is proportional to k.
- If p is very large (say, larger than 40), this approach (all possible search) becomes computationally infeasible.
- An alternative is forward selection, backward elimination and stepwise.

- Variable selection methods are known to be unstable.
- "Unstable" means that small change of data results in large change of the estimator.
- This is because variable selection uses a hard decision rule (survive or die).
- The instability causes sub-optimal prediction accuracy.
- See "Breiman (1996). Heuristics of instability and stabilization in model selection, *Annals of Statistics*, **24**, 2350-2383".
- Shrinkage methods are promising alternatives.

4.1 Theory For Subset Selection

Despite its computational intractability, best subset selection has some attractive risk properties. A classic result is due to [8], on the in-sample risk of best subset selection in penalized form (4), which we will paraphrase here. First, we raise a very simple point: if A denotes the support (also called the active set) of the subset selection solution in (4)—meaning that $\hat{\beta}_j = 0$ for all $j \notin A$, and denoted $A = \text{supp}(\hat{\beta})$—then we have

$$\begin{aligned} \hat{\beta}_A &= (X_A^T X_A)^{-1} X_A^T y, \\ \hat{\beta}_{-A} &= 0. \end{aligned} \tag{7}$$

Here and throughout we write X_A for the columns of matrix X in a set A, and x_A for the components of a vector x in A. We will also use X_{-A} and x_{-A} for the columns or components not in A. The observation in (7) follows from the fact that, given the support set A, the ℓ_0 penalty term in the subset selection criterion doesn't depend on the actual magnitudes of the coefficients (it contributes a constant factor), so the problem reduces to least squares.

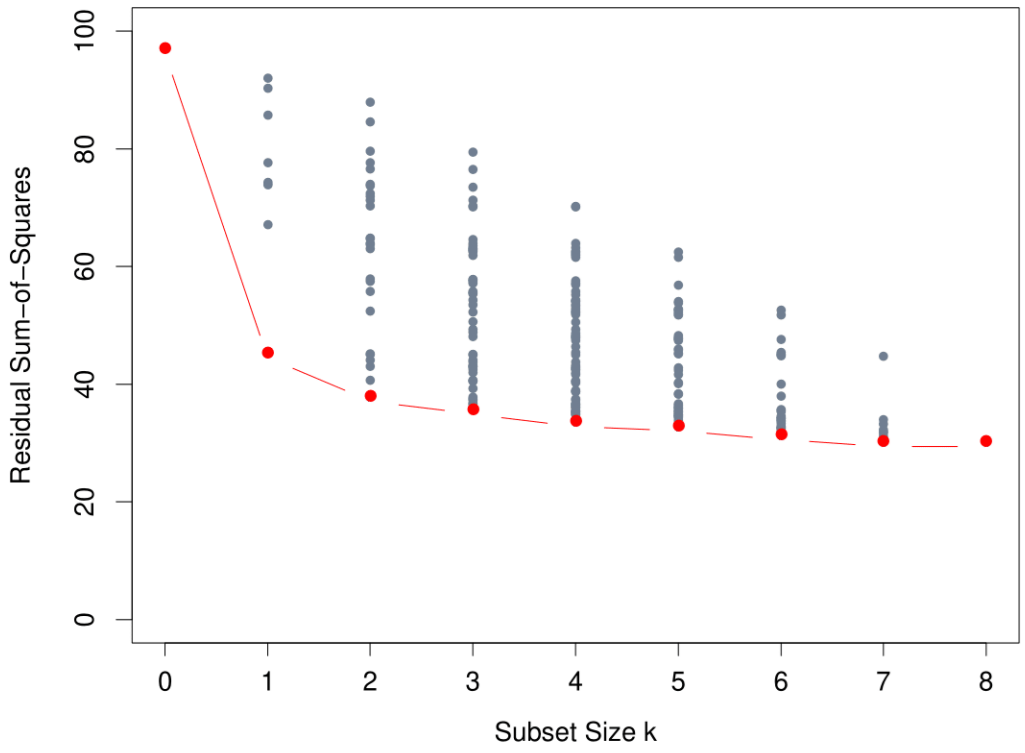

Figure 2: All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size. Figure 3.5 from [11].

Now, consider a standard linear model as with X fixed, and $\epsilon \sim N(0, \sigma^2 I)$. Suppose that the underlying coefficients have support $S = \hat{\beta}(\beta_0)$, and $s_0 = |S|$. Then, the estimator given by least squares on S, i.e.,

$$\hat{\beta}_S^{\text{oracle}} = (X_S^T X_S)^{-1} X_S^T y,$$
$$\hat{\beta}_{-S}^{\text{oracle}} = 0.$$

is is called *oracle estimator*, and as we know from our previous calculations, has in-sample risk

$$\frac{1}{n}\|X\hat{\beta}^{\text{oracle}} - X\beta_0\|_2^2 = \sigma^2 \frac{s_0}{n}.$$

[8] consider this setup, and compare the risk of the best subset selection estimator $\hat{\beta}$ in (4) to the oracle risk of $\sigma^2 s_0/n$. They show that, if we choose $\lambda \asymp \sigma^2 \log p$, then the best subset selection estimator satisfies

$$\frac{\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \leq 4 \log p + 2 + o(1), \tag{8}$$

as $n, p \to \infty$. This holds without any conditions on the predictor matrix X. Moreover, they prove the lower bound

$$\inf_{\hat{\beta}} \sup_{X, \beta_0} \frac{\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \geq 2 \log p - o(\log p),$$

where the infimum is over all estimators $\hat{\beta}$, and the supremum is over all predictor matrices X and underlying coefficients with $\|\beta_0\|_0 = s_0$. Hence, in terms of rate, best subset selection achieves the optimal risk inflation over the oracle risk.

Returning to what was said above, the kicker is that we can't really compute the best subset selection estimator for even moderately-sized problems. As we will in the following, the lasso provides a similar risk inflation guarantee, though under considerably stronger assumptions.

Lastly, it is worth remarking that even if we *could* compute the subset selection estimator at scale, it's not at all clear that we would want to use this in place of the lasso. (Many people assume that we would.) We must remind ourselves that theory provides us an understanding of the performance of various estimators under typically idealized conditions, and it doesn't tell the complete story. It could be the case that the lack of shrinkage in the subset selection coefficients ends up being harmful in practical situations, in a signal-to-noise regime, and yet the lasso could still perform favorably in such settings.

Update. Some nice recent work in optimization [2] shows that we can cast best subset selection as a mixed integer quadratic program, and proposes to solve it (in general this means approximately, though with a certified bound on the duality gap) with an industry-standard mixed integer optimization package like Gurobi. However, in a recent paper, Hastie, Tibshirani and Tibshirani (arXiv:1707.08692) show that best subset selection does not do well statistically unless there is an extremely high signal to noise ratio.

5 Ridge Regression

We bring (3) and (6) here: Definition of Ridge estimator is

$$\beta^{\text{ridge}} = \text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2$$
$$\text{subject to } \sum_{k=1}^{p} \beta_k^2 \leq s,$$

or equivalently,

$$\beta^{\text{ridge}} = \text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2 + \lambda \sum_{k=1}^{p} \beta_k^2. \tag{9}$$

There is one-to-one correspondence between s and λ by Lagrange method. s (or λ) controls the complexity of the model, so the parameter s or λ is called the regularization parameter. If $s = 0$, the model includes only the intercept term while the model becomes the full model when $s = \infty$. The selection of this parameter is the same as model selection. We will learn it later.

The ridge estimator was proposed by [13] to resolve the problem of the least square estimator when $p > n$. Recall that the least square estimator is given as

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

where $X = (x_1, \ldots, x_n)^\top$ and $Y = (y_1, \ldots, y_n)^\top$. When $p > n$, $(X^\top X)^{-1}$ does not exist. The initial motivation of the ridge estimator is to replace $(X^\top X)^{-1}$ by $(X^\top X + \lambda I)^{-1}$. Note that the ridge estimator is the solution of (9).

We can extend the ridge estimator for logistic regression easily by

$$\beta^{\text{ridge}} = \text{argmin} \sum_{i=1}^{n} \left(y_i(\beta_0 + x_i^\top \beta) - \log(1 + \exp(\beta_0 + x_i^\top \beta)) \right)^2 + \lambda \sum_{k=1}^{p} \beta_k^2,$$

with the computation again using the Iteratively Reweighted Least Squares (IRLS) algorithm.

5.1 Bayesian justification

Assume that

$$Y = X\beta + \epsilon,$$

where

$$\epsilon \sim N_n(\mathbf{0}, \sigma^2 I).$$

A priori, we assume

$$\beta \sim N_p(\mathbf{0}, \tau^2 I).$$

Then we can easily see that the log posterior of β is equal to (up to constant)

$$\sum_{i=1}^{n} \frac{1}{2\sigma^2} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_k \beta_k \right)^2 + \frac{1}{2\tau^2} \sum_{k=1}^{p} \beta_k^2.$$

Hence, the ridge estimator is the maximum a posteriori (MAP) estimator with $\lambda = \tau^2/\sigma^2$. In fact, any estimators based on Bayesian methods are shirnkage estimators (shrinkage toward a prior).

6 LASSO

A disadvantage of the ridge regression is that the interpretation is not easy since all input variables are used. A question is whether we can do selection and shrinkage at the same time. Surprisingly, it is possible. The first of such methods is LASSO (Least Absolute Shrinkage and Selection Operator), firstly proposed by [15].

We bring (2) and (5) here: LASSO estimates β by

$$\begin{aligned} \beta^{\text{LASSO}} &= \text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik} \beta_k \right)^2 \\ &\text{subject to } \sum_{k=1}^{p} |\beta_k| \leq s, \end{aligned}$$

or equivalently,

$$\beta^{\text{LASSO}} = \text{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik} \beta_k \right)^2 + \lambda \sum_{k=1}^{p} |\beta_k| .$$

The only difference to ridge is the penalty function. We can say that the l_1 penalty is used in LASSO while the l_2 penalty is used in Ridge. This seemingly tiny difference makes qualitative gaps practically as well as theoretically. As we have seen above, one very interesting property of LASSO is that the predictive model is sparse (i.e. some coefficients are exactly 0).

We can see that a key property of sparse penalty function is that it is nondifferentiable around 0. That is, for sparse learning, we need to optimize a nondifferentiable objective function. Hence, standard numerical optimization methods such as gradient descent and Newton-Raphson can not be applied directly. Since LASSO was proposed firstly, optimization issue has been one of the hottest issues in statistics and machine learning society.

Roughly speaking, there are three algorithms, one based on the QP, the second based on angle, and the last one based on gradient descent.

The optimization problem of LASSO can be written as

$$\begin{aligned} &\text{minimize } \sum_{i=1}^{n} l(y_i, x_i^\top \beta) \\ &\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq s. \end{aligned}$$

When $l(y, a) = (y - a)^2$, this is a quadratic programming problem with linear constraints, and so we can apply any QP algorithm, which was done by [15].

Later, Osborne (2000a, 2000b), Efron et al. (2004) and Rosset and Zhu (2007) developed more efficient algorithms.

Now we turn to subgradient optimality (sometimes called the KKT conditions) for the lasso problem in (5). They tell us that any lasso solution $\hat{\beta}$ must satisfy

$$X^T(y - X\hat{\beta}) = \lambda s, \tag{10}$$

where $s \in \partial\|\hat{\beta}\|_1$, a subgradient of the ℓ_1 norm evaluated at $\hat{\beta}$. Precisely, this means that

$$s_j \in \begin{cases} \{+1\} & \hat{\beta}_j > 0 \\ \{-1\} & \hat{\beta}_j < 0 \\ [-1, 1] & \hat{\beta}_j = 0, \end{cases} \quad j = 1, \ldots, p. \tag{11}$$

From (10) we can read off a straightforward but important fact: even though the solution $\hat{\beta}$ may not be uniquely determined, the optimal subgradient s is a function of the unique fitted value $X\hat{\beta}$ (assuming $\lambda > 0$), and hence is itself unique.

Now from (11), note that the uniqueness of s implies that any two lasso solutions must have the same signs on the overlap of their supports. That is, it cannot happen that we find two different lasso solutions $\hat{\beta}$ and $\tilde{\beta}$ with $\hat{\beta}_j > 0$ but $\tilde{\beta}_j < 0$ for some j, and hence we have no problem interpretating the signs of components of lasso solutions.

Let's assume henceforth that the columns of X are in general position (and we are looking at a nontrivial end of the path, with $\lambda > 0$), so the lasso solution $\hat{\beta}$ is unique. Let $A = \text{supp}(\hat{\beta})$ be the lasso active set, and let $s_A = \text{sign}(\hat{\beta}_A)$ be the signs of active coefficients. From the subgradient conditions (10), (11), we know that

$$X_A^T(y - X_A\hat{\beta}_A) = \lambda s_A,$$

and solving for $\hat{\beta}_A$ gives

$$\begin{aligned} \hat{\beta}_A &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A), \\ \hat{\beta}_{-A} &= 0 \end{aligned} \tag{12}$$

(where recall we know that $X_A^T X_A$ is invertible because X has columns in general position). We see that the active coefficients $\hat{\beta}_A$ are given by taking the least squares coefficients on X_A, $(X_A^T X_A)^{-1} X_A^T y$, and shrinking them by an amount $\lambda(X_A^T X_A)^{-1} s_A$. Contrast this to, e.g., the subset selection solution in (7), where there is no such shrinkage.

Now, how about this so-called shrinkage term $(X_A^T X_A)^{-1} X_A^T y$? Does it always act by moving each one of the least squares coefficients $(X_A^T X_A)^{-1} X_A^T y$ towards zero? Indeed, this is not always the case, and one can find empirical examples where a lasso coefficient is actually larger (in magnitude) than the corresponding least squares coefficient on the active set. Of course, we also know that this is due to the correlations between active variables, because when X is orthogonal, as we've already seen, this never happens.

On the other hand, it is always the case that the lasso solution has a strictly smaller ℓ_1 norm than the least squares solution on the active set, and in this sense, we are (perhaps) justified in always referring to $(X_A^T X_A)^{-1} X_A^T y$ as a shrinkage term. To see this, note that, for any vector b, $||b||_1 = s^T b$ where s is the vector of signs of b. So $||\hat{\beta}||_1 = s^T\hat{\beta} = s_A^T\hat{\beta}_A$ and so

$$\|\hat{\beta}\|_1 = s_A^T(X_A^T X_A)^{-1} X_A^T y - \lambda s_A^T(X_A^T X_A)^{-1} s_A < \|(X_A^T X_A)^{-1} X_A^T y\|_1. \tag{13}$$

The first term is less than or equal to $\|(X_A^T X_A)^{-1} X_A^T y\|_1$, and the term we are subtracting is strictly negative (because $(X_A^T X_A)^{-1}$ is positive definite).

7 Theoretical analysis of the lasso

7.1 Slow rates

There has been an enormous amount theoretical work analyzing the performance of the lasso. Some references (warning: a highly incomplete list) are [10, 9, 7, 5, 14, 17, 4, 16]; a helpful text for these kind of results is [3].

We begin by stating what are called *slow rates* for the lasso estimator. Most of the proofs are simple enough that they are given below. These results don't place any real assumptions on the predictor matrix X, but deliver slow(er) rates for the risk of the lasso estimator than what we would get under more assumptions, hence their name.

We will assume the standard linear model with X fixed, and $\epsilon \sim N(0, \sigma^2)$. We will also assume that $\|X_j\|_2^2 \leq n$, for $j = 1, \ldots, p$. That the errors are Gaussian can be easily relaxed to sub-Gaussianity.

The lasso estimator in bound form (2) is particularly easy to analyze. Suppose that we choose $s = \|\beta_0\|_1$ as the tuning parameter. Then, simply by virtue of optimality of the solution in (2), we find that

$$\|y - X\hat{\beta}\|_2^2 \leq \|y - X\beta_0\|_2^2,$$

or, expanding and rearranging,

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\langle \epsilon, X\hat{\beta} - X\beta_0 \rangle.$$

Here we denote $\langle a, b \rangle = a^T b$. The above is sometimes called the *basic inequality* (for the lasso in bound form). Now, rearranging the inner product, using Holder's inequality, and recalling the choice of bound parameter:

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\langle X^T\epsilon, \hat{\beta} - \beta_0 \rangle \leq 4\|\beta_0\|_1 \|X^T\epsilon\|_\infty.$$

Notice that $\|X^T\epsilon\|_\infty = \max_{j=1,\ldots,p} |X_j^T\epsilon|$ is a maximum of p Gaussians, each with mean zero and variance upper bounded by $\sigma^2 n$. By a standard maximal inequality for Gaussians, for any $\delta > 0$,

$$\max_{j=1,\ldots,p} |X_j^T\epsilon| \leq \sigma\sqrt{2n\log(ep/\delta)},$$

with probability at least $1-\delta$. Plugging this to the second-to-last display and dividing by n, we get the finite-sample result for the lasso estimator

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq 4\sigma\|\beta_0\|_1\sqrt{\frac{2\log(ep/\delta)}{n}}, \tag{14}$$

with probability at least $1 - \delta$.

The high-probability result (14) implies an in-sample risk bound of

$$\frac{1}{n}\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2 \lesssim \|\beta_0\|_1\sqrt{\frac{\log p}{n}}.$$

Compare to this with the risk bound (8) for best subset selection, which is on the (optimal) order of $s_0 \log p/n$ when β_0 has s_0 nonzero components. If each of the nonzero components here has constant magnitude, then above risk bound for the lasso estimator is on the order of $s_0\sqrt{\log p/n}$, which is much slower.

Predictive risk. Instead of in-sample risk, we might also be interested in out-of-sample risk, as after all that reflects actual (out-of-sample) predictions. In least squares, recall, we saw that out-of-sample risk was generally higher than in-sample risk. The same is true for the lasso [**?**] gives a nice, simple analysis of out-of-sample risk for the lasso. He assumes that x_0, x_i, $i = 1, \ldots, n$ are i.i.d. from an arbitrary distribution supported on a compact set in ^{p}, and shows that the lasso estimator in bound form (2) with $t = \|\beta_0\|_1$ has out-of-sample risk satisfying

$$\mathbb{E}(x_0^T\hat{\beta} - x_0^T\beta)^2 \lesssim \|\beta_0\|_1^2\sqrt{\frac{\log p}{n}}.$$

The proof is not much more complicated than the above, for the in-sample risk, and reduces to a clever application of Hoeffding's inequality, though we omit it for brevity. Note here the dependence on $\|\beta_0\|_1^2$, rather than $\|\beta_0\|_1$ as in the in-sample risk. This agrees with the analysis we did in the previous set of notes where we did not assume the linear model. (Only the interpretation changes.)

Oracle inequality. If we don't want to assume linearity of the mean then we can still derive an *oracle inequality* that characterizes the risk of the lasso estimator in excess of the risk of the best linear predictor. For this part only, assume the more general model

$$y = \mu(X) + \epsilon,$$

with an arbitrary mean function $\mu(X)$, and normal errors $\epsilon \sim N(0, \sigma^2)$. We will analyze the bound form lasso estimator (2) for simplicity. By optimality of $\hat{\beta}$, for any other $\tilde{\beta}$ feasible for the lasso problem in (2), it holds that[1]

$$\langle X^T(y - X\hat{\beta}), \tilde{\beta} - \hat{\beta} \rangle \leq 0. \tag{15}$$

Rearranging gives

$$\langle \mu(X) - X\hat{\beta}, X\tilde{\beta} - X\hat{\beta} \rangle \leq \langle X^T\epsilon, \hat{\beta} - \tilde{\beta} \rangle.$$

Now using the polarization identity $\|a\|_2^2 + \|b\|_2^2 - \|a - b\|_2^2 = 2\langle a, b \rangle$,

$$\|X\hat{\beta} - \mu(X)\|_2^2 + \|X\hat{\beta} - X\tilde{\beta}\|_2^2 \leq \|X\tilde{\beta} - \mu(X)\|_2^2 + 2\langle X^T\epsilon, \hat{\beta} - \tilde{\beta} \rangle,$$

and from the exact same arguments as before, it holds that

$$\frac{1}{n}\|X\hat{\beta} - \mu(X)\|_2^2 + \frac{1}{n}\|X\hat{\beta} - X\tilde{\beta}\|_2^2 \leq \frac{1}{n}\|X\tilde{\beta} - \mu(X)\|_2^2 + 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}},$$

with probability at least $1 - \delta$. This holds simultaneously over all $\tilde{\beta}$ with $\|\tilde{\beta}\|_1 \leq s$. Thus, we may write, with probability $1 - \delta$,

$$\frac{1}{n}\|X\hat{\beta} - \mu(X)\|_2^2 \leq \left\{ \inf_{\|\tilde{\beta}\|_1 \leq s} \frac{1}{n}\|X\tilde{\beta} - \mu(X)\|_2^2 \right\} + 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}}.$$

Also if we write $X\tilde{\beta}^{\text{best}}$ as the best linear that predictor of ℓ_1 at most s, achieving the infimum on the right-hand side (which we know exists, as we are minimizing a continuous function over a compact set), then

$$\frac{1}{n}\|X\hat{\beta} - X\tilde{\beta}^{\text{best}}\|_2^2 \leq 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}},$$

with probability at least $1 - \delta$

7.2 Fast rates

Under **very** strong assumptions we can get faster rates. For example, if we assume that X satisfies the *restricted eigenvalue condition* with constant $\phi_0 > 0$, i.e.,

$$\frac{1}{n}\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2 \quad \text{for all subsets } J \subseteq \{1, \ldots, p\} \text{ such that } |J| = s_0$$
$$\text{and all } v \in \mathbb{R}^p \text{ such that } \|v_{J^c}\|_1 \leq 3\|v_J\|_1 \tag{16}$$

then

$$\|\hat{\beta} - \beta_0\|_2^2 \lesssim \frac{s_0 \log p}{n\phi_0^2} \tag{17}$$

with probability tending to 1. (This condition can be slightly weakened, but not much.) The condition is unlikely to hold in any real problem. Nor is it checkable.

7.3 Support recovery

Here we discuss results on support recovery of the lasso estimator. There are a few versions of support recovery results and again [3] is a good place to look for a thorough coverage. Here we describe a result due to [16], who introduced a proof technique called the *primal-dual witness method*. The assumptions are even stronger (and less believable) than in the previous section. In addition to the previous assumptions we need:

Mutual incoherence: for some $\gamma > 0$, we have

$$\|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \leq 1 - \gamma, \quad \text{for } j \notin S,$$

Minimum eigenvalue: for some $C > 0$, we have

$$\Lambda_{\min}\left(\frac{1}{n}X_S^T X_S\right) \geq C,$$

[1] To see this, consider minimizing a convex function $f(x)$ over a convex set C. Let $\hat{x}$ be a minimizer. Let $z \in C$ be any other point in C. If we move away from the solution $\hat{x}$ we can only increase $f(\hat{x})$. In other words, $\langle \nabla f(\hat{x}), z - \hat{z} \rangle \geq 0$.

where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A

Minimum signal:

$$\beta_{0,\min} = \min_{j \in S} |\beta_{0,j}| \geq \lambda \|(X_S^T X_S)^{-1}\|_\infty + \frac{4\gamma\lambda}{\sqrt{C}},$$

where $\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{q} |A_{ij}|$ denotes the ℓ_∞ norm of an $m \times q$ matrix A

Under these assumptions, once can show that, if λ is chosen just right, then

$$P(\text{support}(\hat{\beta}) = \text{support}(\beta)) \to 1. \tag{18}$$

References

[1] E. M. L. Beale, M. G. Kendall, and D. W. Mann. The discarding of variables in multivariate analysis. *Biometrika*, 54(3/4):357–366, 1967.

[2] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

[3] Peter Buhlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

[4] Emmanuel J. Candes and Yaniv Plan. Near ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5):2145–2177, 2009.

[5] Emmanuel J. Candes and Terence Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[6] Scott Chen, David L. Donoho, and Michael Saunders. Atomic decomposition for basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[7] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(12):1289–1306, 2006.

[8] Dean Foster and Edward George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975, 1994.

[9] Jean Jacques Fuchs. Recovery of exact sparse representations in the presense of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

[10] Eitan Greenshtein and Ya'Acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[12] R. R. Hocking and R. N. Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967.

[13] Arthur Hoerl and Robert Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[14] Nicolai Meinshausen and Peter Buhlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[15] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[16] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[17] Peng Zhao and Bi Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2564, 2006.