Shrinkage methods

김지수 (Jisu KIM)

통계적 기계학습(Statistical Machine Learning), 2025 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 [13].

1 Review

1.1 Basic Model for Supervised Learning

- Input(입력) / Covariate(설명 변수) : $x \in \mathbb{R}^d$, so $x = (x_1, \ldots, x_d)$.
- 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) :

  $$y \approx f(x), \qquad f \in \mathcal{M}.$$

  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 $\mathcal{M}$. 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. The most common ones are:
  - square: $\ell(y, a) = (y - a)^2$.
  - $0 - 1$: $\ell(y, a) = I(y \neq a)$.
- 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,

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

  Here, $\mathcal{F}$ can be different from $\mathcal{M}$; $\mathcal{F}$ can be smaller than $\mathcal{M}$.
- 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

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

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

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

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

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

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

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

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:

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

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

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:

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

(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, 14], ridge regression to [15], and the lasso to [19, 7]

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 $\hat{\beta}$ also solves (1), but the converse is not true.

3.2 A Toy Example

It is helpful to first consider a toy example. Let $Y = (Y_1, \ldots, Y_d)^\top \in \mathbb{R}^d$ be a random sample from $N_d(\mu, \sigma^2 I)$.

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

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

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

is better than Y. Note that $\left\|\delta^{JS}\right\| \leq \|Y\|$. The expected prediction risk of Y for L_2 loss is

$$\mathbb{E}\left[\|Y - \mu\|_2^2\right] = d\sigma^2,$$

while the expected prediction risk of δ^{JS} for L_2 loss is

$$\mathbb{E}\left[\left\|\delta^{JS} - \mu\right\|_2^2\right] = d\sigma^2 - (d-2)^2\sigma^2\mathbb{E}\left[\frac{1}{\|Y\|_2^2}\right].$$

See TPE, Example 7.1. So the 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:

$$\frac{1}{2}\|Y-\mu\|_2^2+\lambda\|\mu\|_0\,,\quad \frac{1}{2}\|Y-\mu\|_2^2+\lambda\|\mu\|_1\,,\quad \frac{1}{2}\|Y-\mu\|_2^2+\lambda\|\mu\|_2^2\,.$$

You should verify that the solutions can be obtained coordinate-wise, and are given as

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

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 $j\in\{1,\dots,d\}$. 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

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

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

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^Ty) 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 $p\le 1$ and convexity requires $p\ge 1$. The only norm that gives sparsity and convexity is $p=1$. The appendix has a brief review of convexity.

4 Variable Selection

- For given $k\le d$, 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.

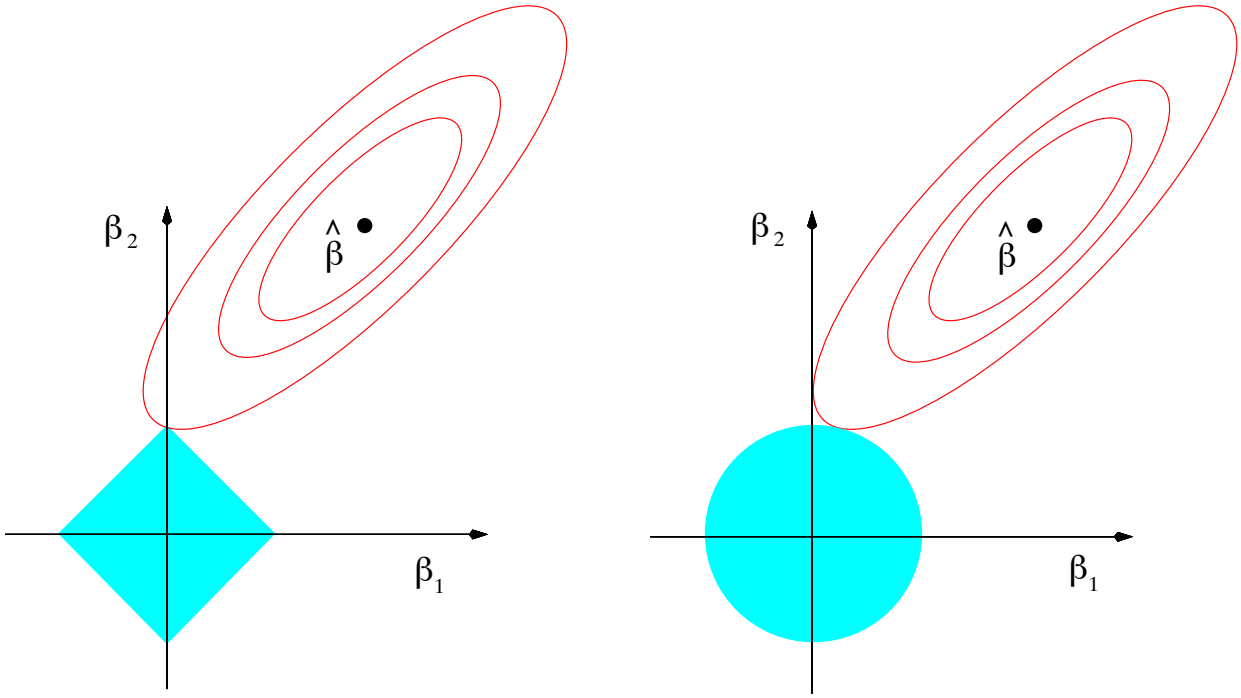


Figure 1: *The "classic" illustration comparing lasso and ridge constraints. From Figure 3.11 of* [13].

- Select the optimal model among $M_0, \ldots, M_d$. (we will see this later in Model selection.)
- The complexity of the model is proportional to k.
- If d 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 [10], 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 $\hat{\beta}$ in (4)—meaning that $\hat{\beta}_j = 0$ for all $j \notin A$, and denoted $A = \mathrm{supp}(\hat{\beta})$—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

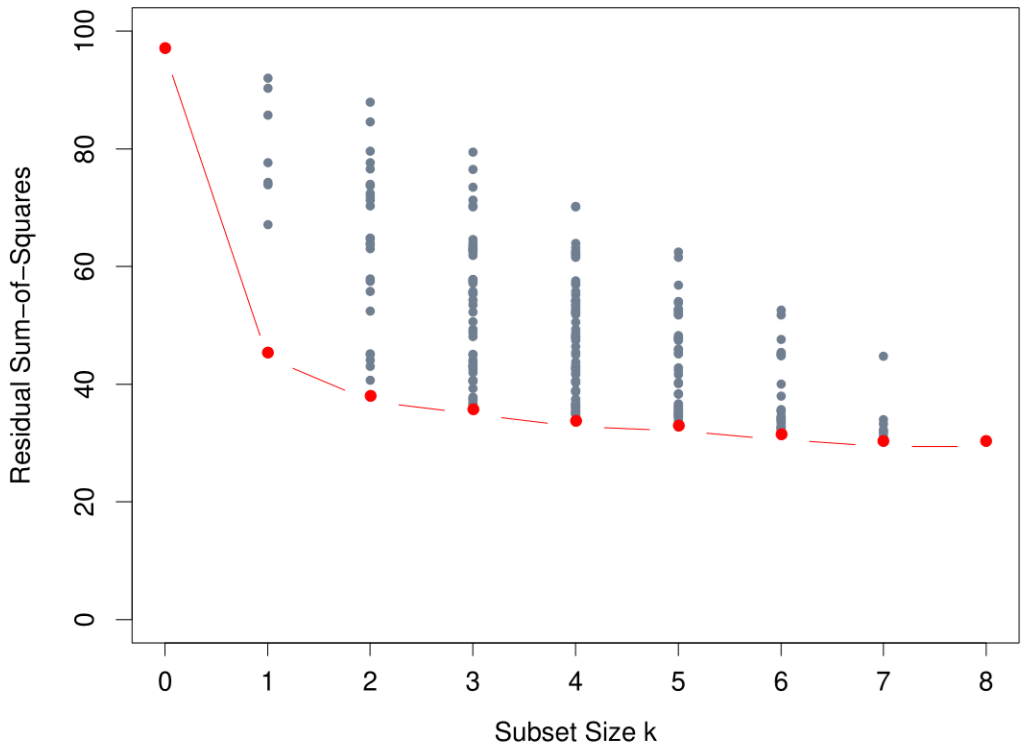


Figure 2: All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size. Figure 3.5 from [13].

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.

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

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

[10] 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 $\lambda \asymp \sigma^2 \log d$, then the best subset selection estimator satisfies

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

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

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

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 (능선회귀)

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

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

or equivalently,

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

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 (능선추정량) was proposed by [15] 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 $d > n$, $(X^\top X)^{-1}$ 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 (로지스틱 회귀) easily by

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

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

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

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

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

Hence, the ridge estimator (능선추정량) 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 (라쏘)

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) (라쏘), firstly proposed by [19].

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

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

or equivalently,

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

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 (라쏘) can be written as

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

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 [19].

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

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

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 (라쏘). Some references (warning: a highly incomplete list) are [12, 11, 8, 5, 16, 22, 4, 21]; a helpful text for these kind of results is [3].

We begin by stating what are called *slow rates* for the lasso estimator (라쏘추정량). 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.

First, note the maximal Gaussian inequality:

Lemma. *Suppose we have random variables $X_1, \ldots, X_d$ with $X_j \sim N(0, \sigma^2)$, but X_j's are not necessarily independent. Then*

$$P\left(\max_{i=1,\ldots,d} |X_j| \geq t\right) \leq 2d \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

or equivalently, with probability $1-\delta$,

$$\max_{i=1,\ldots,p} |X_j| \leq \sigma\sqrt{2\log(2d/\delta)}.$$

The Gaussian condition $X_j \sim N(0, \sigma^2)$ can be relaxed to the subgaussian condition.

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 $j = 1, \ldots, d$. That the errors are Gaussian can be easily relaxed to sub-Gaussianity.

Theorem. *Suppose the linear model $Y = X\beta_0 + \epsilon$ with ϵ_i i.i.d. from $N(0, \sigma^2)$. If we choose $s = \|\beta_0\|_1$ as the tuning parameter, then*

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \lesssim \frac{\|\beta_0\|_1 \|X^T\epsilon\|_\infty}{n}.$$

Suppose $\|X_j\|_2 = \sqrt{n}$. Then with high probability,

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

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 $\hat{\beta}$ in (2), we find that

$$\|y - X\hat{\beta}\|_2^2 \leq \|y - X\beta_0\|_2^2 = \|\epsilon\|_2^2.$$

By developing squares,

$$\begin{aligned}\|y - X\hat{\beta}\|_2^2 &= \left\|X\beta_0 + \epsilon - X\hat{\beta}\right\|_2^2 \\ &= \left\|X(\hat{\beta} - \beta_0)\right\|_2^2 + \|\epsilon\|_2^2 - 2\left\langle \epsilon, X(\hat{\beta} - \beta_0)\right\rangle,\end{aligned}$$

and combining gives

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

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 $\delta > 0$,

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

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

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

with probability at least $1-\delta$.

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

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

Compare to this with the risk bound (8) for best subset selection, which is on the (optimal) order of $s_0 \log d/n$ 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 $s_0\sqrt{\log d/n}$, 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 [6] 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 $\mathbb{R}^p$, and shows that the lasso estimator in bound form (2) with $t = \|\beta_0\|_1$ has out-of-sample risk satisfying

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

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

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

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 $1-\delta$,

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

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

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

with probability at least $1-\delta$

[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, $\langle \nabla f(\hat{x}), z - \hat{z}\rangle \geq 0$.

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.,

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

Theorem. *Suppose $Y = X\beta_0 + \epsilon$ with X fixed, ϵ_i i.i.d. from $N(0,\sigma^2)$, with $\|\beta_0\|_0 = s_0$. Suppose X satisfies the restricted eigenvalue condition in (16). If $\lambda \geq 2\left\|X^\top\epsilon\right\|_\infty$, then*

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \lesssim \frac{s_0\lambda^2}{n^2\phi_0^4},$$

and

$$\|\hat{\beta} - \beta_0\|_2^2 \lesssim \frac{s_0\lambda^2}{n^2\phi_0^4}.$$

Suppose $\|X_j\|_2 = \sqrt{n}$, and λ is chosen as $\lambda = 2\sigma\sqrt{2n\log(2d/\delta)}$. Then with high probability,

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \lesssim \frac{s_0\log(d)}{n\phi_0^2},$$

and

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

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.

We may interpret the restricted eigenvalue condition roughly as follows: the requirement $(1/n)\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2$ for all $v \in \mathbb{R}^n$ would be a lower bound of ϕ_0^2 on the smallest eigenvalue of $(1/n)X^TX$; we don't require this (as this would of course mean that X was full column rank, and couldn't happen when $d > n$), but instead that require that the same inequality hold for v that are "mostly" supported on small subsets J of variables, with $|J| = s_0$

Here is a proof of the theorem. There are many flavors of fast rates, and the conditions required are all very closely related. [20] provides a nice review and discussion.

For the lasso of the penalized form, we have

$$\|y - X\hat{\beta}\|_2^2 + 2\lambda\left\|\hat{\beta}\right\|_1 \leq \|y - X\beta_0\|_2^2 + 2\lambda\|\beta_0\|_1.$$

Hence for this, the basic inequality becomes

$$\begin{aligned}\|X\hat{\beta} - X\beta_0\|_2^2 &\leq 2\langle\epsilon, X^\top(\hat{\beta} - \beta_0)\rangle + 2\lambda\left(\|\beta_0\|_1 - \left\|\hat{\beta}\right\|_1\right) \\ &\leq 2\left\|X^\top\epsilon\right\|_\infty\left\|\hat{\beta} - \beta_0\right\|_1 + 2\lambda\left(\|\beta_0\|_1 - \left\|\hat{\beta}\right\|_1\right).\end{aligned}$$

Then using the condition $\lambda \geq 2\left\|X^\top\epsilon\right\|_\infty$ gives

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq \lambda\|\hat{\beta} - \beta_0\|_1 + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1).$$

Let $\hat{\Delta} = \hat{\beta} - \beta_0$ for convenience. Then since $\beta_{0,-S} = 0$,

$$\begin{aligned}\|\beta_0\|_1 - \|\hat{\beta}\|_1 &= \|\beta_{0,S}\|_1 - \|\beta_{0,S} + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{-S}\|_1 \\ &\leq \|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{-S}\|_1,\end{aligned}$$

where the inequality followed from the triangle inequality. Hence

$$\begin{aligned}\|X\hat{\beta} - X\beta_0\|_2^2 &\leq \lambda\|\hat{\beta} - \beta_0\|_1 + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \\ &\leq \lambda\left(\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{-S}\|_1\right) + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{-S}\|_1) \\ &= 3\lambda\|\hat{\Delta}_S\|_1 - \lambda\|\hat{\Delta}_{-S}\|_1,\end{aligned}$$

As $\|X\hat{\beta} - X\beta_0\|_2^2 \geq 0$, we have shown

$$\|\hat{\Delta}_{-S}\|_1 \leq 3\|\hat{\Delta}_S\|_1,$$

and thus we may apply the restricted eigenvalue condition (**??**) to the vector $\hat{\Delta} = \hat{\beta} - \beta_0$. This gives us two bounds: one on the fitted values, and the other on the coefficients. Both start with the key inequality (from the second-to-last display)

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 3\lambda\|\hat{\Delta}_S\|_1. \tag{18}$$

For the fitted values, we upper bound the right-hand side of the key inequality (18) with

$$\|\hat{\Delta}_S\|_1 \leq \sqrt{s_0}\|\hat{\Delta}_S\|_2 \leq \sqrt{s_0}\|\hat{\Delta}\|_2,$$

to get

$$\begin{aligned}\|X\hat{\beta} - X\beta_0\|_2^2 &\leq 3\lambda\sqrt{s_0}\|\hat{\beta} - \beta_0\|_2 \\ &\leq 3\lambda\sqrt{\frac{s_0}{n\phi_0^2}}\|X\hat{\beta} - X\beta_0\|_2.\end{aligned}$$

And dividing through both sides by $\|X\hat{\beta} - X\beta_0\|_2$, then squaring both sides, and dividing by n,

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{9s_0\lambda^2}{n^2\phi_0^2}.$$

Now, we have seen that when $\|X_j\|_2 = \sqrt{n}$, with high probability, $\left\|X^\top\epsilon\right\|_\infty \leq \sigma\sqrt{2n\log(2d/\delta)}$ of probability at least $1-\delta$. Hence by choosing $\lambda = 2\sigma\sqrt{2n\log(2d/\delta)}$, with probability $1-\delta$, we have that

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{72\sigma^2 s_0\log(2d/\delta)}{n\phi_0^2}, \tag{19}$$

with probability at least $1-\delta$. Notice the similarity between (19) and (8): both provide us in-sample risk bounds on the order of $s_0\log p/n$, but the bound for the lasso requires a strong compability assumption on the predictor matrix X, which roughly means the predictors can't be too correlated

For the coefficients, we have from the restricted eigenvalue condition that

$$\begin{aligned}\|\hat{\beta} - \beta_0\|_2^2 &\leq \frac{1}{n\phi_0^2}\left\|X\hat{\beta} - X\beta_0\right\|_2^2 \\ &\leq \frac{9s_0\lambda^2}{n^2\phi_0^4}.\end{aligned}$$

Plugging in $\lambda = 2\sigma\sqrt{2n\log(ed/\delta)}$, we have shown that

$$\|\hat{\beta} - \beta_0\|_2^2 \leq \frac{72\sigma^2 s_0\log(2d/\delta)}{n\phi_0^4}, \tag{20}$$

with probability at least $1-\delta$. This is an error bound on the order of $s_0\log d/n$ for the lasso coefficients.

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 [21], 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

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

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

$$P(\text{support}(\hat{\beta}) = \text{support}(\beta)) \to 1. \tag{21}$$

8 Geometry of the solutions

One undesirable feature of the best subset selection solution (7) is the fact that it behaves discontinuously with y. As we change y, the active set A must change at some point, and the coefficients will jump discontinuously, because we are just doing least squares onto the active set. So, does the same thing happen with the lasso solution (12)? The answer it not immediately clear. Again, as we change y, the active set A must change at some point; but if the shrinkage term were defined "just right", then perhaps the coefficients of variables to leave the active set would gracefully and continously drop to zero, and coefficients of variables to enter the active set would continuously move form zero. This would make whole the lasso solution continuous. Fortuitously, this is indeed the case, and the lasso solution $\hat{\beta}$ is continuous as a function of y. It might seem a daunting task to prove this, but a certain perspective using convex geometry provides a very simple proof. The geometric perspective in fact proves that the lasso fit $X\hat{\beta}$ is nonexpansive in y, i.e., 1-Lipschitz continuous, which is a very strong form of continuity. Define the convex polyhedron $C = \{u : \|X^T u\|_\infty \leq \lambda\} \subseteq \mathbb{R}^n$. The dual problem for the penalized lasso

$$\text{minimize}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{22}$$

can be obtained using convex optimization as

$$\begin{aligned} &\text{minimize}_{u \in \mathbb{R}^n} \|y - u\|_2^2 \\ &\text{subject to } \left\|X^\top u\right\|_\infty \leq \lambda. \end{aligned} \tag{23}$$

And from the KKT conditions, the solution $\hat{\beta}$ of (22) and the solution $\hat{u}$ of (23) are linked as

$$X\hat{\beta} = y - \hat{u}.$$

Hence the lasso solution $\hat{\beta}$ satisfies

$$X\hat{\beta} = (I - P_C)(y),$$

the residual from projecting y onto C. A picture to show this (just look at the left panel for now) is given in Figure 3.

The projection onto any convex set is nonexpansive, i.e., $\|P_C(y) - P_C(y')\|_2 \leq \|y - y'\|_2$ for any y, y'. This should be visually clear from the picture. Actually, the same is true with the residual map: $I - P_C$ is also nonexpansive, and hence the lasso fit is 1-Lipschitz continuous. Viewing the lasso fit as the residual from projection onto a convex polyhedron is actually an even more fruitful perspective. Write this polyhedron as

$$C = (X^T)^{-1}\{v : \|v\|_\infty \leq \lambda\},$$

where $(X^T)^{-1}$ denotes the preimage operator under the linear map X^T. The set $\{v : \|v\|_\infty \leq \lambda\}$ is a hypercube in $\mathbb{R}^p$. Every face of this cube corresponds to a subset $A \subseteq \{1, \ldots p\}$ of dimensions (that achieve the maximum value $|\lambda|$) and signs $s_A \in \{-1, 1\}^{|A|}$ (that tell which side of the cube the face will lie on, for each dimension). Now, the faces of C are just faces of $\{v : \|v\|_\infty \leq \lambda\}$ run through the (linear) preimage transformation, so each face of C can also indexed by a set $A \subseteq \{1, \ldots p\}$ and signs $s_A \in \{-1, 1\}^{|A|}$. The picture in Figure 3 attempts to convey this relationship with the colored black face in each of the panels.

Now imagine projecting y onto C; it will land on some face. We have just argued that this face corresponds to a set A and signs s_A. One can show that this set A is exactly the active set of the lasso solution at y, and s_A are exactly the active signs. The size of the active set $|A|$ is the co-dimension of the face. Looking at the picture: we can that see that as we wiggle y around, it will project to the same face. From the correspondence between faces and active set and signs of lasso solutions, this means that A, s_A do not change as we perturb y, i.e., they are locally

constant. But this isn't true for all points y, e.g., if y lies on one of the rays emanating from the lower right corner of the polyhedron in the picture, then we can see that small perturbations of y do actually change the face that it projects to, which invariably changes the active set and signs of the lasso solution. However, this is somewhat of an exceptional case, in that such points can be form a of Lebesgue measure zero, and therefore we can assure ourselves that the active set and signs A, s_A are locally constant for almost every y.

From the lasso KKT conditions (10), (11), it is possible to compute the lasso solution in (5) as a function of λ, which we will write as $\hat{\beta}(\lambda)$, for all values of the tuning parameter $\lambda \in [0, \infty]$. This is called the *regularization path* or *solution path* of the problem (5). Path algorithms like the one we will describe below are not always possible; the reason that this ends up being feasible for the lasso problem (5) is that the solution path $\hat{\beta}(\lambda)$, $\lambda \in [0, \infty]$ turns out to be a piecewise linear, continuous function of λ. Hence, we only need to compute and store the *knots* in this path, which we will denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$, and the lasso solution at these knots. From this information, we can then compute the lasso solution at any value of λ by linear interpolation.

The knots $\lambda_1 \geq \ldots \geq \lambda_r$ in the solution path correspond to λ values at which the active set $A(\lambda) = \mathrm{supp}(\hat{\beta}(\lambda))$ changes. As we decrease λ from ∞ to 0, the knots typically correspond to the point at which a variable enters the active set; this connects the lasso to an incremental variable selection procedure like forward stepwise regression. Interestingly though, as we decrease λ, a knot in the lasso path can also correspond to the point at which a variables leaves the active set. See Figure 4.

The lasso solution path was described by [17, 18, 9]. Like the construction of all other solution paths that followed these seminal works, the lasso path is essentially given by an iterative or inductive verification of the KKT conditions; if we can maintain that the KKT conditions holds as we decrease λ, then we know we have a solution. The trick is to start at a value of λ at which the solution is trivial; for the lasso, this is $\lambda = \infty$, at which case we know the solution must be $\hat{\beta}(\infty) = 0$.

Why would the path be piecewise linear? The construction of the path from the KKT conditions is actually rather technical (not difficult conceptually, but somewhat tedious), and doesn't shed insight onto this matter. But we can actually see it clearly from the projection picture in Figure 3.

As λ decreases from ∞ to 0, we are shrinking (by a multiplicative factor λ) the polyhedron onto which y is projected; let's write $C_\lambda = \{u : \|X^T u\|_\infty \leq \lambda\} = \lambda C_1$ to make this clear. Now suppose that y projects onto the relative interior of a certain face F of C_λ, corresponding to an active set A and signs s_A. As λ decreases, the point on the boundary of C_λ onto which y projects, call it $\hat{u}(\lambda) = P_{C_\lambda}(y)$, will move along the face F, and change linearly in λ (because we are equivalently just tracking the projection of y onto an affine space that is being scaled by λ). Thus, the lasso fit $X\hat{\beta}(\lambda) = y - \hat{u}(\lambda)$ will also behave linearly in λ. Eventually, as we continue to decrease λ, the projected point $\hat{u}(\lambda)$ will move to the relative boundary of the face F; then, decreasing λ further, it will lie on a different, neighboring face F'. This face will correspond to an active set A' and signs $s_{A'}$ that (each) differ by only one element to A and s_A, respectively. It will then move linearly across F', and so on.

Now we will walk through the technical derivation of the lasso path, starting at $\lambda = \infty$ and $\hat{\beta}(\infty) = 0$, as indicated above. Consider decreasing λ from ∞, and continuing to set $\hat{\beta}(\lambda) = 0$ as the lasso solution. The KKT conditions (10) read

$$X^T y = \lambda s,$$

where s is a subgradient of the ℓ_1 norm evaluated at 0, i.e., $s_j \in [-1, 1]$ for every $j = 1, \ldots, p$. For large enough values of λ, this is satisfied, as we can choose $s = X^T y/\lambda$. But this ceases to be a valid subgradient if we decrease λ past the point at which $\lambda = |X_j^T y|$ for some variable $j = 1, \ldots, p$. In short, $\hat{\beta}(\lambda) = 0$ is the lasso solution for all $\lambda \geq \lambda_1$, where

$$\lambda_1 = \max_{j=1,\ldots,d} |X_j^T y|. \tag{24}$$

What happens next? As we decrease λ from λ_1, we know that we're going to have to change $\hat{\beta}(\lambda)$ from 0 so that the KKT conditions remain satisfied. Let j_1 denote the variable that achieves the maximum in (24). Since the subgradient was $|s_{j_1}| = 1$ at $\lambda = \lambda_1$, we see that we are "allowed" to make $\hat{\beta}_{j_1}(\lambda)$ nonzero. Consider setting

$$\begin{aligned} \hat{\beta}_{j_1}(\lambda) &= (X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \\ \hat{\beta}_j(\lambda) &= 0, \quad \text{for all } j \neq j_1, \end{aligned} \tag{25}$$

as λ decreases from λ_1, where $s_{j_1} = \mathrm{sign}(X_{j_1}^T y)$. Note that this makes $\hat{\beta}(\lambda)$ a piecewise linear and continuous function of λ, so far. The KKT conditions are then

$$X_{j_1}^T \Big(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \Big) = \lambda s_{j_1},$$

which can be checked with simple algebra, and

$$\Big|X_j^T\Big(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1})\Big)\Big| \leq \lambda,$$

for all $j \neq j_1$. Recall that the above held with strict inequality at $\lambda = \lambda_1$ for all $j \neq j_1$, and by continuity of the constructed solution $\hat{\beta}(\lambda)$, it should continue to hold as we decrease λ for at least a little while. In fact, it will hold until one of the piecewise linear paths

$$X_j^T(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1})), \quad j \neq j_1$$

becomes equal to $\pm\lambda$, at which point we have to modify the solution because otherwise the implicit subgradient

$$s_j = \frac{X_j^T(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}))}{\lambda}$$

will cease to be in $[-1, 1]$. It helps to draw yourself a picture of this.

Thanks to linearity, we can compute the critical "hitting time" explicitly; a short calculation shows that, the lasso solution continues to be given by (25) for all $\lambda_1 \geq \lambda \geq \lambda_2$, where

$$\lambda_2 = \max +_{j \neq j_1,\, s_j \in \{-1,1\}} \frac{X_j^T(I - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}X_{j_1})y}{s_j - X_j^T X_{j_1}(X_{j_1}^T X_{j_1})^{-1}s_{j_1}}, \tag{26}$$

and $\max +$ denotes the maximum over all of its arguments that are $< \lambda_1$.

To keep going: let j_2, s_2 achieve the maximum in (26). Let $A = \{j_1, j_2\}$, $s_A = (s_{j_1}, s_{j_2})$, and consider setting

$$\begin{aligned} \hat{\beta}_A(\lambda) &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A) \\ \hat{\beta}_{-A}(\lambda) &= 0, \end{aligned} \tag{27}$$

as λ decreases from λ_2. Again, we can verify the KKT conditions for a stretch of decreasing λ, but will have to stop when one of

$$X_j^T(y - X_A(X_A^T X_A)^{-1}(X_A^T y - \lambda s_A), \quad j \notin A$$

becomes equal to $\pm\lambda$. By linearity, we can compute this next "hitting time" explicitly, just as before. Furthermore, though, we will have to check whether the active components of the computed solution in (27) are going to cross through zero, because past such a point, s_A will no longer be a proper subgradient over the active components. We can again compute this next "crossing time" explicitly, due to linearity. Therefore, we maintain that (27) is the lasso solution for all $\lambda_2 \geq \lambda \geq \lambda_3$, where λ_3 is the maximum of the next hitting time and the next crossing time. For convenience, the lasso path algorithm is summarized below.

As we decrease λ from a knot λ_k, we can rewrite the lasso coefficient update in Step 1 as

$$\begin{aligned} \hat{\beta}_A(\lambda) &= \hat{\beta}_A(\lambda_k) + (\lambda_k - \lambda)(X_A^T X_A)^{-1}s_A, \\ \hat{\beta}_{-A}(\lambda) &= 0. \end{aligned} \tag{28}$$

We can see that we are moving the active coefficients in the direction $(\lambda_k - \lambda)(X_A^T X_A)^{-1}s_A$ for decreasing λ. In other words, the lasso fitted values proceed as

$$X\hat{\beta}(\lambda) = X\hat{\beta}(\lambda_k) + (\lambda_k - \lambda)X_A(X_A^T X_A)^{-1}s_A,$$

for decreasing λ. [9] call $X_A(X_A^T X_A)^{-1}s_A$ the *equiangular direction*, because this direction, in $\mathbb{R}^n$, takes an equal angle with all $X_j \in \mathbb{R}^n$, $j \in A$.

For this reason, the lasso path algorithm in Algorithm **??** is also often referred to as the *least angle regression* path algorithm in "lasso mode", though we have not mentioned this yet to avoid confusion. Least angle regression is considered as another algorithm by itself, where we skip Step 3 altogether. In words, Step 3 disallows any component path to cross through zero. The left side of the plot in Figure 4 visualizes the distinction between least angle regression and lasso estimates: the dotted black line displays the least angle regression component path, crossing through zero, while the lasso component path remains at zero.

Lastly, an alternative expression for the coefficient update in (28) (the update in Step 1) is

$$\begin{aligned} \hat{\beta}_A(\lambda) &= \hat{\beta}_A(\lambda_k) + \frac{\lambda_k - \lambda}{\lambda_k}(X_A^T X_A)^{-1}X_A^T r(\lambda_k), \\ \hat{\beta}_{-A}(\lambda) &= 0, \end{aligned} \tag{29}$$

where $r(\lambda_k) = y - X_A\hat{\beta}_A(\lambda_k)$ is the residual (from the fitted lasso model) at λ_k. This follows because, recall, $\lambda_k s_A$ are simply the inner products of the active variables with the residual at λ_k, i.e., $\lambda_k s_A = X_A^T(y - X_A\hat{\beta}_A(\lambda_k))$. In words, we can see that the update for the active lasso coefficients in (29) is in the direction of the least squares coefficients of the residual $r(\lambda_k)$ on the active variables X_A.

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] Sourav Chatterjee. Assumptionless consistency of the lasso. arXiv: 1303.5817, 2013.

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

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

[9] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

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

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

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

[13] 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.

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

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

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

[17] Michael Osborne, Brett Presnell, and Berwin Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–404, 2000.

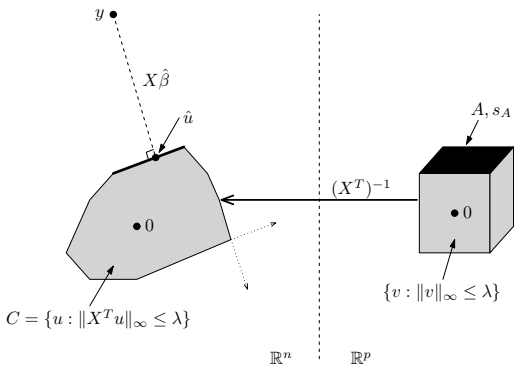
[18] Michael Osborne, Brett Presnell, and Berwin Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

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

[20] Sara van de Geer and Peter Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[21] 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.

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



1

Figure 3: *A geometric picture of the lasso solution. The left panel shows the polyhedron underlying all lasso fits, where each face corresponds to a particular combination of active set A and signs s; the right panel displays the "inverse" polyhedron, where the dual solutions live*

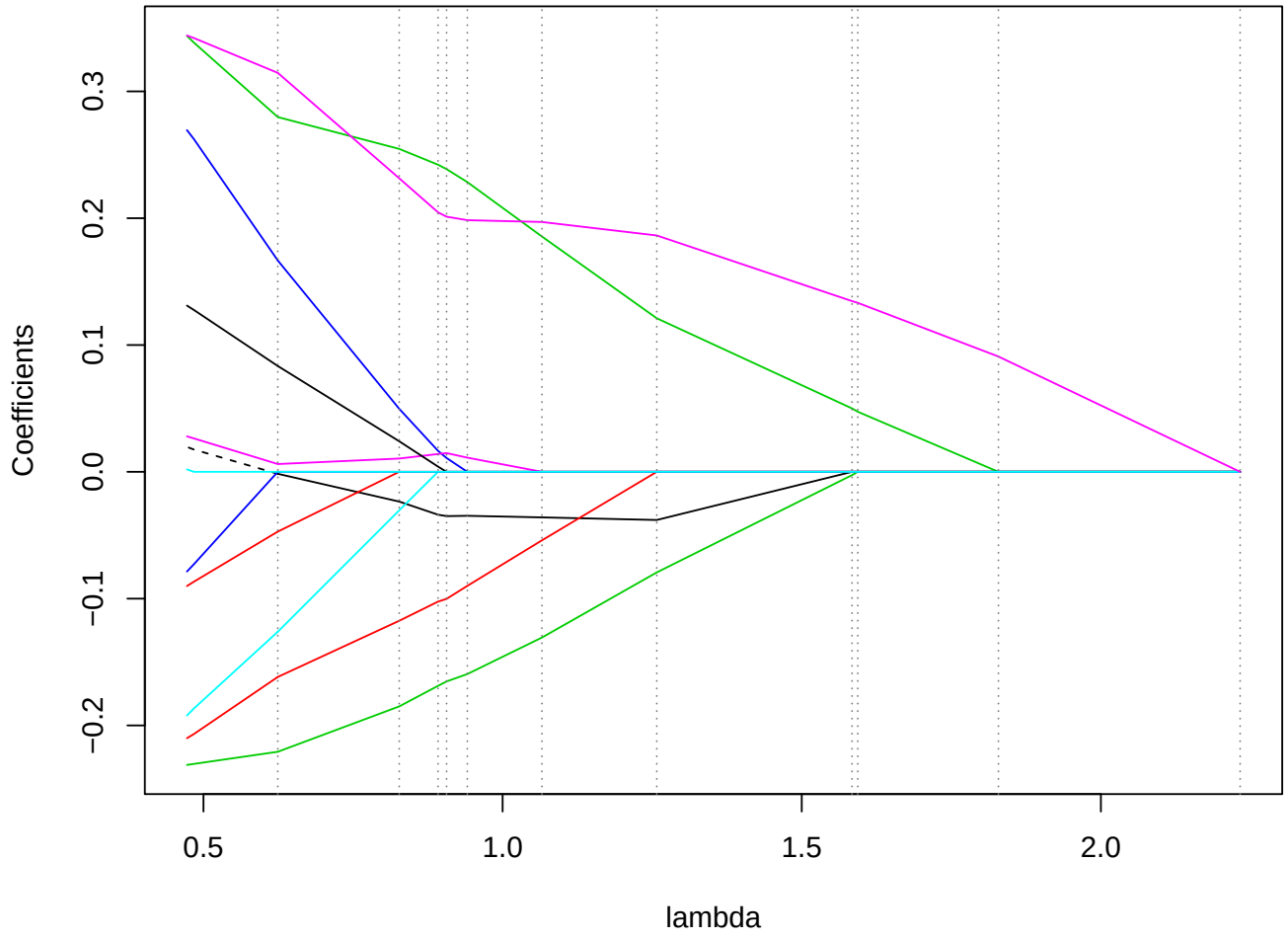


Figure 4: *An example of the lasso path. Each colored line denotes a component of the lasso solution* $\hat{\beta}_j(\lambda)$, $j = 1, \ldots, p$ *as a function of* λ. *The gray dotted vertical lines mark the knots* $\lambda_1 \geq \lambda_2 \geq \ldots$