Statistical Learning with Sparsity

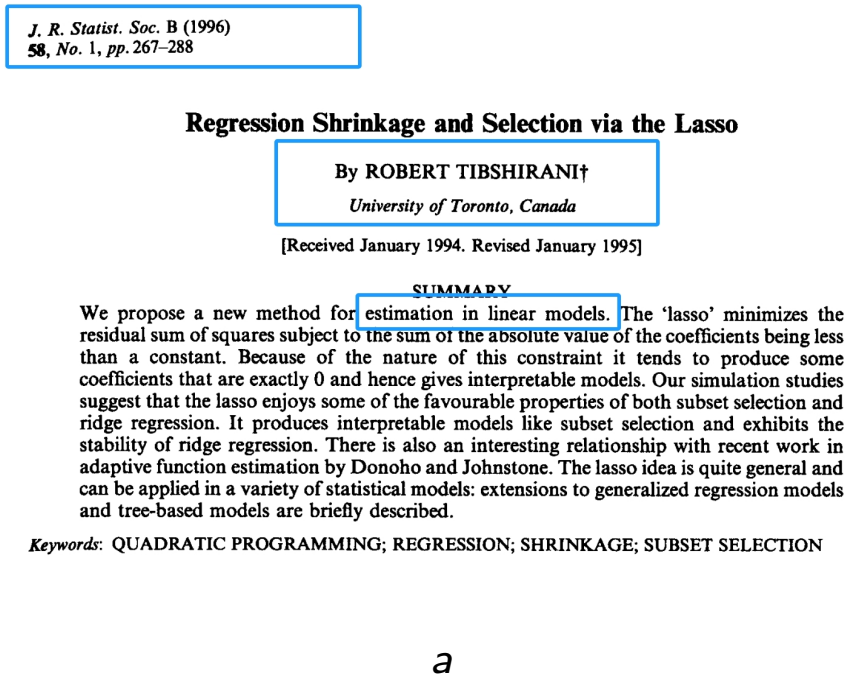
A Brief Introduction

Boen Jiang

Fudan University

November 11, 2024

Brief Histories: Robert Tibshirani



J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†
University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: QUADRATIC PROGRAMMING; REGRESSION; SHRINKAGE; SUBSET SELECTION

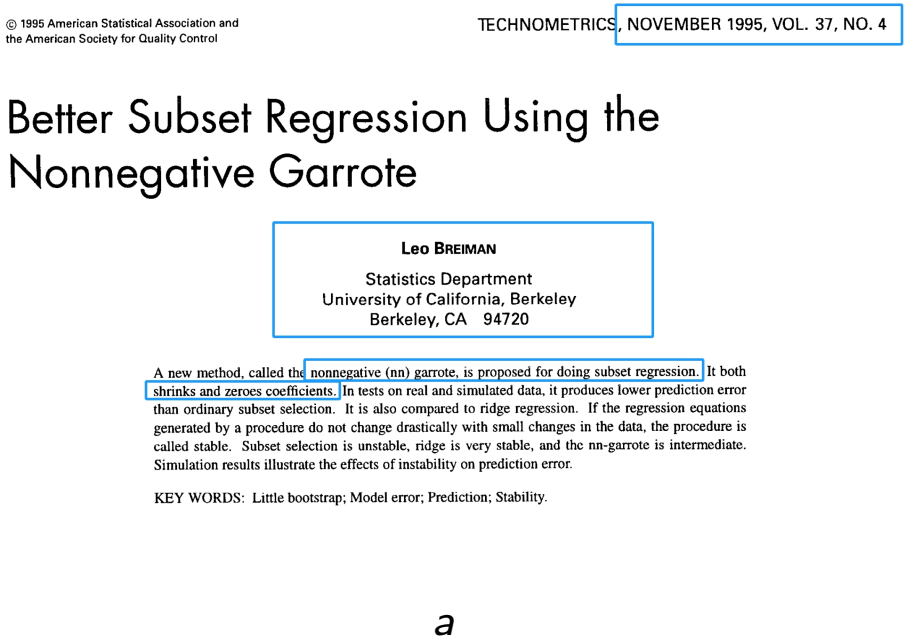
a

[a] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (Jan. 1996), pp. 267–288

Rob Tibshirani

- B. Math. in statistics and computer science from the University of Waterloo in 1979.
- Master's degree in Statistics from the University of Toronto in 1980.
- Received his Ph.D. in 1984 under the supervision of Bradley Efron.

Breif Histories: Leo Breiman



TECHNOMETRICS, NOVEMBER 1995, VOL. 37, NO. 4

Better Subset Regression Using the Nonnegative Garrote

Leo Breiman
Statistics Department
University of California, Berkeley
Berkeley, CA 94720

A new method, called the nonnegative (nn) garrote, is proposed for doing subset regression. It both shrinks and zeroes coefficients. In tests on real and simulated data, it produces lower prediction error than ordinary subset selection. It is also compared to ridge regression. If the regression equations generated by a procedure do not change drastically with small changes in the data, the procedure is called stable. Subset selection is unstable, ridge is very stable, and the nn-garrote is intermediate. Simulation results illustrate the effects of instability on prediction error.

KEY WORDS: Little bootstrap; Model error; Prediction; Stability.

a

[a] Leo Breiman. "Better Subset Regression Using the Nonnegative Garrote". en. In: 37.4 (1995)

- Let $\left\{\widehat{\beta}_k\right\}$ be the original OLS estimates. Take $\{c_k\}$ to minimize

$$\sum_k \left(y_n - \sum_k c_k \widehat{\beta}_k x_{kn} \right)^2$$

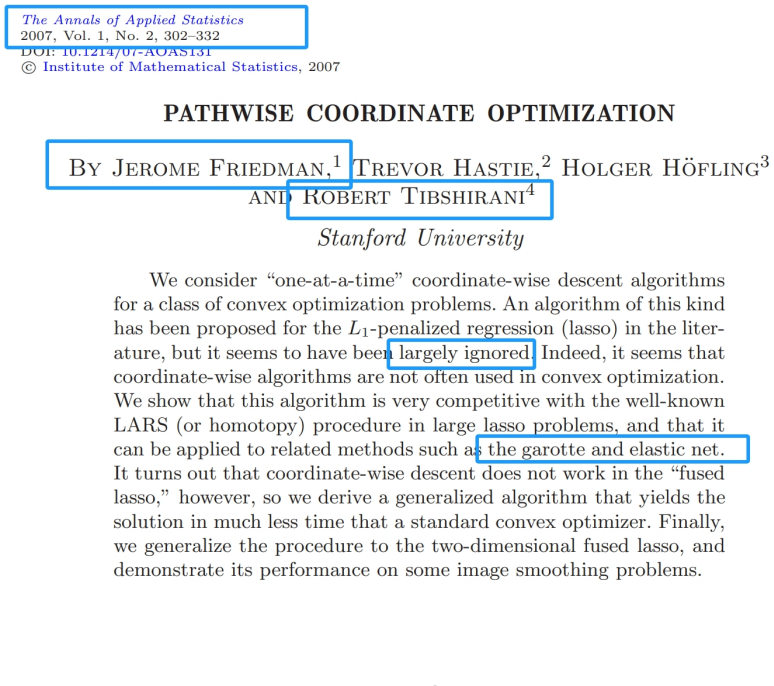
under the constraints

$$c_k \geq 0, \quad \sum_k c_k \leq s.$$

The $\widetilde{\beta}_k(s) = c_k \widehat{\beta}_k$ are the new predictor coefficients.

- "Lasso is a softer term for Canadian than Garrote".

Brief Histories: Jerome Friedman



The Annals of Applied Statistics
2007, Vol. 1, No. 2, 302–332
DOI: 10.1214/07-AOAS131
© Institute of Mathematical Statistics, 2007

PATHWISE COORDINATE OPTIMIZATION

BY JEROME FRIEDMAN,[1] TREVOR HASTIE,[2] HOLGER HÖFLING[3] AND ROBERT TIBSHIRANI[4]

Stanford University

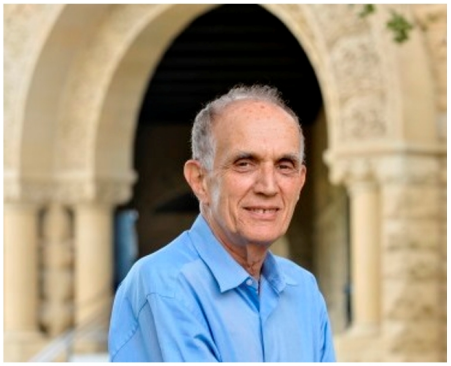
We consider "one-at-a-time" coordinate-wise descent algorithms for a class of convex optimization problems. An algorithm of this kind has been proposed for the L_1-penalized regression (lasso) in the literature, but it seems to have been largely ignored. Indeed, it seems that coordinate-wise algorithms are not often used in convex optimization. We show that this algorithm is very competitive with the well-known LARS (or homotopy) procedure in large lasso problems, and that it can be applied to related methods such as the garotte and elastic net. It turns out that coordinate-wise descent does not work in the "fused lasso," however, so we derive a generalized algorithm that yields the solution in much less time that a standard convex optimizer. Finally, we generalize the procedure to the two-dimensional fused lasso, and demonstrate its performance on some image smoothing problems.

a

[a] Jerome Friedman et al. "Pathwise coordinate optimization". en. In: *The Annals of Applied Statistics* 1.2 (Dec. 2007). arXiv:0708.1485 [stat]. ISSN: 1932-6157

- In history, the coordinate descent idea appeared in Fu (1998), then in Daubechies et al. (2004), but it was ignored then. Later in 2007 it gained popularity.
- This work focused on linear regression. Friedman and Hastie focused the implementation of R package `glmnet`.

Interesting Connections



Bradley Efron

Rob Tibshirani

Jerome Friedman

Leo Breiman

Ryan Tibshirani

Trevor Hastie

Geof Hinton

Linear Regression

Given:

- N samples $\{(x_i, y_i)\}_{i=1}^{N}$,
- $x_i = (x_{i1}, \ldots, x_{ip})$ is a p-dimensional vector of predictors and each $y_i \in \mathbb{R}$ is the associated response variable.

Linear Regression Model:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$$

where

- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are unknown parameters and e_i are error terms.

Ordinary Least Squares for parameter β

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

Trouble in High Dimensions

- Least squares risk degrades as $p \to n$:

$$\frac{1}{n}\mathbb{E}\left[\left\|X\hat{\beta} - X\beta_0\right\|_2^2 \middle| X\right] = \sigma^2 \frac{p}{n}$$

- When $p > n$, least squares has no unique solution:

$$\hat{\beta} = \left(X^\top X\right)^+ X^\top Y + \eta, \quad \eta \in \ker(X)$$

where A^+ is pseudo-inverse, $\ker(A)$ is the null space

  - Coefficient interpretation impossible $\rightsquigarrow$ for any $\hat{\beta}_j > 0$, exists $\tilde{\beta}_j < 0$ solution
  - Sign consistency unattainable

Remarks on Sparsity

"We are drowning in information and starving for knowledge."

- sparse statistical model is one in which only a relatively small number of parameters (or predictors) play an important role
- reasonable interpretation of the fitted model and computational convenience
- Use a procedure that does well in sparse problems, since no procedure does well in dense problems.

The "Bet on Sparsity" Principle

Use methods that perform well under sparsity, as no method performs well in dense problems.

An Example for lasso-regularized multinomial classifier

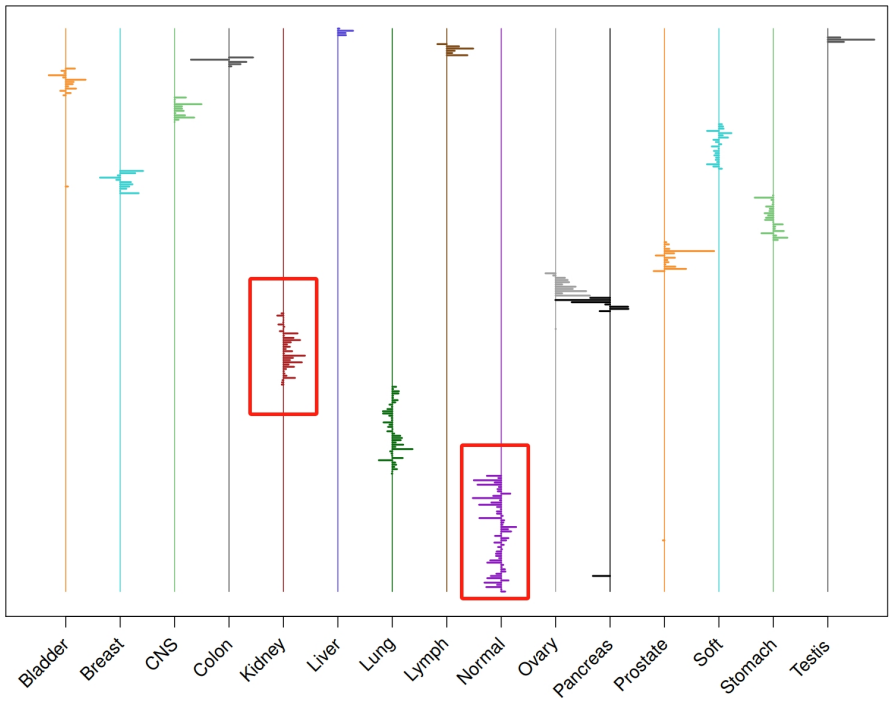
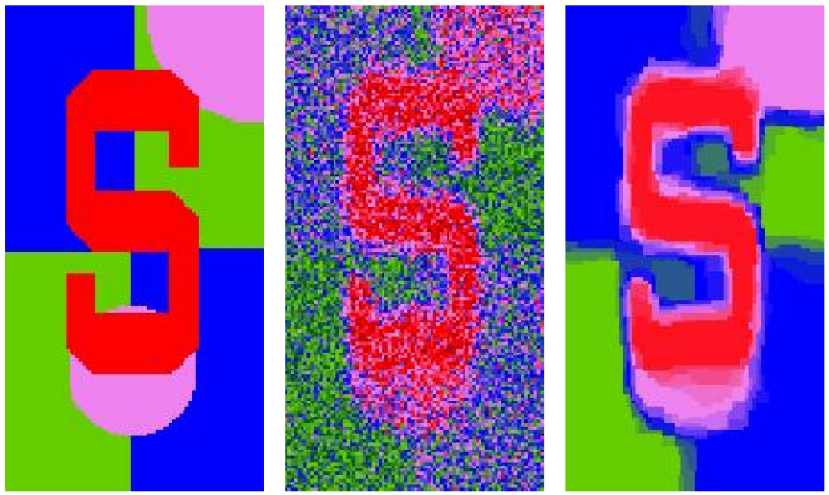


图: 15-class gene expression cancer data: estimated nonzero feature weights from a lasso-regularized multinomial classifier. Shown are the 254 genes (out of 4718) with at least one nonzero weight among the 15 classes. The genes (unlabelled) run from top to bottom. Line segments pointing to the right indicate positive weights, and to the left, negative weights. We see that only a handful of genes are needed to characterize each class.

An example of the 2d fused lasso for image denoising



(a) Original (b) Noisy (c) Denoised

图: We started with a toy signal, shown in (a). The colors green, blue, purple, red in the image correspond to the numeric levels 1, 2, 3, 4, respectively. We then added noise, shown in (b), interpolating between colors to display the intermediate values. This is used as the data y in the $2d$ fused lasso problem. The solution (for $\lambda = 1$) is shown in (c), and it is a fairly accurate reconstruction. The fused lasso is effective here because the original image is piecewise constant. [1]

[1] Ryan J. Tibshirani and Jonathan Taylor. "The solution path of the generalized lasso". en. In: *The Annals of Statistics* 39.3 (June 2011)

Lasso Estimator

Lasso for Linear Regression (optimization with ℓ_1-norm constraint)

Given a collection of N predictor-response pairs $\{(x_i, y_i)\}_{i=1}^N$, the lasso finds the solution $\hat{\beta}$ to the optimization problem

$$\underset{\beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

The constraint $\sum_{j=1}^{p} |\beta_j| \leq t$ can be written more compactly as the ℓ_1-norm constraint $\|\beta\|_1 \leq t$.

Remarks on Lasso Problem

- Typically, we first standardize the predictors $\mathbf{X}$ so that each column is centered $\left(\frac{1}{N}\sum_{i=1}^{N} x_{ij} = 0\right)$ and has unit variance $\left(\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1\right)$.
- Without standardization, the lasso solutions would depend on the units (e.g., feet versus meters) used to measure the predictors. On the other hand, we typically would not standardize if the features were measured in the same units.
- For convenience, we also assume that the outcome values y_i have been centered, meaning that $\frac{1}{N}\sum_{i=1}^{N} y_i = 0$. These centering conditions are convenient, since they mean that we can omit the intercept term β_0 in the lasso optimization.

Lagrangian Formulation (Not Dual Form)

$$\underset{\beta}{\text{minimize}} \left\{ \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t,$$

is equivalent to the Lagrangian form:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

for some $\lambda \geq 0$. The two forms are equivalent in the sense that for a given λ, there exists a t with the same solution.

- For each value of t in the range where the constraint $\|\beta\|_1 \leq t$ is active, there is a corresponding value of λ that yields the same solution from the Lagrangian form.
- Conversely, the solution $\widehat{\beta}_\lambda$ to Lagrangian problem solves the bound problem with $t = \left\| \widehat{\beta}_\lambda \right\|_1$.

Canonical Regularizers in Regression

- ℓ_0 pseudo-norm: subset selection

$$\|\beta\|_0 = \sum_{j=1}^{p} 1\{\beta_j \neq 0\}$$

  - fails positive homogeneity
  - not convex $\rightsquigarrow$ NP-hard

- ℓ_1 norm: lasso

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

- ℓ_2 norm: ridge

$$\|\beta\|_2 = \left(\sum_{j=1}^{p} \beta_j^2 \right)^{1/2}$$

Convex Optimization Conditions

- Since lasso problem has a convex objective function, and a convex constraint, the solution to the lasso problem is guaranteed to be a global minimum. The lasso problem is a convex optimization problem.
- the ℓ_1-norm $g(\beta) = \sum_{j=1}^{p} |\beta_j|$ is a convex function, but it fails to be differentiable at any point where at least one coordinate β_j is equal to zero $\rightsquigarrow$ subgradient.
- a vector $z \in \mathbb{R}^p$ is said to be a subgradient of f at β if $f(\beta') \geq f(\beta) + \langle z, \beta' - \beta \rangle \quad$ for all $\beta' \in \mathbb{R}^p$.
- At points of nondifferentiability, the subdifferential is a convex set containing all possible subgradients.
- For absolute value function $f(\beta) = |\beta|$, we have

$$\partial f(\beta) = \begin{cases} \{+1\} & \text{if } \beta > 0 \\ \{-1\} & \text{if } \beta < 0 \\ [-1, +1] & \text{if } \beta = 0 \end{cases}$$

Subgradient

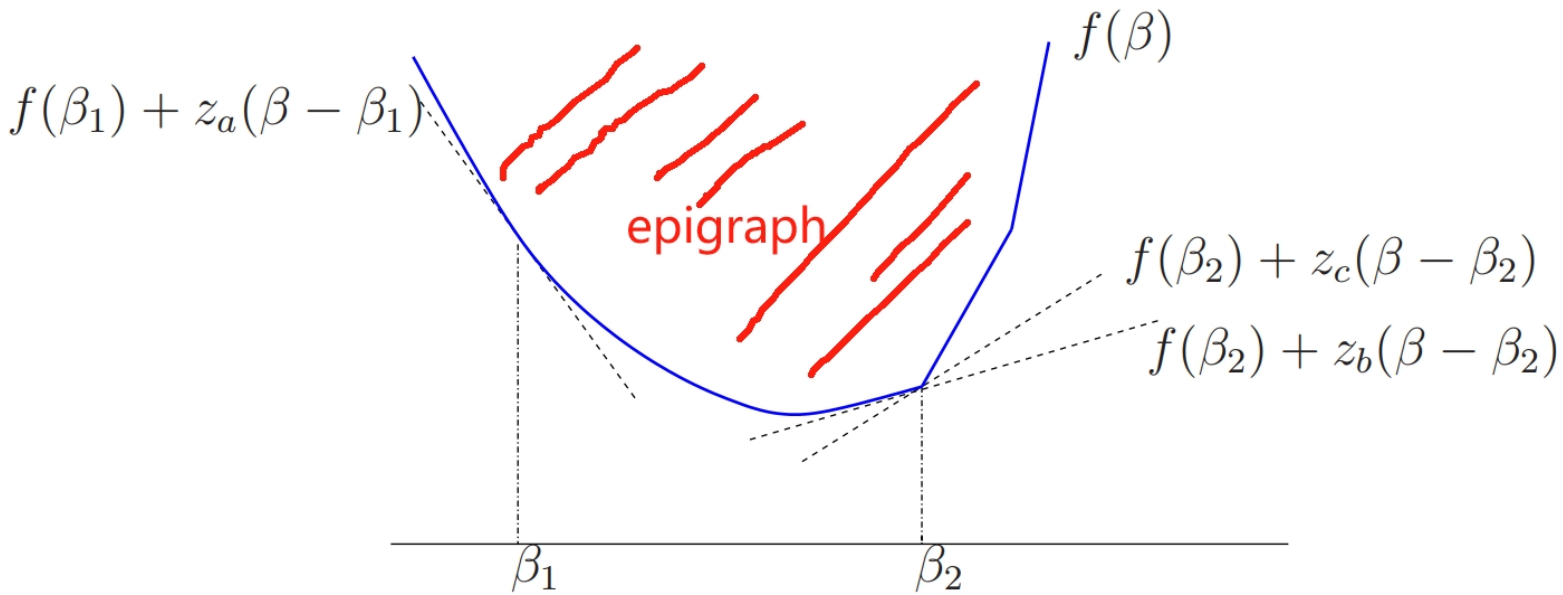


图: A convex function $f : \mathbb{R} \to \mathbb{R}$, along with some examples of subgradients at β_1 and β_2. In geometric terms, the subgradient vector z is the normal to a (nonvertical) hyperplane that supports the epigraph of f.

Generalized KKT's Stationary Condition

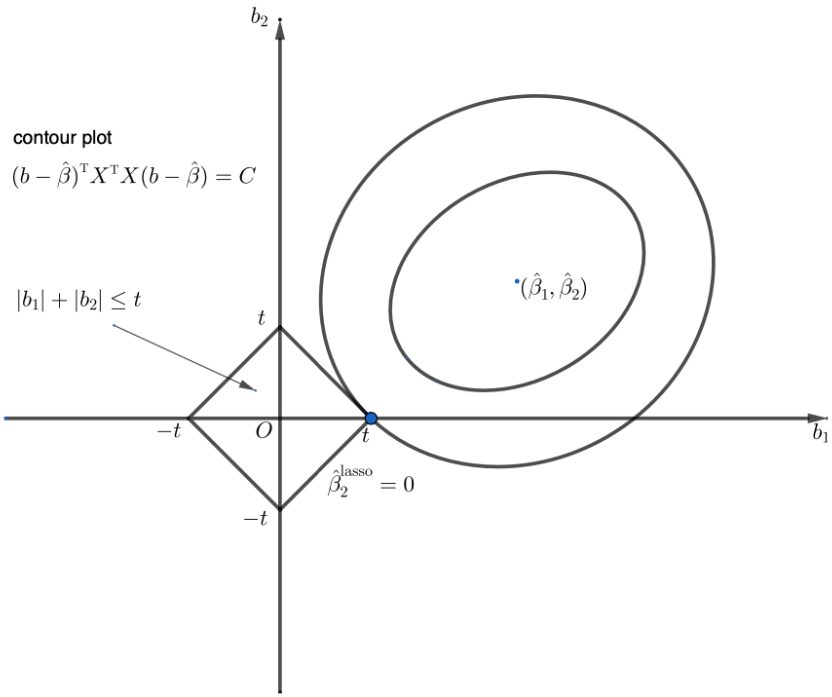
$$0 \in \partial f(\beta^*) + \sum_{j=1}^{m} \lambda_j^* \partial g_j(\beta^*)$$

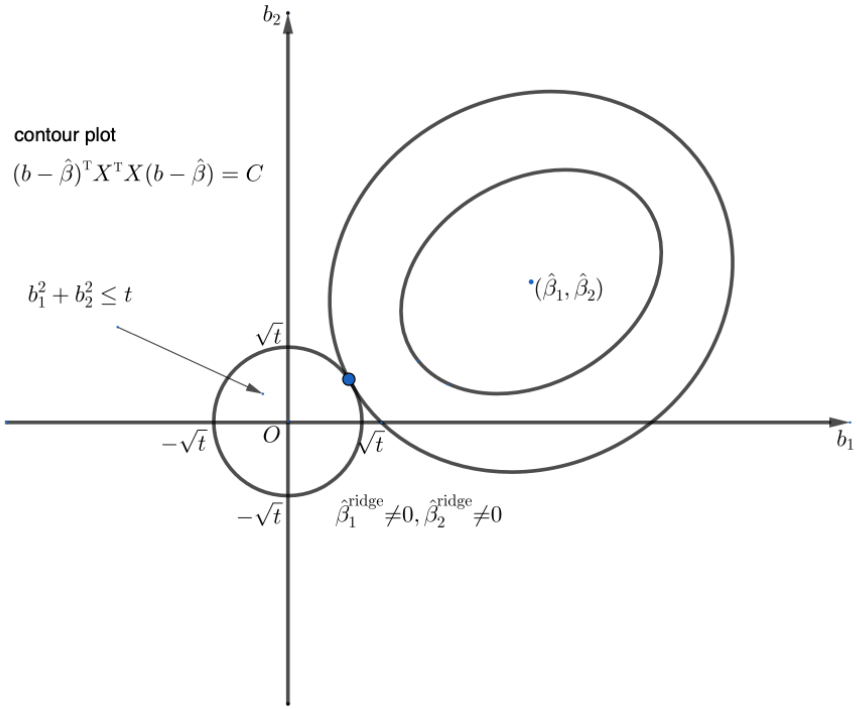
Hence, a necessary condition for β^* to be a solution to the lasso problem is that

$$-\frac{1}{N} \langle \boldsymbol{x}_j, \boldsymbol{y} - \boldsymbol{X}\beta \rangle + \lambda s_j = 0, j = 1, \ldots, p$$

Here each s_j is an unknown quantity equal to $\operatorname{sign}(\beta_j)$ if $\beta_j \neq 0$ and some value lying in $[-1, 1]$ otherwise – that is, it is a subgradient for the absolute value function.

Contour Plots: Lasso vs Ridge





Why are the plot like these?

$$(Y-Xb)^{\top}(Y-Xb)=(Y-X\hat{\beta})^{\mathrm{T}}(Y-X\hat{\beta})+(b-\hat{\beta})^{\mathrm{T}}X^{\mathrm{T}}X(b-\hat{\beta})$$

Generalization – Overfitting Tradeoff

- larger values of t free up more parameters and allow the model to adapt more closely to the training data,
- smaller values of t restrict the parameters more, leading to sparser, more interpretable models that fit the data less closely.
- A value of t that is too small can prevent the lasso from capturing the main signal in the data, while too large a value can lead to overfitting.
- $\rightsquigarrow$ K-fold cross-validation to select the best value of t.
- $K = 5, 10$ in general. K can be N for leave-one-out cross-validation(LOOCV).

K-fold Cross-validation

1. Randomly divide the set of observations into $K > 1$ groups, or folds, of approximately equal size.
2. Fix one fold as test set and remaining $k - 1$ folds as training sets.
3. Apply the lasso to the training data for a range of different t values.
4. Use each fitted model to predict the response in the test set.
5. Determine the mean-squared prediction errors for each value of t
   Mean-squared error for test fold j :

$$MSE_j = \frac{1}{|F_j|} \sum_{i \in F_j} \left(y_i - \hat{f}(x_i) \right)^2 .$$

6. Repeat process k times such that each fold is once the test set.
7. Average k mean-squared errors for each value of t

$$CV_{(k)}(t) = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

"cross-validation error curve".

Baseball Data Analysis

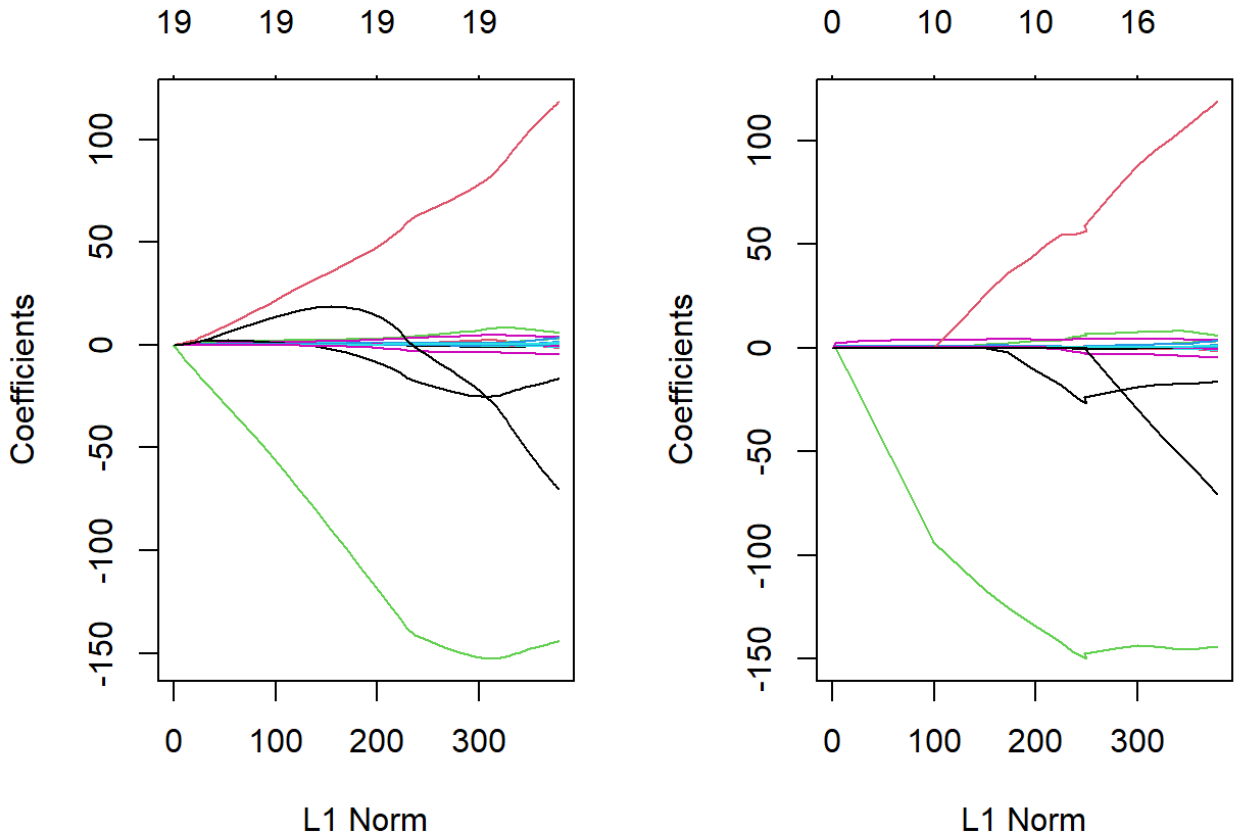
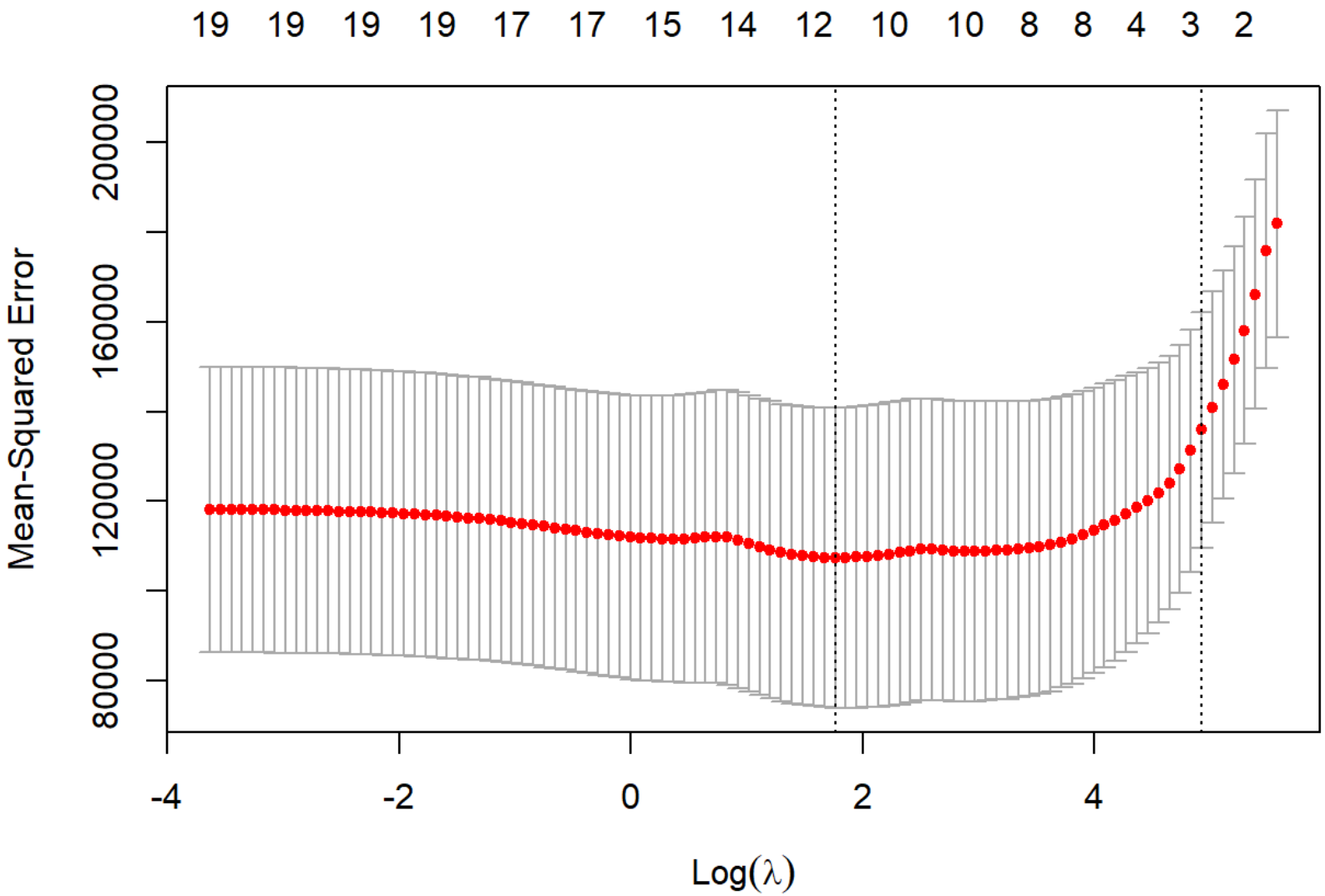


图: we can see from the coefficient plot that depending on lambda, some of the coefficients will be exactly equal to zero

Baseball Data Analysis: Cross-validation



Computing the lasso using commercial solvers

The lasso problem is a convex program, specifically a quadratic program (QP) with a convex constraint. As such, there are many commercial solvers for solving the lasso. For example, `Gurobi`, `Mosek` .

```python
# 创建一个Gurobi模型
model = gp.Model("l1_regularization")
# 创建变量
x = model.addVars(n, lb=-GRB.INFINITY, name="x")  # x的下界设为负无穷
# 设置目标函数
# 目标是最小化 (1/2) * ||Ax - b||_2^2 + mu * ||x||_1
# 我们需要引入一个额外的变量来处理||x||_1
z = model.addVars(n, name="z")  # z用于表示|x|
# 添加目标函数
model.setObjective(
    (1/2) * gp.quicksum((gp.quicksum(A[i, j] * x[j] for j in range(n)) - b[i]) ** 2 for i in range(m)) +
    mu * gp.quicksum(z[j] for j in range(n)),
    GRB.MINIMIZE
)
# 添加约束: z[j] >= x[j] 和 z[j] >= -x[j]
for j in range(n):
    model.addConstr(z[j] >= x[j])
    model.addConstr(z[j] >= -x[j])

# 求解模型
model.optimize()
```

图: Gurobi is a commercial optimization solver that can be used to solve the lasso problem. The above code snippet shows how to use Gurobi to solve the lasso problem in Python.

Coordinate descent lemma

Let's begin with toy case. Let $\operatorname{sign}(x)$ denote the sign of a real number x, which equals $1, 0, -1$ if $x > 0, x = 0, x < 0$, respectively. Let $(x)_+ = \max(x, 0)$ denote the positive part of a real number x.
By classification discussion, given b_0 and $\lambda \geq 0$, we have

$$\begin{aligned}\arg\min_{b\in\mathbb{R}} \frac{1}{2}(b - b_0)^2 + \lambda|b| &= \operatorname{sign}(b_0)(|b_0| - \lambda)_+ \\ &= \begin{cases} b_0 - \lambda, & \text{if } b_0 \geq \lambda, \\ 0 & \text{if } -\lambda \leq b_0 \leq \lambda, \\ b_0 + \lambda & \text{if } b_0 \leq -\lambda. \end{cases} \\ &\triangleq \mathcal{S}_\lambda(b_0).\end{aligned}$$

Here the soft-thresholding operator

$$\mathcal{S}_\lambda(x) = \operatorname{sign}(x)(|x| - \lambda)_+$$

translates its argument x toward zero by the amount λ, and sets it to zero if $|x| \leq \lambda$.

Soft Thresholding

证明.

$$\frac{1}{2}(b-b_0)^2+\lambda|b| = \frac{1}{2}b^2+(\lambda|b|-b_0b)+\frac{1}{2}b_0^2$$
$$= \frac{1}{2}b^2+(\lambda\text{sign}(b)-b_0)b+\frac{1}{2}b_0^2$$

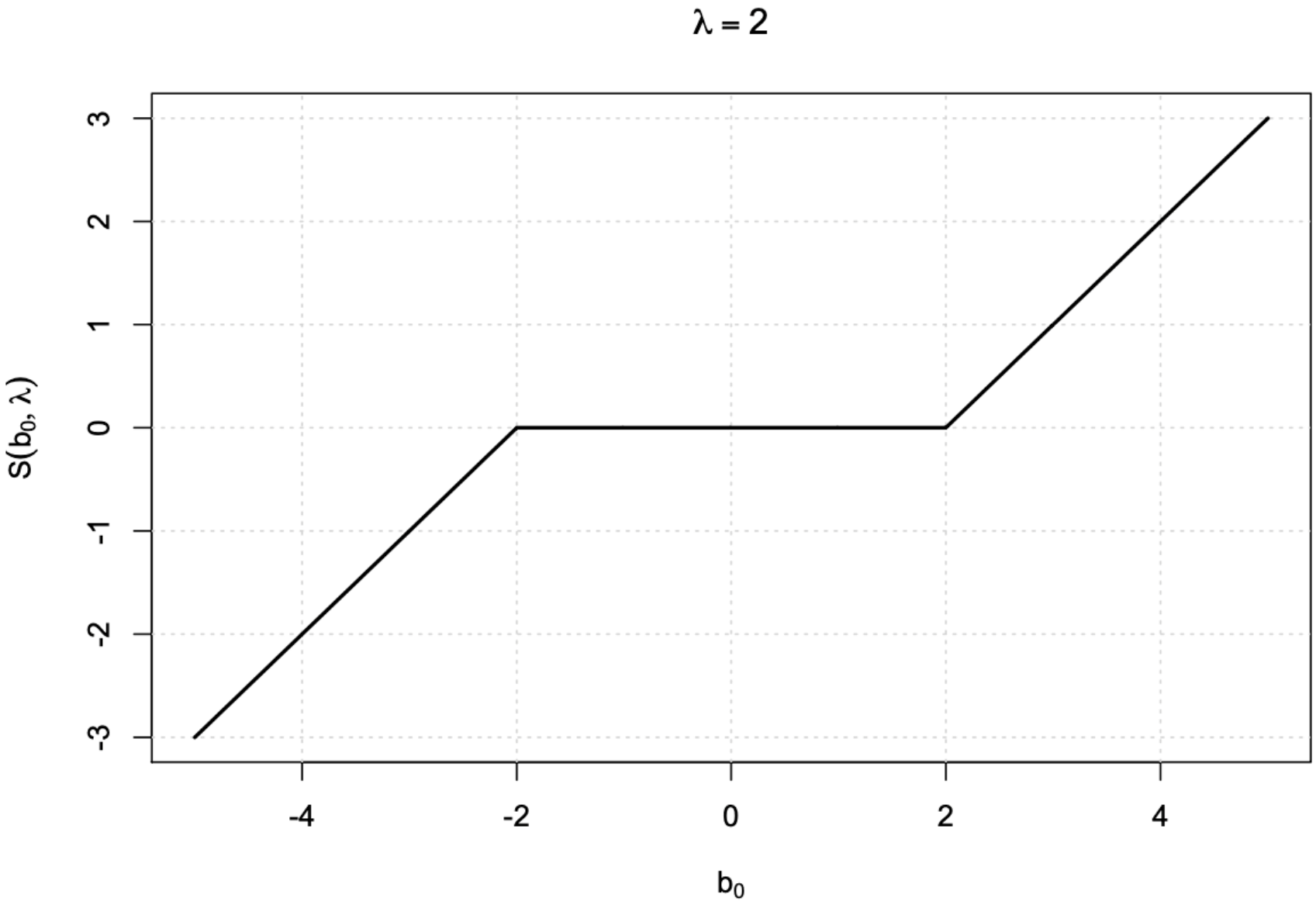
在$b>0$部分, 上式$=\frac{1}{2}b^2+(\lambda-b_0)b+\frac{1}{2}b_0^2$,

1. $b_0 \geq \lambda \Rightarrow \arg\min = b_0-\lambda$;
2. $b_0 \leq \lambda \Rightarrow \arg\min = 0$

在$b<0$部分, 上式$=\frac{1}{2}b^2+(-\lambda-b_0)b+\frac{1}{2}b_0^2$,

1. $b_0 \geq -\lambda \Rightarrow \arg\min = 0$;
2. $b_0 \leq -\lambda \Rightarrow \arg\min = b_0+\lambda$

Soft Thresholding



Computing the lasso via coordinate descent

- The `glmnet` package in R uses the coordinate descent algorithm.

Coordinate descent vs proximal gradient for lasso regression: 100 random instances with $n = 200$, $p = 50$ (all methods cost $O(np)$ per iter)

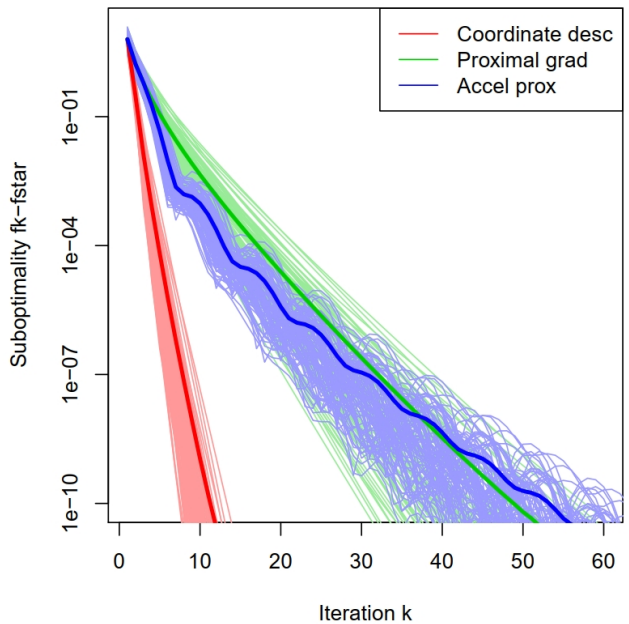


图: In Ryan Tibshirani's slides, the coordinate descent algorithm is outperforming. Trever also compares (among others) his two R packages: `glmnet` (using coordinate descent) and lars (using `LARS`). It's shown that coordinate descent is faster in each setting: $p > N$,$p < N$ and sparse or dense data.

Coordinate Descent Algorithm

For a general case, write the Lagrangian form of the lasso problem as

$$\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\sum_{k\neq j}x_{ik}\beta_k-x_{ij}\beta_j\right)^2+\lambda\sum_{k\neq j}|\beta_k|+\lambda|\beta_j|$$

Denote partial residual as

$$r_{ij}=y_i-\sum_{k\neq j}\hat{\beta}_k x_{ik}$$

Then we need to optimize

$$\frac{1}{2N}\sum_{i=1}^{N}\left(r_{ij}-\beta_j x_{ij}\right)^2+\lambda|\beta_j|\,.$$

Coordinate Descent Algorithm

Run OLS of r_{ij} on x_{ij} to get the estimate $\hat{\beta}_{j,0}$. Here,

$$\hat{\beta}_{j,0} = \frac{\sum_{i=1}^{n} x_{ij} r_{ij}}{\sum_{i=1}^{n} x_{ij}^2 (= n)} = n^{-1} \sum_{i=1}^{n} x_{ij} r_{ij}$$

By adding and substracting,

$$\begin{aligned} \frac{1}{2N} \sum_{i=1}^{N} (r_{ij} - \beta_j x_{ij})^2 &= \frac{1}{2N} \sum_{i=1}^{N} \left(r_{ij} - \hat{\beta}_{j,0} x_{ij} \right)^2 + \frac{1}{2N} \sum_{i=1}^{N} x_{ij}^2 \left(\beta_j - \hat{\beta}_{j,0} \right)^2 \\ &= \text{ constant } + \frac{1}{2} \left(\beta_j - \hat{\beta}_{j,0} \right)^2 . \end{aligned}$$

Using the soft-thresholding operator, we have

$$\hat{\beta}_j = \mathcal{S}_\lambda \left(\hat{\beta}_{j,0} \right) .$$

Comments on Coordinate Descent Algorithm

- The theory of Tseng (2001) ensures the convergence of the algorithm.[2]
- We can start with a large λ and all zero coefficients.
- We then gradually decrease λ, for each λ, we apply the algorithm.
- We finally select λ via K-fold CV.
- Since we gradually decrease λ, the initial values of from the last step are very close to the minimizer and the algorithm converges fairly fast.

[2] P. Tseng. "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization". en. In: *Journal of Optimization Theory and Applications* 109.3 (June 2001), pp. 475–494. ISSN: 1573-2878

Degree of Freedom

Suppose we have an additive-error model, with

$$y_i = f(x_i) + \epsilon_i, i = 1, \ldots, N,$$

for some unknown f and with the errors ϵ_i id $\left(0, \sigma^2\right)$. If the N sample predictions are denoted by $\widehat{\mathbf{y}}$, then we define

$$\mathrm{df}(\widehat{\mathbf{y}}) := \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathrm{Cov}\left(\widehat{y}_i, y_i\right).$$

This is consistent with the degrees of freedom in the usual definition of the t-statistic for a linear model.

The degrees of freedom corresponds to the total amount of self-influence that each response measurement has on its prediction.

LARS Algorithm

Algorithm 5.1 LEAST ANGLE REGRESSION.

1. Standardize the predictors to have mean zero and unit ℓ_2 norm. Start with the residual $\boldsymbol{r}_0 = \mathbf{y} - \bar{\mathbf{y}}$, $\beta^0 = (\beta_1, \beta_2, \ldots, \beta_p) = \mathbf{0}$. set empty
2. Find the predictor $\mathbf{x}_j$ most correlated with $\boldsymbol{r}_0$; i.e., with largest value for $|\langle \mathbf{x}_j, \boldsymbol{r}_0 \rangle|$. Call this value λ_0, define the active set $\mathcal{A} = \{j\}$, and $\mathbf{X}_\mathcal{A}$, the matrix consisting of this single variable. variable best correlated with the residual
3. For $k = 1, 2, \ldots, K = \min(N-1, p)$ do: when active vars is not everything
   (a) Define the least-squares direction $\delta = \frac{1}{\lambda_{k-1}}(\mathbf{X}_\mathcal{A}^T \mathbf{X}_\mathcal{A})^{-1} \mathbf{X}_\mathcal{A}^T \boldsymbol{r}_{k-1}$, and define the p-vector Δ such that $\Delta_\mathcal{A} = \delta$, and the remaining elements are zero.
   lambda here represents kind of turning point
   (b) Move the coefficients β from β^{k-1} in the direction Δ toward their least-squares solution on $\mathbf{X}_\mathcal{A}$: $\beta(\lambda) = \beta^{k-1} + (\lambda_{k-1} - \lambda)\Delta$ for $0 < \lambda \leq \lambda_{k-1}$, keeping track of the evolving residuals $\boldsymbol{r}(\lambda) = \mathbf{y} - \mathbf{X}\beta(\lambda) = \boldsymbol{r}_{k-1} - (\lambda_{k-1} - \lambda)\mathbf{X}\Delta$.
   (c) Keeping track of $|\langle \mathbf{x}_\ell, \boldsymbol{r}(\lambda) \rangle|$ for $\ell \notin \mathcal{A}$, identify the largest value of λ at which a variable "catches up" with the active set; if the variable has index j, that means $|\langle \mathbf{x}_j, \boldsymbol{r}(\lambda) \rangle| = \lambda$. This defines the next "knot" λ_k.
   (d) Set $\mathcal{A} = \mathcal{A} \cup \{j\}$, $\beta^k = \beta(\lambda_k) = \beta^{k-1} + (\lambda_{k-1} - \lambda_k)\Delta$, and $\boldsymbol{r}_k = \mathbf{y} - \mathbf{X}\beta^k$.
4. Return the sequence $\{\lambda_k, \beta^k\}_0^K$. update parameters

The Annals of Statistics
2004, Vol. 32, No. 2, 407–499

LEAST ANGLE REGRESSION

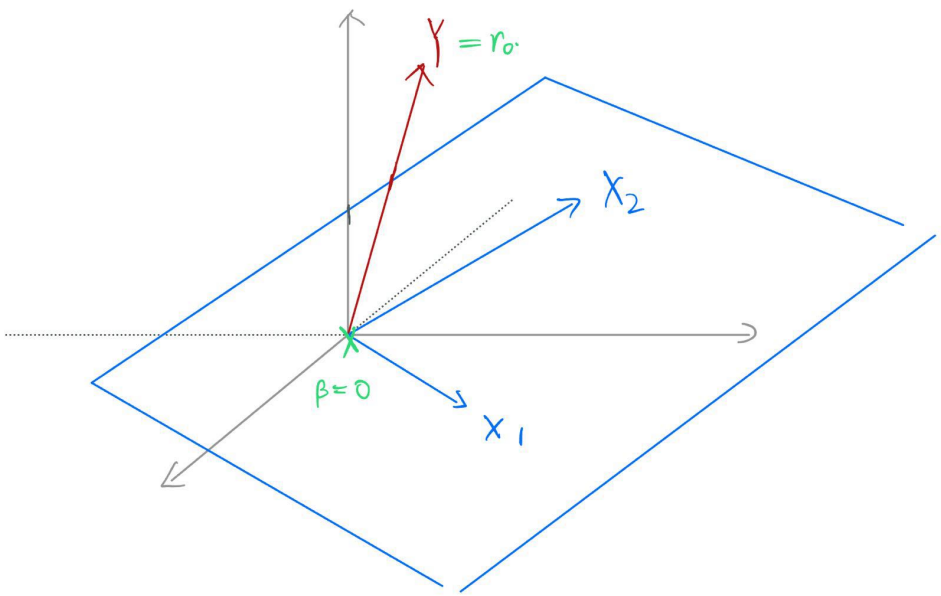
BY BRADLEY EFRON,[1] TREVOR HASTIE,[2] IAIN JOHNSTONE[3] AND ROBERT TIBSHIRANI[4]

Stanford University

The purpose of model selection algorithms such as *All Subsets*, *Forward Selection* and *Backward Elimination* is to choose a linear model on the basis of the same set of data to which the model will be applied. Typically we have available a large collection of possible covariates from which we hope to select a parsimonious set for the efficient prediction of a response variable. *Least Angle Regression* (LARS), a new model selection algorithm, is a useful and less greedy version of traditional forward selection methods. Three main properties are derived: (1) A simple modification of the LARS algorithm implements the Lasso, an attractive version of ordinary least squares that constrains the sum of the absolute regression coefficients; the LARS modification calculates all possible Lasso estimates for a given problem, using an order of magnitude less computer time than previous methods. (2) A different LARS modification efficiently implements Forward Stagewise linear regression, another promising new model selection method; this connection explains the similar numerical results previously observed for the Lasso and Stagewise, and helps us understand the properties of both methods, which are seen as constrained versions of the simpler LARS algorithm. (3) A simple approximation for the degrees of freedom of a LARS estimate is available, from which we derive a C_p estimate of prediction error; this allows a principled choice among the range of possible LARS estimates. LARS and its variants are computationally efficient: the paper describes a publicly available algorithm that requires only the same order of magnitude

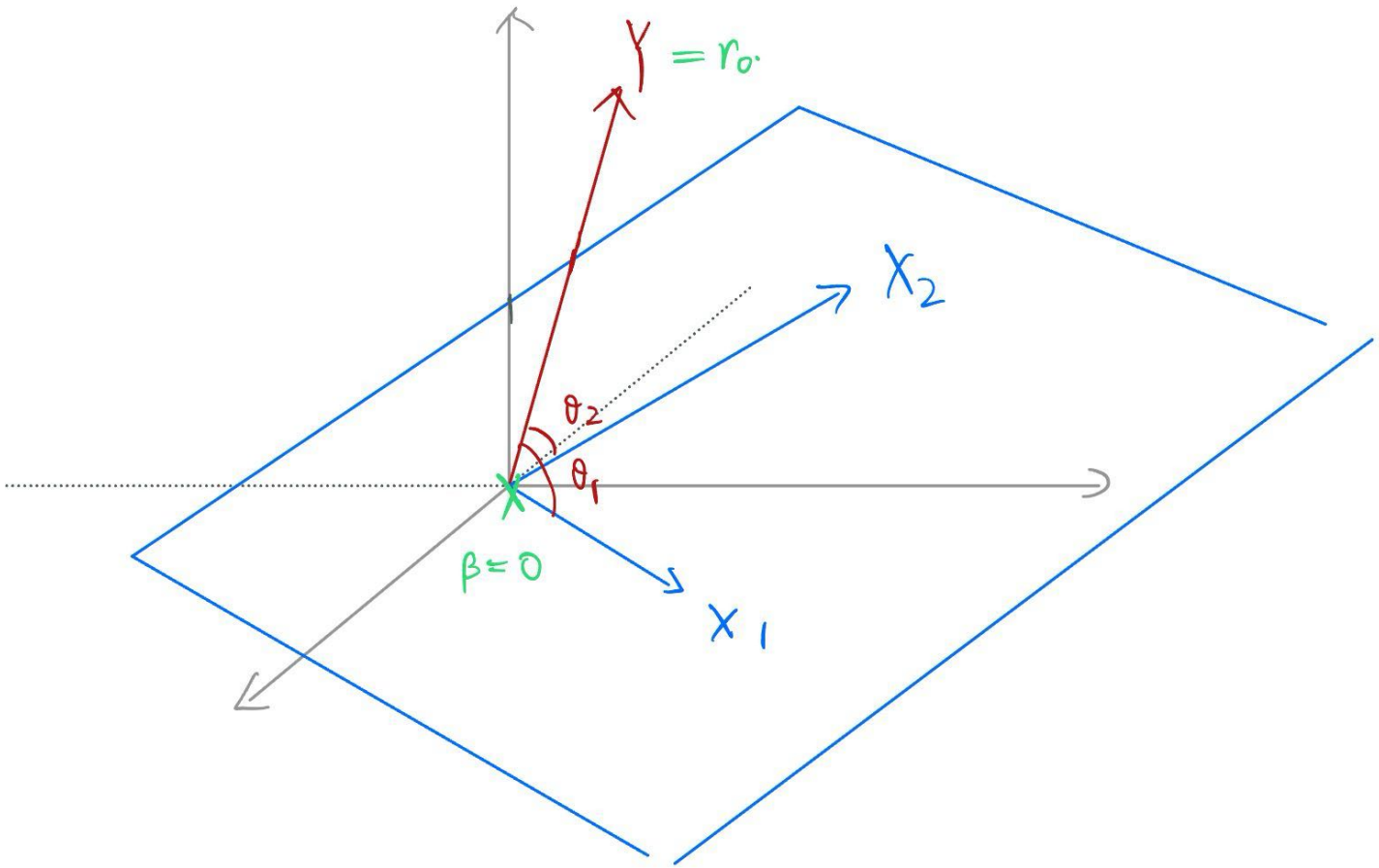
Step 1

- Standardize the predictors to have mean zero and unit ℓ^2 norm. Start with the residual $r_0 = \mathbf{y} - \bar{\mathbf{y}}, \beta^0 = (\beta_1, \beta_2, ..., \beta_p) = \mathbf{0}$.
- $\sum_{i=1}^n y_i = 0, \quad \sum_{i=1}^n x_{ij} = 0, \quad \sum_{i=1}^n x_{ij}^2 = 1 \quad$ for $j = 1, 2, \ldots, m$
- Angle:
  $\langle Y, X \rangle = \mathbb{E}\{YX\} = \mathrm{Cov}(X, Y) \rightsquigarrow \cos(\theta) = \mathrm{corr}(X, Y) = \langle X, Y \rangle$.



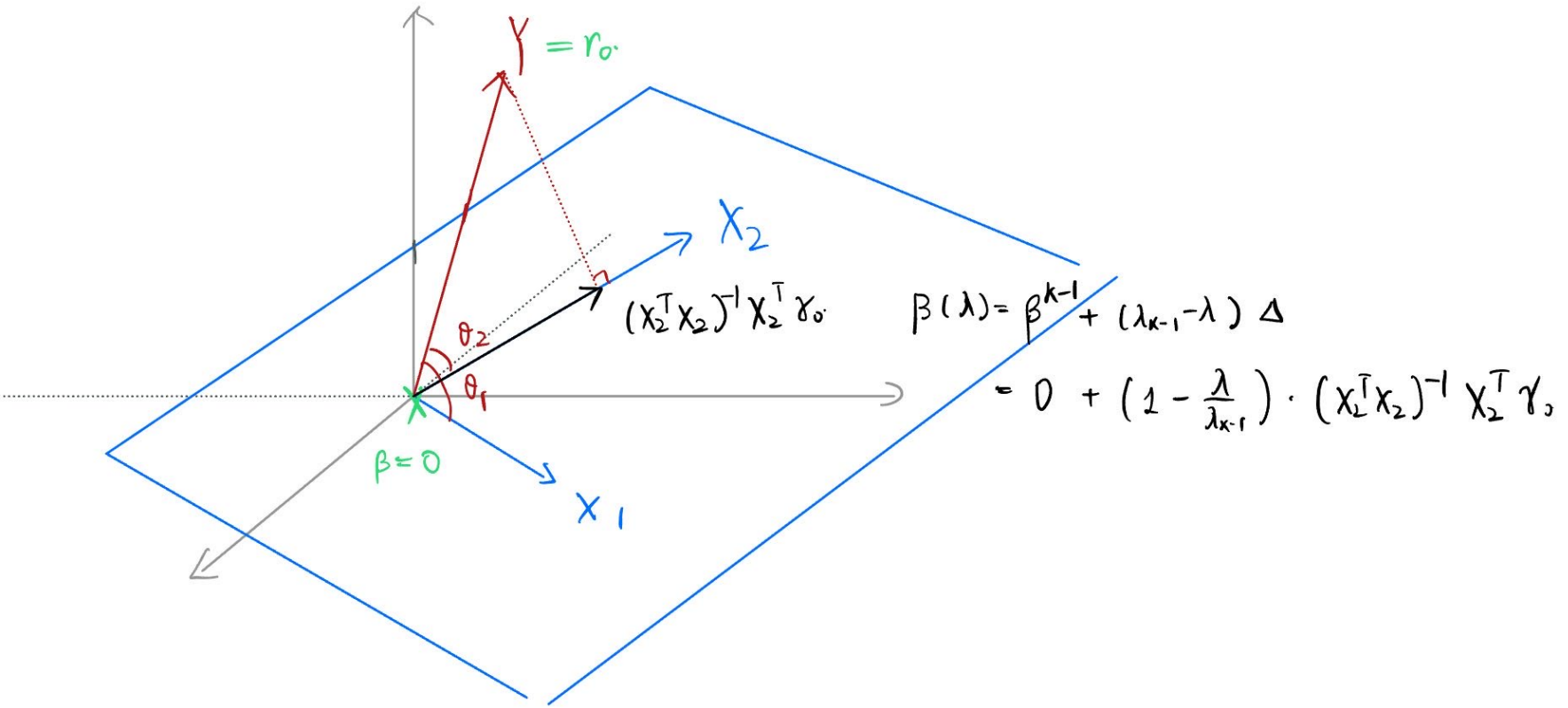
Step 2

- $\mathcal{A} = \{2\}$



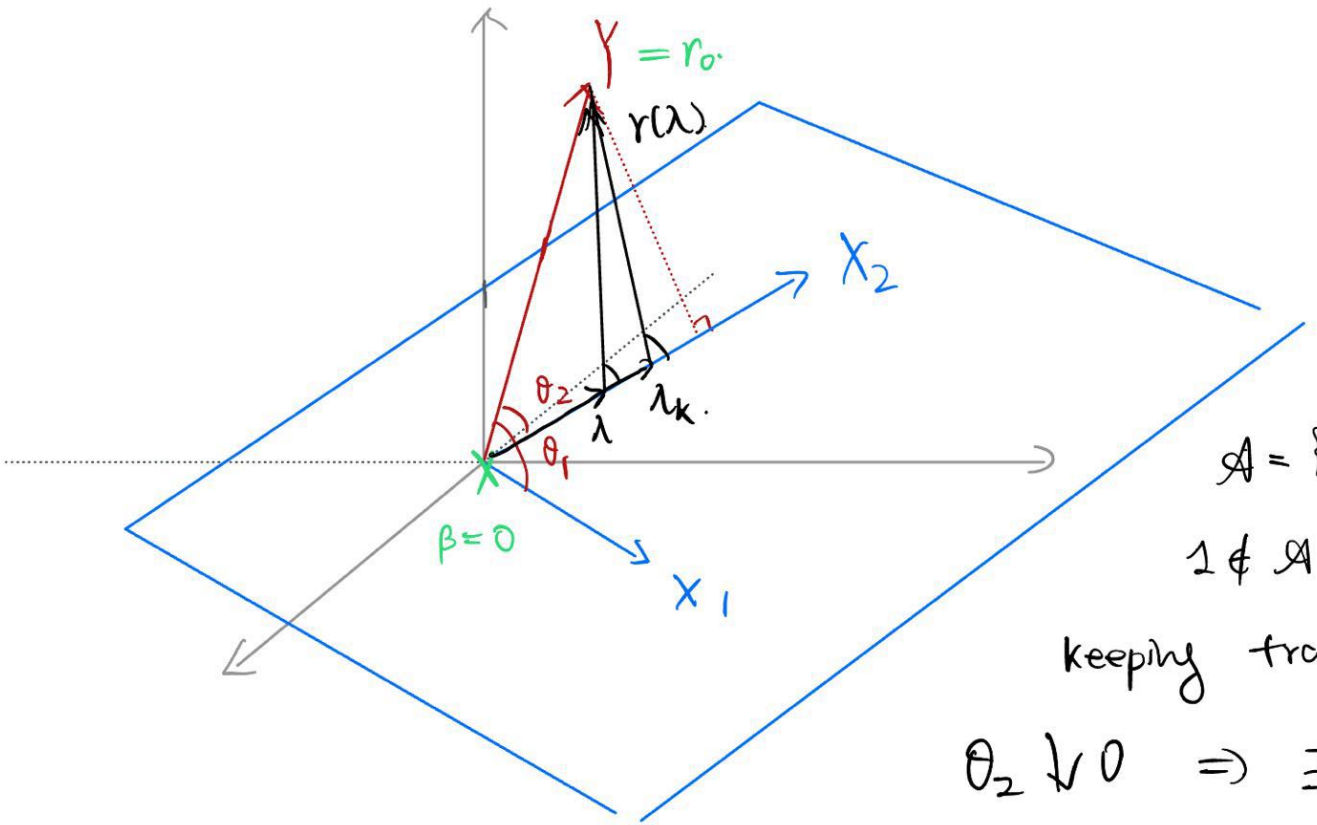
Step 3

- least-squares directions



Step 4

- decrease λ from λ_{k-1} to 0
- add new variable to the active set



$\mathcal{A} = \{2\}$

$1 \notin \mathcal{A}$

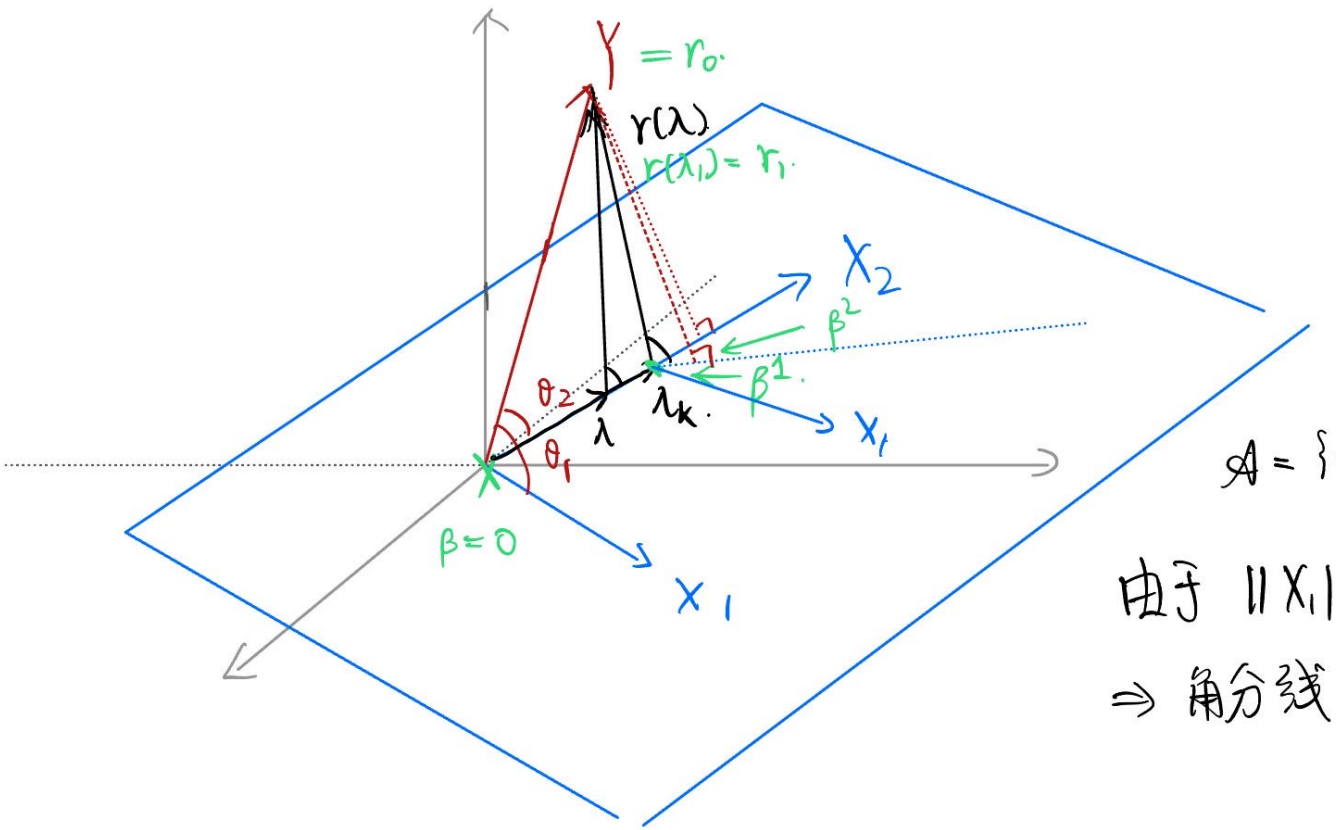
keeping track of $|\langle x_1, r(\lambda)\rangle|$.

$\theta_2 \neq 0 \Rightarrow \exists\ \underbrace{\text{a bit } \lambda}$ s.t. $\theta_2 = \theta_1$.

$\lambda_k \Leftarrow \lambda$.

Step 5

- continue until all variables are in the model (OLS)



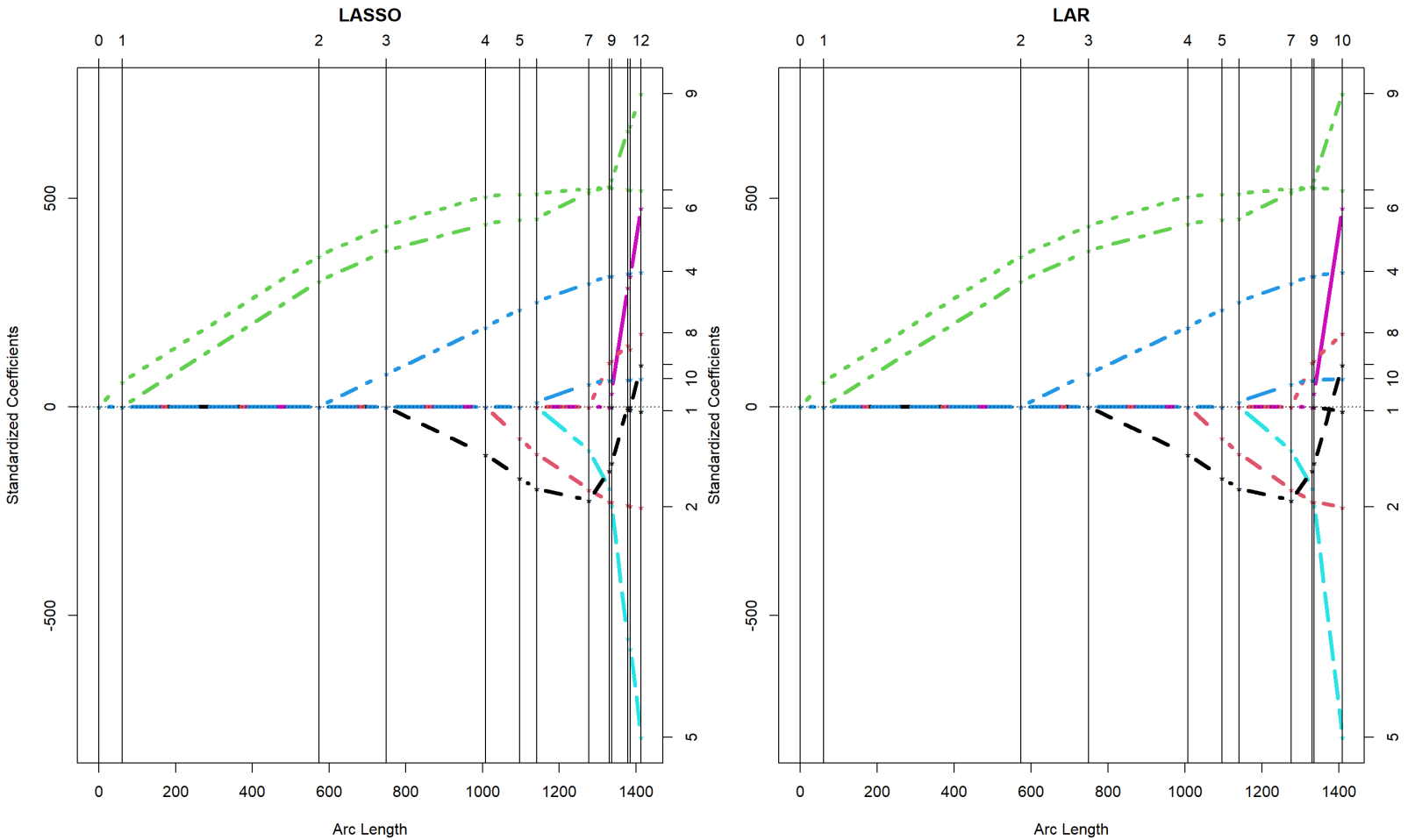
$A = \{1, 2\}$

由于 $\|X_1\| = \|X_2\|$

$\Rightarrow$ 角分线

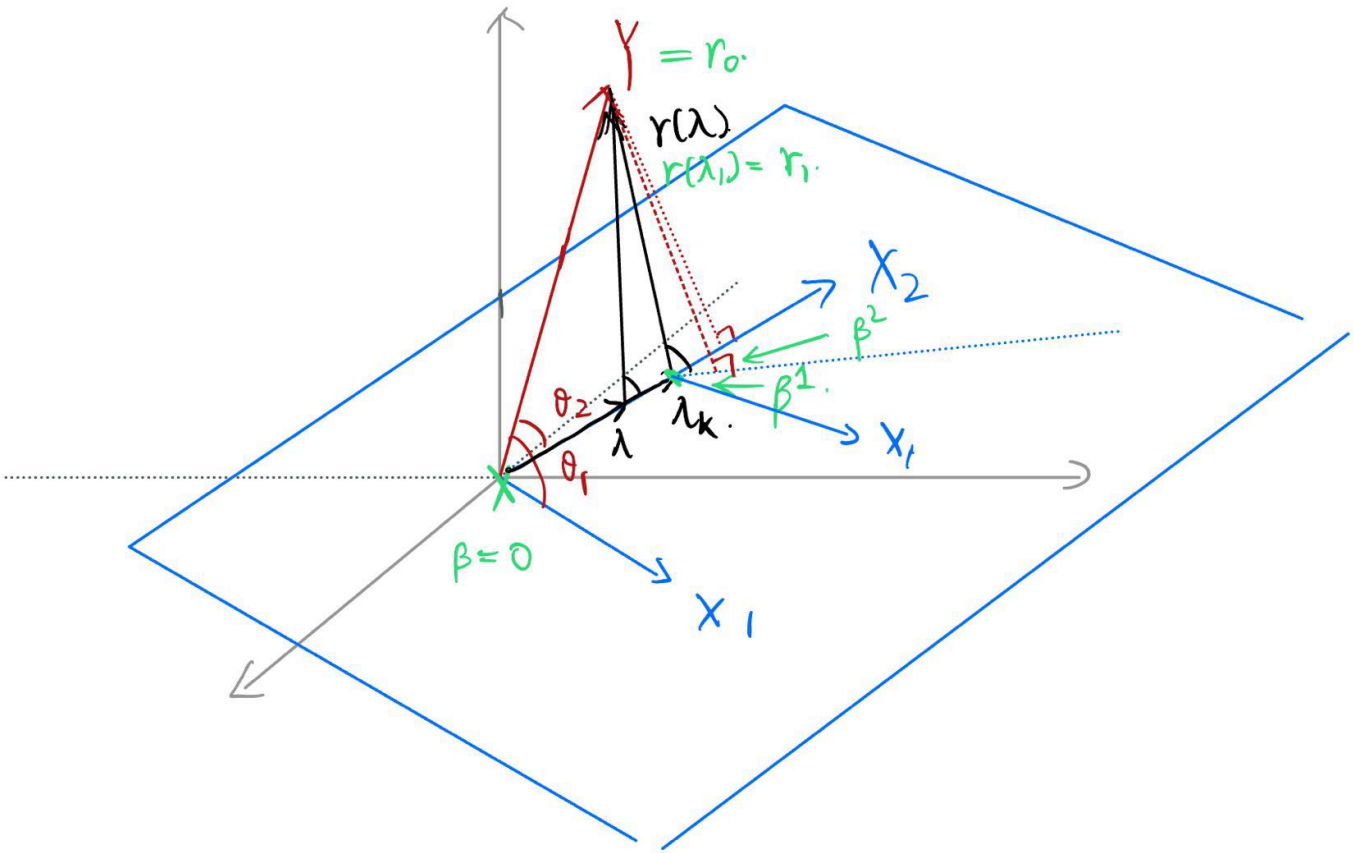
LARS comments

- 为什么叫最小角回归? $\rightsquigarrow$ 每次找到的角 (相关系数的绝对值) 是最小的.
- 和 lasso 解的轨迹十分相似. arc length 就是走的弧长.



LARS modification

- 共同性: 一阶条件.
- lasso: $X_j^\top (y - X\beta) = \lambda \cdot \partial|\beta_j|, \forall j$
- LARS: $X_j^\top (y - X\beta) = \nu \cdot s_j, s_j \in \{-1, 1\}$



LARS modification

3(c)+ *lasso modification*: If a nonzero coefficient crosses zero before the next variable enters, drop it from $\mathcal{A}$ and recompute the current joint least-squares direction.

LASSO MODIFICATION. If $\tilde{\gamma} < \hat{\gamma}$, stop the ongoing LARS step at $\gamma = \tilde{\gamma}$ and remove $\tilde{j}$ from the calculation of the next equiangular direction. That is,

$$\hat{\boldsymbol{\mu}}_{\mathcal{A}_+} = \hat{\boldsymbol{\mu}}_{\mathcal{A}} + \tilde{\gamma}\mathbf{u}_{\mathcal{A}} \quad \text{and} \quad \mathcal{A}_+ = \mathcal{A} - \{\tilde{j}\} \tag{3.6}$$

rather than (2.12).

THEOREM 1. *Under the Lasso modification, and assuming the "one at a time" condition discussed below, the LARS algorithm yields all Lasso solutions.*

Proof: refer P437 of the LARS paper.

Uniqueness of the Lasso Solutions

Ryan Tibshirani [3] considered the uniqueness of the lasso solutions. The paper focused on LARS algorithm based solutions and gives a proof of the uniqueness of the lasso solutions.
Using theories in standard convex optimization, we can show that for any y, X, and $\lambda \geq 0$, the lasso problem (in Lagrangian) has the following properties:

1. There is either a unique lasso solution or an (uncountably) infinite number of solutions.
2. Every lasso solution $\hat{\beta}$ gives the same fitted value $X\hat{\beta} \rightsquigarrow$ unique optimal value.
3. If $\lambda > 0$, then every lasso solution $\hat{\beta}$ has the same ℓ_1 norm, $\|\hat{\beta}\|_1$.

[3] Ryan J. Tibshirani. *The Lasso Problem and Uniqueness*. arXiv:1206.0313. Nov. 2012

Uniqueness of the Lasso Solutions

Proof. (i) The lasso criterion is convex and has no directions of recession (strictly speaking, when $\lambda = 0$ the criterion can have directions of recession, but these are directions in which the criterion is constant). Therefore it attains its minimum over $\mathbb{R}^p$ (see, for example, Theorem 27.1 of Rockafellar (1970)), that is, the lasso problem has at least one solution. Suppose now that there are two solutions $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$, $\hat{\beta}^{(1)} \neq \hat{\beta}^{(2)}$. Because the solution set of a convex minimization problem is convex, we know that $\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)}$ is also a solution for any $0 < \alpha < 1$, which gives uncountably many lasso solutions as α varies over $(0, 1)$.

(ii) Suppose that we have two solutions $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ with $X\hat{\beta}^{(1)} \neq X\hat{\beta}^{(2)}$. Let c^* denote the minimum value of the lasso criterion obtained by $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}$. For any $0 < \alpha < 1$, we have

$$\frac{1}{2}\|y - X(\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)})\|_2^2 + \lambda\|\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)}\|_1 < \alpha c^* + (1-\alpha)c^* = c^*,$$

where the strict inequality is due to the strict convexity of the function $f(x) = \|y - x\|_2^2$ along with the convexity of $f(x) = \|x\|_1$. This means that $\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)}$ attains a lower criterion value than c^*, a contradiction.

(iii) By (ii), any two solutions must have the same fitted value, and hence the same squared error loss. But the solutions also attain the same value of the lasso criterion, and if $\lambda > 0$, then they must have the same ℓ_1 norm. □

Theoretical Properties of the Lasso

Mean-Squared Error Consistency

With high probability:

$$\frac{1}{N}\left\|X\left(\hat{\beta}-\beta_0\right)\right\|_2^2 \lesssim \|\beta_0\|_1 \sqrt{\frac{\log(p)}{N}}$$

- Requires $\|\beta_0\|_1 = o(\sqrt{N/\log(p)})$ (sparse true parameters) to obtain MSE consistency
- No additional conditions on design matrix X

Proof of MSE Consistency

For any coefficient vector $\beta \in \mathbb{R}^d$,

$$\frac{1}{2}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Simply rearranging,

$$\frac{1}{2}\|Y - X\hat{\beta}\|_2^2 - \frac{1}{2}\|Y - X\beta\|_2^2 \leq \lambda\left(\|\beta\|_1 - \|\hat{\beta}\|_1\right)$$

Then adding and subtracting $X\beta$ in the leftmost term, and expanding the square, we get

$$\frac{1}{2}\|X\hat{\beta} - X\beta\|_2^2 \leq \langle Y - X\beta, X\hat{\beta} - X\beta\rangle + \lambda\left(\|\beta\|_1 - \|\hat{\beta}\|_1\right)$$

where we have moved the inner product term to the right-hand side. This is true for any vector β. Taking $\beta = \beta_0$ in particular and recognizing $Y - X\beta_0 = \epsilon$, the noise vector, we get from the last display our basic inequality for $\hat{\beta}$,

$$\frac{1}{2}\left\|X\hat{\beta} - X\beta_0\right\|_2^2 \leq \left\langle \epsilon, X\hat{\beta} - X\beta_0\right\rangle + \lambda\left(\|\beta_0\|_1 - \|\hat{\beta}\|_1\right)$$

Proof of MSE Consistency(Cont'd)

A result on the in-sample prediction risk for the lasso is only a few lines away. Observe that

$$\begin{aligned}\left\langle \epsilon, X\hat{\beta} - X\beta_0 \right\rangle &= \left\langle X^\top \epsilon, \hat{\beta} - \beta_0 \right\rangle \\ &\leq \left\| X^\top \epsilon \right\|_\infty \left\| \hat{\beta} - \beta_0 \right\|_1 .\end{aligned}$$

Thus from the last page, we learn that

$$\frac{1}{2}\left\| X\hat{\beta} - X\beta_0 \right\|_2^2 \leq \left\| X^\top \epsilon \right\|_\infty \left\| \hat{\beta} - \beta_0 \right\|_1 + \lambda \left(\|\beta_0\|_1 - \|\hat{\beta}\|_1 \right),$$

and using the triangle inequality,

$$\begin{aligned}\frac{1}{2}\left\| X\hat{\beta} - X\beta_0 \right\|_2^2 &\leq \left\| X^\top \epsilon \right\|_\infty \left(\|\hat{\beta}\|_1 + \|\beta_0\|_1 \right) + \lambda \left(\|\beta_0\|_1 - \|\hat{\beta}\|_1 \right) \\ &\leq 2\lambda \|\beta_0\|_1\end{aligned}$$

where the second line holds if we take $\lambda \geq \left\| X^\top \epsilon \right\|_\infty$.

Proof of MSE Consistency(Cont'd)

Then we can conduct concentration analysis on $\left\|X^\top \epsilon\right\|_\infty$ to get a high-probability bound on the quantity.
Note that $X^\top \epsilon$ has sub-Gaussian entries with mean zero and variance proxy $\max_{j=1,\ldots,p} \|X_j\|_2^2 \sigma^2 \leq n\sigma^2$. By a result on the maximum of sub-Gaussian random variables,

$$\mathbb{P}\left(\left\|X^\top \epsilon\right\|_\infty \geq \sigma\sqrt{2n(\log(2p)+u)}\right) \leq e^{-u}$$

for any $u > 0$. Therefore, taking $\lambda = \sigma\sqrt{2n(\log(2p)+u)}$, we get

$$\frac{1}{N}\left\|X\hat{\beta} - X\beta_0\right\|_2^2 \leq 4\sigma \|\beta_0\|_1 \sqrt{\frac{2(\log(2p)+u)}{N}}$$

with probability at least $1 - e^{-u}$. This bound yields what is called the "slow" rate for the penalized lasso estimator: the in-sample prediction risk scales as $\|\beta_0\|_1 \sqrt{(\log p)/N}$.

Variable-Selection Consistency for the Lasso

Theorem

Under regular conditions[a], For a noise vector $\varepsilon \in \mathbb{R}^N$ with i.i.d. $N\left(0, \sigma^2\right)$ entries, consider the regularized (Lagrangian) lasso program with sufficient large λ:

$$\lambda_N \geq \frac{8 K_{\mathrm{clm}} \sigma}{\gamma} \sqrt{\frac{\log p}{N}}.$$

Then with probability greater than $1 - c_1 e^{-c_2 N \lambda_N^2}$, the lasso has the following properties:

1. Uniqueness: The optimal solution $\widehat{\beta}$ is unique.
2. No false inclusion: The unique optimal solution has its support $S(\widehat{\beta})$ contained within the true support $S(\beta^*)$.

[a] Martin J. Wainwright. *Sharp thresholds for high-dimensional and noisy recovery of sparsity*. arXiv:math/0605740. May 2006

Variable-Selection Consistency for the Lasso(Cont'd)

Theorem(Cont'd)

With probability greater than $1 - c_1 e^{-c_2 N \lambda_N^2}$, the lasso has the following properties:

4. ℓ_∞-bounds: The error $\widehat{\beta} - \beta^*$ satisfies the ℓ_∞ bound

$$\left\| \widehat{\beta}_S - \beta^*_S \right\|_\infty \leq \underbrace{\lambda_N \left[\frac{4\sigma}{\sqrt{C_{\min}}} + \left\| \left(\mathbf{X}_S^T \mathbf{X}_S / N \right)^{-1} \right\|_\infty \right]}_{B(\lambda_N, \sigma; \mathbf{X})}$$

   where for a matrix $\mathbf{A}$, its ∞-norm is given by $\|\mathbf{A}\|_\infty = \max_{\|u\|_\infty = 1} \|\mathbf{A}u\|_\infty$.

5. No false exclusion: The lasso solution includes all indices $j \in S(\beta^*)$ such that $\left|\beta_j^*\right| > B(\lambda_N, \sigma; \mathbf{X})$, and hence is variable selection consistent as long as $\min_{j \in S} \left|\beta_j^*\right| > B(\lambda_N, \sigma; \mathbf{X})$.

Consistency Conditions: Strong Convexity

- In OLS, analysis of the convergence of the optimal value to the true value is based on the strong convexity of the loss function (objective function).
- It requires that the objective function not to be too flat around the true value.
- $$f\left(\theta'\right) - f(\theta) \geq \nabla f(\theta)^T \left(\theta' - \theta\right) + \frac{\gamma}{2} \left\|\theta' - \theta\right\|_2^2$$
- Standard convergence rates for optimization algorithms are based on the strong convexity of the objective function[4].
- However, in high-dimensional statistics, the loss function is not strongly convex due to $\mathbf{X}^\top \mathbf{X}$ is rank-deficient.

[4] Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. en. Version 29. Cambridge New York Melbourne New Delhi Singapore: Cambridge University Press, 2023

Strong Convexity Violated

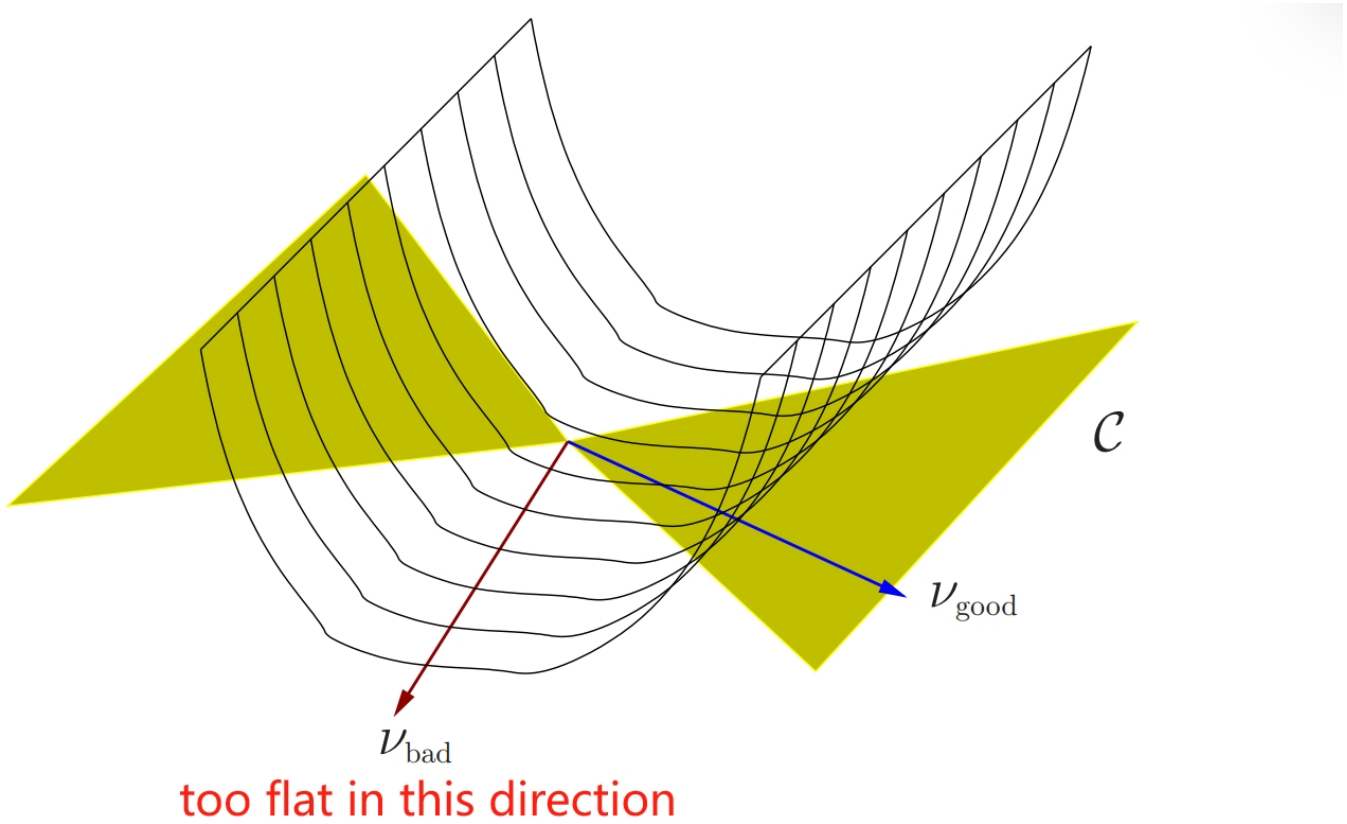


图: A convex loss function in high-dimensional settings (with $p \gg N$) cannot be strongly convex; rather, it will be curved in some directions but flat in others. It can be shown that, the lasso error $\widehat{\nu} = \widehat{\beta} - \beta^*$ must lie in a restricted subset $\mathcal{C}$ of $\mathbb{R}^p$. For this reason, it is only necessary that the loss function be curved in certain directions of space.

Restricted Eigenvalue Condition

For some subset $\mathcal{C} \subset \mathbb{R}^p$ of possible perturbation vectors $\nu \in \mathbb{R}^p$. In particular, we say that a function f satisfies restricted strong convexity at β^* with respect to $\mathcal{C}$ if there is a constant $\gamma > 0$ such that

$$\frac{\nu^T \nabla^2 f(\beta) \nu}{\|\nu\|_2^2} \geq \gamma \text{ for all nonzero } \nu \in \mathcal{C},$$

and for all $\beta \in \mathbb{R}^p$ in a neighborhood of β^*.
Let $\widehat{\nu}_S \in \mathbb{R}^{|S|}$ denote the subvector indexed by elements of S, with $\widehat{\nu}_{S^c}$ defined in an analogous manner. For appropriate choices of the ℓ_1- ball radius-or equivalently, of the regularization parameter λ_N-it turns out that the lasso error satisfies a cone constraint of the form

$$\|\widehat{\nu}_{S^c}\|_1 \leq \alpha \|\widehat{\nu}_S\|_1 ,$$

for some $\alpha \geq 1$.

Consistency Results

Consistency

Under suitable restricted eigenvalue condition,

1. Then any estimate $\widehat{\beta}$ based on the constrained lasso (11.2) with $R = \|\beta^*\|_1$ satisfies the bound

$$\left\|\widehat{\beta} - \beta^*\right\|_2 \leq \frac{4}{\gamma}\sqrt{\frac{k}{N}}\left\|\frac{\mathbf{X}^T\mathbf{w}}{\sqrt{N}}\right\|_\infty$$

2. Given a regularization parameter $\lambda_N \geq 2\left\|\mathbf{X}^T\mathbf{w}\right\|_\infty / N > 0$, any estimate $\widehat{\beta}$ from the regularized lasso (11.3) satisfies the bound

$$\left\|\widehat{\beta} - \beta^*\right\|_2 \leq \frac{3}{\gamma}\sqrt{\frac{k}{N}}\sqrt{N}\lambda_N$$

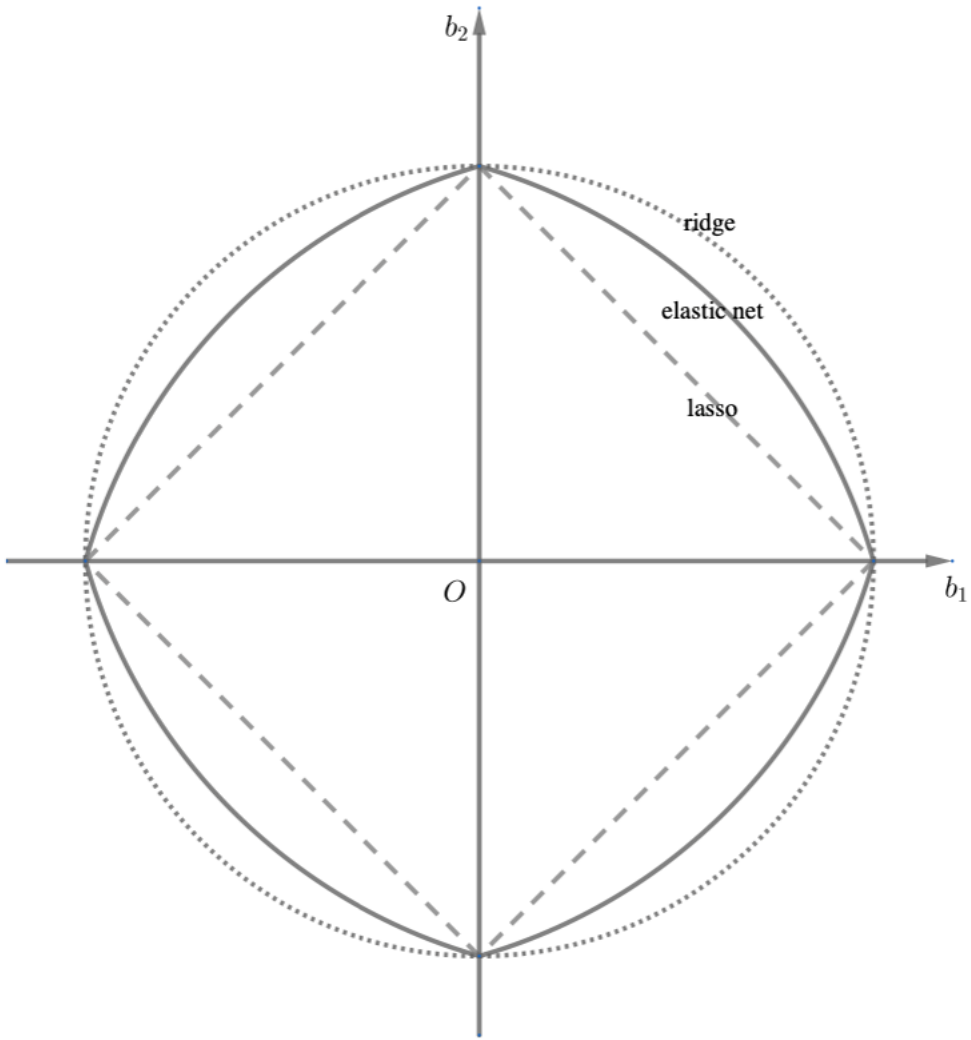
Elastic Net

- Zou and Hastie (2005) proposed the elastic net, which combines the penalties of the lasso and ridge gives $\hat{\beta}^{\text{enet}}(\lambda, \alpha)$

$$\min_{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p} \left\{ \frac{1}{2}\sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T\beta\right)^2 + \lambda \left[\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right] \right\}.$$

- Because the constraint is not smooth, it encourages sparse solution as the lasso.
- Due to the ridge penalty, it can deal with collinearity of the covariates better than the lasso.
- Implemented in `glmnet` package

Elastic Net



Nonnegative Garrote

- Obtain initial estimate $\tilde{\beta}$ (OLS, lasso, ridge, elastic net)
- Solve optimization problem:

$$\min_{c \in \mathbb{R}^p} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} c_j x_{ij} \tilde{\beta}_j \right)^2 \right\}$$

  subject to $c \succeq 0$ and $\|c\|_1 \leq t$
- Final estimate: $\widehat{\beta}_j = \hat{c}_j \cdot \tilde{\beta}_j$

Orthogonal Case

When columns of X are orthogonal:

$$\widehat{c}_j = \left(1 - \frac{\lambda}{\tilde{\beta}_j^2}\right)_+, \quad j = 1, \ldots, p$$

where λ is chosen so that $\|\hat{c}\|_1 = t$.

- Large coefficients: minimal shrinkage (near 1)
- Small coefficients: severe shrinkage (toward 0)
- Exhibits nonconvex penalty behavior