Sparse Linear Model

Wei Fan, Jingguo Lan

University of Science and Technology of China

November 27, 2021

Outline

1. Variable Selection
2. Folded-Concave Penalization
3. Lasso and its Generalizations
   - Lasso
   - Adaptive Lasso
   - SLOPE
   - Elastic Net
   - Group Lasso
   - Fused Lasso
4. Optimization Algorithms for Lasso
   - Coordinate Descent
   - ADMM
   - Proximal Gradient Methods
   - Least Angle regression

I. Introduction

- This is often realized by putting a penalty on the objective function, e,g, the Tikhonov regularization

$$\min_{\boldsymbol{\beta}} ||\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}||_2^2, \text{ s.t. } ||\beta||_2 \le \mathrm{C} \tag{1}$$

- The regularization term could significant reduce the condition number of the above question, hence enhance the computational stability.

I. Introduction

- This is often realized by putting a penalty on the objective function, e,g, the Tikhonov regularization

$$\min_{\boldsymbol{\beta}} ||\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}||_2^2, \text{ s.t. } ||\beta||_2 \leq \mathrm{C} \tag{2}$$

- The regularization term could significant reduce the condition number of the above question, hence enhance the computational stability.

I. Introduction

- The penalty term could be generalized the L^q case, which is

$$||\beta||_q = \left(\sum_{j=1}^{p} |\beta_j|^q \right)^q .$$

- The case with $q = 2$ is essential the Tikhonov regularization (or ridge regression). Although, this lacks interpretation (except for Bayesian explanation).
- When $0 \leq p \leq 1$, the estimator of β become sparse, which is a more reasonable solution for the practical case like:
  - Prostate Cancer Data (see [1]).
  - Genomes with certain features.

I. Introduction

- If q = 0, we are aimed to solve

$$\min_{\beta} \|\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}\|^2, \text{s.t. } \|\boldsymbol{\beta}\|_0 = \mathrm{m} \tag{3}$$

which is the best subset section.

- It's worth noting that $0 \leq \mathrm{q} < 1$ is not a convex optimization problem, which possesses certain difficulties.

- No matter what method you use to construct the estimation of β, we always need to proceed with the model selection step.
- We define

$$\mathrm{RSS_m} = ||\mathrm{Y} - \mathrm{X}\hat{\beta}_\mathrm{m}||_2^2,$$

where m is the model complexity term here, which is defined as

$$\mathrm{m} = \frac{1}{\sigma^2}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Cov}(\hat{\mathrm{Y}}_\mathrm{i}, \mathrm{Y}_\mathrm{i})$$

One can show that m is the trace of the projection matrix for Y on the ridge regression case. And m = s if we assume that only s predictors are used in LS.

II. Variable Selection

- Mallow's C_p criterion: $C_p(m) = RSS_m + 2\sigma^2 m$.
  - (X^*, Y^*) be a completely new observation, the prediction error using model $\mathcal{M}_m$ is

$$PE(\mathcal{M}_m) = E(Y^* - \hat{\boldsymbol{\beta}}_m^T X^*_{\boldsymbol{\mathcal{M}}_m})^2$$

  - An unbiased estimation of prediction error is $C_p(m)$.

II. Variable Selection

- Information Criterion: $\mathrm{IC(m)} = \log(\mathrm{RSS_m/n}) + \lambda \mathrm{m/n}$, is asymptotically equivalent to $\mathrm{C_p}$ criterion with Taylor expansion.
  - Akaike information criterion (AIC): $\lambda = 2$.
  - Bayesian information criterion (BIC): $\lambda = \log(\mathrm{n})\sigma^2$.
  - $\phi-$criterion: $\lambda = \mathrm{c}(\log\log \mathrm{n})$.
  - Risk Inflation Criterion (RIC): $\lambda = 2\log(\mathrm{p})$.
- Other possible methods: Cross Validation, adjusted R^2.

- In general, the L_p penalty with $0 \le p \le 1$ is not the unique way to produce the sparse solution. We can generalize the regularization functions as

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\|Y - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{4}$$

$$= \frac{1}{2n}\|Y - X\boldsymbol{\beta}\|^2 + \|p_\lambda(|\boldsymbol{\beta}|)\|_1 \tag{5}$$

- $p_\lambda(\cdot)$ is a penalty function in which the regularization parameter is λ.

III. Folded-Concave Penalization - Introduction

- Examples of choosing $\mathrm{p}_\lambda(\cdot)$ includes
  - L_0 penalty: $\mathrm{p}_\lambda(\theta) = \lambda \mathrm{I}(\theta \neq 0)$.
  - L_2 penalty: $\mathrm{p}_\lambda(\theta) = \lambda \theta^2/2$, whose solution is ridge regression

$$\hat{\boldsymbol{\beta}}_{\mathrm{ridge}} = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \mathrm{n}\lambda \mathrm{I}_{\mathrm{p}})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}.$$

  - Bridge Regression: $\mathrm{p}_\lambda(\theta) = \lambda|\theta|^{\mathrm{q}}$ $(0 < \mathrm{q} < 2)$, $\mathrm{q} = 1$ correspond to Lasso.
  - SCAD, MCP…

III. Folded-Concave Penalization - Choice of Penalization

- We consider the case when $X^T X = nI_p$, then Equation (4) will reduce to
$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\|Y - X\hat{\boldsymbol{\beta}}\|^2 + \frac{1}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 + \|p_\lambda(|\boldsymbol{\beta}|)\|_1 \tag{6}$$
where $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y = n^{-1} X^T Y$ is the OLS estimate.
- The minimization of Equation (6) is equivalent to minimizing
$$\sum_{j=1}^{p} \left\{ \frac{1}{2}(\hat{\beta}_j - \beta_j)^2 + p_\lambda(|\beta_j|) \right\}.$$
- Here we define
$$\phi_{z,\lambda}(\theta) = \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|) \tag{7}$$
and the univariate PLS problem as
$$\hat{\theta}(z) = \arg\min_{\theta} \phi_{z,\lambda}(\theta) \tag{8}$$

- We assume $p_\lambda(t)$ is non-decreasing and continuously differentiable on $[0, \infty]$. For $\theta > 0$, the derivative is

$$\phi'_{z,\lambda}(\theta) = \theta + p'_\lambda(\theta) - z.$$

A good penalty function give estimators with following properties:

1. Sparsity: The estimator automatically set small estimated coefficients to zero and reduce model complexity.
   - Correspond penalty: $\min_{t \geq 0} \{t + p'_\lambda(t)\} > 0$, which holds if $p'_\lambda(0+) > 0$.
2. Approximate unbiasedness: The estimator is nearly unbiased, especially when true β_j is large.
   - Correspond penalty: If $p'_\lambda(t) = 0$ for large t (i.e. for $t > a\lambda$ for some a).
3. Continuity: The estimator is continuous in the data to reduce instability.
   - Correspond penalty: If and only if $\arg\min_{t \geq 0} \{t + p'_\lambda(t)\} = 0$.

Property 1 and 2 is required for a folded-concave function.

III. Folded-Concave Penalization - Choice of Penalization

Some common choice of folded-concave penalty functions

- Smoothly Clipped Absolute Deviation (SCAD, Fan, [2])

$$p'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}.$$

  Choice of parameter: a=3.7.

- Hard-Thresholding Penalty (Bogdan, [3])

$$p'_\lambda(t) = (\lambda - t)_+.$$

- Minimax Concave Penalty (MCP, Zhang, [4])

$$p'_\lambda(t) = (\lambda - t/a)_+.$$

  Choice of parameter: a=2.

III. Folded-Concave Penalization - Properties

- Properties of Solution
  - Soft Thresholding (Lasso penalty)

$$\hat{\theta}_{\text{soft}}(\mathrm{z}) = \operatorname{sgn}(\mathrm{z})(|\mathrm{z}| - \lambda)_+$$

  - SCAD Penalty

$$\hat{\theta}_{\text{SCAD}}(\mathrm{z}) = \begin{cases} \operatorname{sgn}(\mathrm{z})(|\mathrm{z}| - \lambda)_+, & |\mathrm{z}| \le 2\lambda. \\ \operatorname{sgn}(\mathrm{z})[(\mathrm{a}-1)|\mathrm{z}| - \mathrm{a}\lambda]/(\mathrm{a}-2), & 2\lambda < |\mathrm{z}| \le \mathrm{a}\lambda. \\ \mathrm{z}, & |\mathrm{z}| > \mathrm{a}\lambda. \end{cases} \tag{9}$$

    As $\mathrm{a} = \infty$, SCAD estimator becomes soft-thresholding estimator.
  - MCP Penalty

$$\hat{\theta}_{\text{MCP}}(\mathrm{z}) = \begin{cases} \operatorname{sgn}(\mathrm{z})(|\mathrm{z}| - \lambda)_+/(1 - 1/\mathrm{a}), & |\mathrm{z}| \le \mathrm{a}\lambda. \\ \mathrm{z}, & |\mathrm{z}| > \mathrm{a}\lambda. \end{cases} \tag{10}$$

    It discontinues at $|\mathrm{z}| = \lambda$, making the model instable. As $\mathrm{a} = 1$, MCP estimator becomes hard thresholding estimator.

III. Folded-Concave Penalization - Choice of Penalization

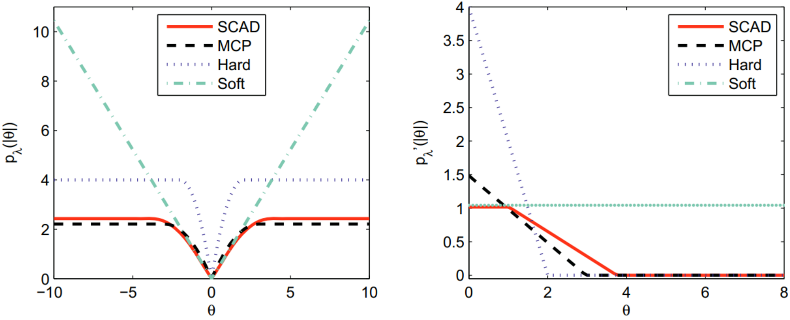


Figure: Some commonly used penalty functions (left panel) and their derivatives (right panel). More precisely, $\lambda = 2$ for hard thresholding penalty, $\lambda = 1.04$ for L_1-penalty, $\lambda = 1.02$ for SCAD with a $= 3.7$, and $\lambda = 1.49$ for MCP with a $= 2$. Taken from [5].

III. Folded-Concave Penalization - Properties

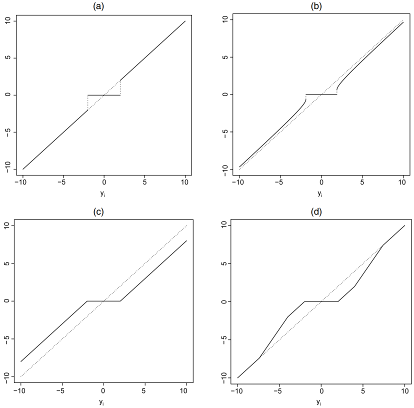


Figure: Plot of thresholding functions with $\lambda = 2$ for (a) the hard; (b) the Bridge ($L_{0.5}$); (c) the Lasso; (d) the SCAD.

- Risk Properties
  - Let $\mathrm{R}(\theta) = \mathrm{E}(\hat{\theta}(\mathrm{Z}) - \theta)^2$ be the risk function of estimator $\hat{\theta}(\mathrm{Z})$.
  - We investigate the risk function of threshold-shrinkage estimators under the model $\mathrm{N}(0, 1)$.

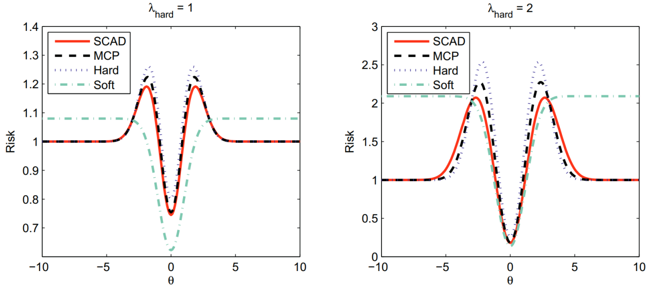


Figure: The risk functions for penalized least squares under the Gaussian model. The left panel corresponds to $\lambda = 1$ and the right panel corresponds to $\lambda = 2$ for the hard-thresholding estimator. Adapted from [5].

- For least square regression, we hope our estimator have following properties
  - Prediction accuracy: we also need to reduce the variance, by shrinking the values of regression coefficients or setting some to be zero.
  - Interpretation: we often would like to identify a smaller subset of these predictors that exhibit the strongest effects.
- Solution: Lasso(Tibshirani, [6]), consider the optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|Y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{11}$$

- Advantage of Lasso: convexity, selection of coefficients.
- Necessary and sufficient condition for solution:

$$-\frac{1}{n}\langle x_j, y - X\beta\rangle + \lambda s_j = 0,\ j = 1, \dots, p.$$

Here s_j is the (sub-)gradient of $|\beta_j|$, which is $\mathrm{sign}(\beta_j)$ for $\beta_j \neq 0$ and every number within $[-1, 1]$ for $\beta_j = 0$.

A direct interpretation of selection property of Lasso.

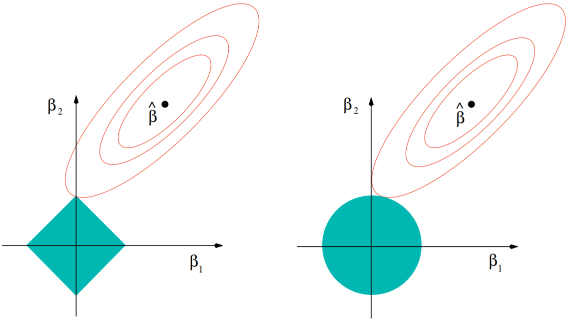


Figure: Estimation picture for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq$ t and $\beta_1^2 + \beta_2^2 \leq \mathrm{t}_2$, respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point $\widehat{\beta}$ depicts the usual (unconstrained) least-squares estimate.

An example of Lasso vs Ridge regression.

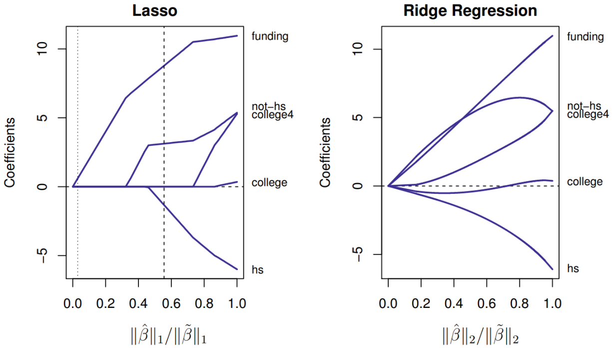


Figure: Left: Coefficient path for the lasso, plotted versus the L_1 norm of the coefficient vector, relative to the norm of the unrestricted least-squares estimate $\tilde{\beta}$. Right: Same for ridge regression, plotted against the relative L_2 norm.

- Irrepresentable condition of Lasso
  - Let $\boldsymbol{\beta}_0$ be the true regression coefficient and $\mathcal{S}_0 = \text{supp}(\boldsymbol{\beta}_0)$, if the Sign consistency holds (i.e. $\text{sgn}(\hat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta}_0)$). We have the following irrepresentable condition

$$\left\| \mathrm{X}_2^{\mathrm{T}} \mathrm{X}_1 \left(\mathrm{X}_1^{\mathrm{T}} \mathrm{X}_1 \right)^{-1} \text{sgn}\left(\boldsymbol{\beta}_{\mathcal{S}_0} \right) \right\|_\infty \leq 1. \tag{12}$$

  - $\left(\mathrm{X}_1^{\mathrm{T}} \mathrm{X}_1\right)^{-1} \mathrm{X}_1^{\mathrm{T}} \mathrm{X}_2$ is the matrix of the regression coefficients of each 'unimportant' variable X_j ($\mathrm{j} \notin \mathcal{S}_0$) regressed on the important variables $\mathrm{X}_1 = \mathrm{X}_{\mathcal{S}_0}$, showing how strongly the important and unimportant variables can be correlated.
  - When irrepresentable condition fails, Lasso does not have sign consistency and this cannot be rescued by using a different value of λ.

- Drawbacks of Lasso:
  - Irrepresentable conditions.
  - Lack of unbiasedness for large coefficients.
- Fix: Adaptive Lasso(Zou, [7]), using adaptive weighted L_1 penalty.

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|Y - X\boldsymbol{\beta}\|^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j|. \tag{13}$$

  Here w_j is non-negative and not fixed.
- Redefine $X_j^w = X_j/w_j$ and $\theta_j = w_j\beta_j$, then we rewrite (13) as

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2n}\|Y - X^w\boldsymbol{\theta}\|^2 + \lambda\sum_{j=1}^{p} |\theta_j|. \tag{14}$$

- Irrepresentable condition:
  - Original Form:

$$\|(X_2^w)^T X_1^w [(X_1^w)^T X_1^w]^{-1} \mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}_0})\|_\infty \leq 1 \tag{15}$$

  - We rewrite $W = (w_1, w_2, \ldots, w_p)^T = (W_1, W_2)^T$. We express (15) with original variables

$$\|[(X_2)^T X_1 (X_1^T X_1)^{-1} W_1 \circ \mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}_0})] \circ W_2^{-1}\|_\infty \leq 1 \tag{16}$$

  - As $\max W_1 / \min W_2 \to 0$, this representable condition can be satisfied for general X_1, X_2 and $\mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}_0})$.
- Construction of w_j: $w_j = |\hat{\beta}_j|^{-\gamma}$ for some $\gamma > 0$, where $\hat{\beta}_j$ is a preliminary estimation of β_j.
  - $p < n$: least-square estimate.
  - $p \gg n$: Lasso estimate.
- Adaptive Lasso Penalty can be seen as approximation to L_q penalty for $q = 1 - \nu$.

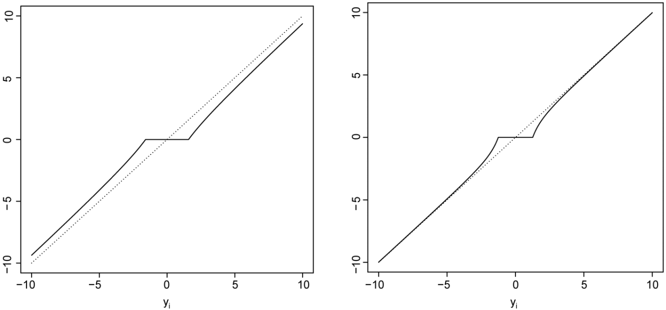


Figure: Plot of thresholding functions with $\lambda = 2$ for (left panel) Adaptive Lasso with $\gamma = 0.5$; (right panel) Adapive Lasso with $\gamma = 2$.

- We consider the model selection problem as $p > n$, which in some senses reduces to the problem of testing p hypotheses

$$H_{0,j} : \beta_j = 0 \quad \text{versus} \quad H_{1,j} : \beta_j \neq 0.$$

- The False Discovery Rate is defined as

$$FDR = \mathbb{E}[\frac{FP}{FP + TP}].$$

where "FP" stands for the case when $\beta_j = 0$ but $\hat{\beta}_j \neq 0$, "TP" stands for the case when $\beta_j \neq 0$ and $\hat{\beta}_j \neq 0$.

- SLOPE(Bogdan, [8]): Given a sequence of penalty levels $\lambda_1 \geq \lambda_2 \geq \lambda_p \geq 0$, it finds the solution of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n} \|Y - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} \lambda_j |\beta|_{(j)}. \tag{17}$$

  where $|\beta|_{(1)} \geq \cdots \geq |\beta|_{(p)}$ are order statistics of $\{|\beta|_j\}_{j=1}^{p}$.
- The FDR is controlled at level q when choosing

$$\lambda_j = \Phi^{-1}(1 - jq/2p)\sigma/\sqrt{n} \approx \sigma\sqrt{(2/n)\log(p/j)}.$$

- Moreover, we may introduce the sorted folded concave penalties as

$$\frac{1}{2n} \|Y - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p_{\lambda_j}(|\beta|_{(j)}). \tag{18}$$

IV. Lasso and its Generalizations - Elastic Net

- Introduction
  - Problem: High variability caused by spurious correlation in high dimensional data.
  - Goal: handle the strong correlations among high-dimensional variables while keeping the continuous shrinkage and selection property of the Lasso.
- Elastic Net(Zou, [9]): Find the minimization of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \|Y - X\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1. \tag{19}$$

The penalty can be rewritten as

$$p_{\lambda,\alpha}(t) = \lambda J_0(t) = \lambda \left[(1-\alpha)t^2 + \alpha|t| \right].$$

- Adaptive Elastic Net penalty

$$p(|\beta_j|) = \lambda_1 w_j |\beta_j| + \lambda_2 |\beta_j|^2,$$

where $w_j = |\hat{\beta}^{enet} + 1/n|^{-\gamma}$.

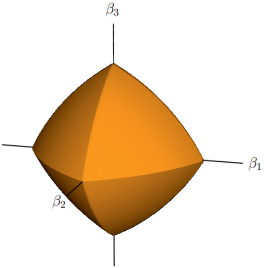


Figure: The elastic-net ball with $\alpha = 0.7$ (left panel) in $\mathbb{R}^3$, compared to the L_1 ball (right panel).

- Simulation Example (sample size $\mathrm{N} = 100$):
  - $\mathrm{Z_1}, \mathrm{Z_2} \sim \mathrm{N}(0, 1)$ independent.
  - $\mathrm{Y} = 3\mathrm{Z_1} - 1.5\mathrm{Z_2} + 2\epsilon$, with $\epsilon \sim \mathrm{N}(0, 1)$.
  - $\mathrm{X_j} = \mathrm{Z_1} + \xi_\mathrm{j}/5$, with $\xi_\mathrm{j} \sim \mathrm{N}(0, 1)$ for $\mathrm{j} = 1, 2, 3$, and
  - $\mathrm{X_j} = \mathrm{Z_2} + \xi_\mathrm{j}/5$, with $\xi_\mathrm{j} \sim \mathrm{N}(0, 1)$ for $\mathrm{j} = 4, 5, 6$.

IV. Lasso and its Generalizations - Elastic Net

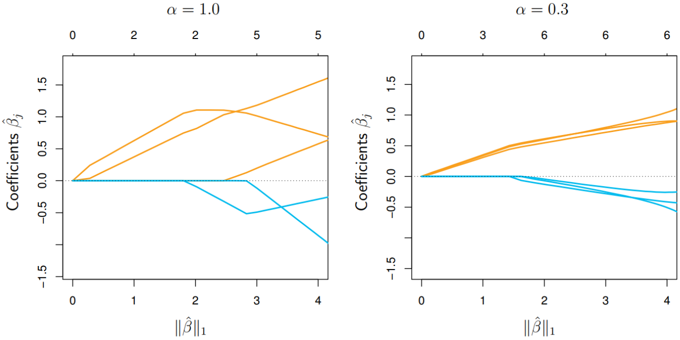


Figure: Six variables, highly correlated in groups of three. The lasso estimates ($\alpha = 1$), as shown in the left panel, exhibit somewhat erratic behavior as the regularization parameter λ is varied. In the right panel, the elastic net with ($\alpha = 0.3$) includes all the variables, and the correlated groups are pulled together.

- Introduction
  - Problem: Covariates have a natural group structure.
  - It is desirable to have all coefficients within a group become nonzero (or zero) simultaneously.
- Consider the linear regression involving J groups of covariate. Let $X = (X_1, \ldots, X_J)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$, where $X_j \in \mathbb{R}^{n \times p_j}$ represents the covariate in group j and $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$ represents its corresponding regression coefficients. Here, our linear model (??) can be rewritten as

$$Y = \sum_{j=1}^{J} X_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}. \tag{20}$$

- Group Lasso(Yuan, [10]): Consider the optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|Y - \sum_{j=1}^{J} X_j\boldsymbol{\beta}_j\|^2 + \lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2. \tag{21}$$

  where $\|\boldsymbol{\beta}_j\|_2$ is the Euclidean norm of vector $\boldsymbol{\beta}_j$.
- Properties
  - depending on $\lambda \geq 0$, in most cases either the entire vector $\boldsymbol{\beta}_j$ will be zero, or all its elements will be nonzero.
  - When $p_j = 1$ for all $1 \leq j \leq J$, the optimization problem reduces to Lasso.

IV. Lasso and its Generalizations - Group Lasso

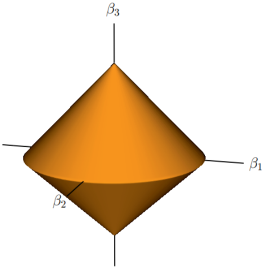


Figure: The group lasso ball (left panel) in $\mathbb{R}^3$, compared to the L_1 ball (right panel). In this case, there are two groups with coefficients $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$ and $\theta_2 = \beta_3 \in \mathbb{R}^1$.

- Example: In gene-expression arrays, we might have a set of highly correlated genes from the same biological pathway. Selecting the group amounts to selecting a pathway.

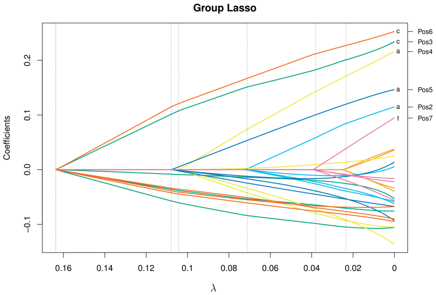


Figure: Coefficient profiles from the group lasso, fit to splice-site detection data. The coefficients come in groups of four, corresponding to the nucleotides A, G, C, T. The vertical lines indicate when a group enters. On the right-hand side we label some of the variables; for example, “Pos6” and the level “c”. The coefficients in a group have the same color, and they always average zero.

Computation for Group Lasso

- Zero subgradient equations

$$-X_j^T(Y - \sum_{\ell=1}^{J} X_\ell \hat{\boldsymbol{\beta}}_\ell) + \lambda \hat{s}_j = 0, \quad \text{for } j = 1, \ldots, J.$$

  with $\hat{s}_j$ is the subdifferential of Euclidean norm at $\hat{\boldsymbol{\beta}}_j$, we have $\hat{s}_j = \hat{\boldsymbol{\beta}}_j / \|\hat{\boldsymbol{\beta}}_j\|_2$ for $\hat{\boldsymbol{\beta}}_j \neq 0$, and any vector with $||\hat{s}_j||_2 \leq 1$ if $\hat{\boldsymbol{\beta}}_j = 0$.
- Fixing all $\left\{\hat{\boldsymbol{\beta}}_k,\ k \neq j\right\}$, we have

$$-X_j^T(r_j - X_j \hat{\boldsymbol{\beta}}_j) + \lambda \hat{s}_j = 0, \quad \text{for } j = 1, \ldots, J.$$

  where $r_j = Y - \sum_{k \neq j} X_k \hat{\boldsymbol{\beta}}_k$ is the j^{th} partial residue.
- We have $\hat{\boldsymbol{\beta}}_j = 0$ for $\|X_j^T r_j\|_2 < \lambda$ and otherwise the minimizer satisfies

$$\hat{\boldsymbol{\beta}}_j = \left(X_j^T X_j + \frac{\lambda}{\|\hat{\boldsymbol{\beta}}_j\|_2} I \right)^{-1} X_j^T r_j.$$

Generalizations of Group Lasso

- Sparse Group Lasso(Simon, [11]):
  - Motivation: We would like sparsity both with respect to which groups are selected, and which coefficients are nonzero within a group.
  - Sparse Group Lasso: Consider the Optimization Problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|Y - \sum_{j=1}^{J} X_j\boldsymbol{\beta}_j\|^2 + \lambda\sum_{j=1}^{J}[(1-\alpha)\|\boldsymbol{\beta}_j\|_2 + \alpha\|\boldsymbol{\beta}_j\|_1]. \tag{22}$$

  with $\alpha \in [0,1]$. This creates a bridge between group Lasso ($\alpha = 0$) and Lasso ($\alpha = 1$).

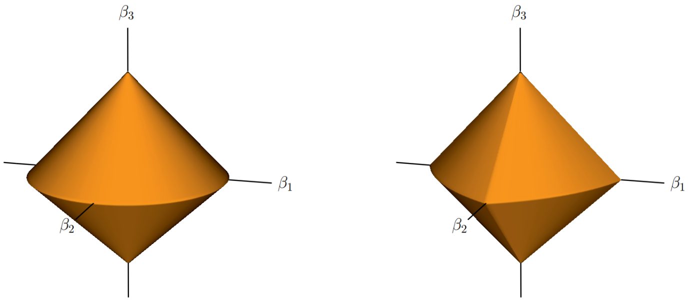


Figure: The group lasso ball (left panel) in $\mathbb{R}^3$, compared to the sparse group Lasso ball with $\alpha = 0.5$ (right panel). Depicted are two groups with coefficients $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$ and $\theta_2 = \beta_3 \in \mathbb{R}^1$.

- The Overlap Group Lasso(Jacob, [12]):
  - In some cases, variables can belong to more than one group.
  - We consider an example of partition $\mathrm{p} = 5$ variables into 2 groups, say

$$\mathrm{Z}_1 = (\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3), \text{ and } \mathrm{Z}_2 = (\mathrm{X}_3, \mathrm{X}_4, \mathrm{X}_5).$$

  - We fit coefficient vectors $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13})$ and $\theta_2 = (\theta_{21}, \theta_{22}, \theta_{23})$ using group Lasso, using a group penalty of $\|\theta_1\|_2 + \|\theta_2\|_2$, we have the following approaches for determining $\hat{\beta}_3$.
    - Let $\hat{\beta}_3 = \theta_{13} = \theta_{21}$.
    - Let $\hat{\beta}_3 = \theta_{13} + \theta_{21}$.
  - The first choice is not desirable, the nonzero combinations can only be $\{1, 2\}$, $\{4, 5\}$ and $\{1, 2, 3, 4, 5\}$, destroying its original group structure.
  - We formalize the second choice in next page.

- The Overlap Group Lasso (continued):
  - Define $\boldsymbol{\nu}_{\mathrm{j}} \in \mathbb{R}^{\mathrm{p}}$ a vector which is zero everywhere except in those positions corresponding to the members of group j. Let $\mathcal{V} \in \mathbb{R}^{\mathrm{p}}$ be the subspace of all such vectors.
  - Here overlap group Lasso solves the minimization problem

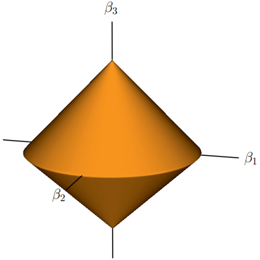
$$\min_{\boldsymbol{\nu}_{\mathrm{j}} \in \mathcal{V},\, \mathrm{j}=1,\ldots,\mathrm{J}} \frac{1}{2\mathrm{n}} \|\mathrm{Y} - \mathrm{X}(\sum_{\mathrm{j}=1}^{\mathrm{J}} \boldsymbol{\nu}_{\mathrm{j}})\|^2 + \lambda \sum_{\mathrm{j}=1}^{\mathrm{J}} \|\boldsymbol{\nu}_{\mathrm{j}}\|_2. \tag{23}$$

  - We define the following penalty

$$\Omega_{\mathcal{V}}(\boldsymbol{\beta}) = \inf_{\boldsymbol{\nu}_{\mathrm{j}} \in \mathcal{V},\, \boldsymbol{\beta} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \boldsymbol{\nu}_{\mathrm{j}}} \sum_{\mathrm{j}=1}^{\mathrm{J}} \|\boldsymbol{\nu}_{\mathrm{j}}\|_2$$

    and solving the overlap group Lasso (23) is equivalent to solving the following minimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{\mathrm{p}}} \frac{1}{2\mathrm{n}} \|\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda \Omega_{\mathcal{V}}(\boldsymbol{\beta}). \tag{24}$$



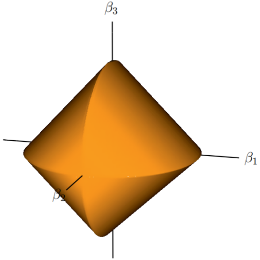


Figure: The group-lasso ball (left panel) in $\mathbb{R}^3$, compared to the overlap-group lasso ball (right panel). Depicted are two groups in both. In the left panel the groups are $\{X_1, X_2\}$ and X_3; in the right panel the groups are $\{X_1, X_2\}$ and $\{X_2, X_3\}$. There are two rings corresponding to the two groups in the right panel.When β_2 is close to zero, the penalty on the other two variables is much like the lasso. When β_2 is far from zero, the penalty on the other two variables “softens” and resembles the L_2 penalty.

IV. Lasso and its Generalizations - The Fused Lasso

- Given a sequence of very noisy data, we expect the true copy numbers need to be piecewise-constant
- The fused Lasso signal approximater(Tibshirani, [13]) exploits such structure within a signal, and is the solution of the following optimization problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \frac{1}{2n}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda_1\sum_{i=1}^{n}|\theta_i|+\lambda_2\sum_{i=2}^{n}|\theta_i-\theta_{i-1}| \tag{25}$$

where the first penalty serve to shrink θ_i towards zero and the second penalty encourage the neighboring coefficients θ_i to be similar.
- We consider a regression problem, the optimization problem will become

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|Y-X\boldsymbol{\beta}\|^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=2}^{p}|\beta_j-\beta_{j-1}| \tag{26}$$

Example

- Consider the results of a comparative genomic hybridization (CGH) experiment.
- Each of these represents the (log base 2) relative copy number of a gene in a cancer sample relative to a control sample; these copy numbers are plotted against the chromosome order of the gene.
- These data are very noisy, so that some kind of smoothing is essential.
- Typically segments of a chromosome— rather than individual genes—that are replicated, and we might expect that the underlying vector of true copy numbers to be piecewise-constant over contiguous regions of a chromosome.

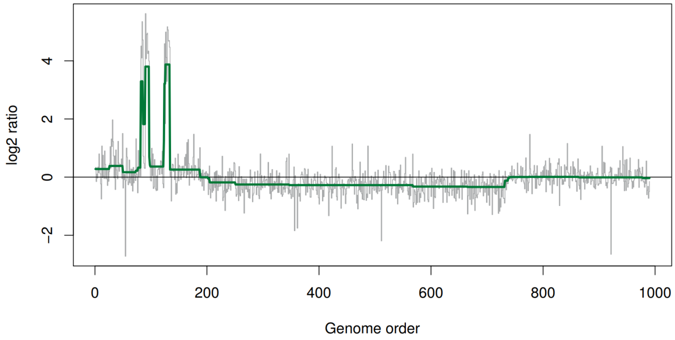


Figure: Fused lasso applied to CGH data. Each spike represents the copy number of a gene in a tumor sample, relative to that of a control (on the log base-2 scale). The piecewise-constant green curve is the fused lasso estimate.

Generalization

- We may generalize the notion of neighbors from a linear ordering to more general neighborhoods, for examples adjacent pixels in an image.
- This leads the penalty of the form

$$\lambda_2 \sum_{\mathrm{i}\sim\mathrm{i}'} |\theta_{\mathrm{i}} - \theta_{\mathrm{i}'}|.$$

  where we sum over all neighboring pairs $\mathrm{i} \sim \mathrm{i}'$.
- Example: Recover from a noisy image., taken from [14]

(b) (c)

Figure: (b) Noisy image. (c) Fused lasso estimate using 2d lattice prior.

Introduction

This section, we will introduce some optimization algorithms. And use these methods to solve the Lasso problem.
Our target is the following problem:

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|^2 \quad \text{subject to} \quad \|\beta\|_1 \le c. \tag{27}$$

It is equal to minimize:

$$f(\beta) = \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1. \tag{28}$$

Coordinate Descent

This suggests that for the problem

$$\min_{x} f(x)$$

we can use coordinate descent (CD): let $x^{(0)} \in \mathbb{R}^n$, and repeat

$$x_i^{(k)} = \underset{x_i}{\operatorname{argmin}}\ f\left(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)}\right) \qquad i = 1, \ldots, n$$

for $k = 1, 2, 3, \ldots$. Important note: we always use most recent information possible.

CD for Lasso

Consider the lasso problem:

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Note that nonsmooth part here is separable: $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$.
Minimizing over β_j, given $\beta_k, k \neq j$:

$$0 = X_j^T X_j \beta_j + X_j^T \left(X_{-j}\beta_{-j} - y\right) + \lambda s_j$$

where $s_j \in \partial |\beta_j|$ is the (sub-)gradient mentioned above. Solution is simply given by soft-thresholding

$$\beta_j = S_{\lambda/\|X_j\|_2^2}\left(\frac{X_j^T\left(y - X_{-j}\beta_{-j}\right)}{X_j^T X_j}\right)$$

Repeat this for $j = 1, 2, \ldots, p$.

Augmented Lagrangian method

Consider the problem

$$\min_{x} f(x) \text{ subject to } Ax = b$$

where f is strictly convex and closed. Denote Lagrangian

$$L(x, u) = f(x) + u^T(Ax - b)$$

Dual gradient ascent repeats, for $k = 1, 2, 3, \ldots$

$$x^{(k)} = \underset{x}{\operatorname{argmin}}\, L\left(x, u^{(k-1)}\right)$$
$$u^{(k)} = u^{(k-1)} + t_k\left(Ax^{(k)} - b\right)$$

Good: x can update separably when f does.
Bad: require stringent assumptions (strong convexity of f) to ensure convergence.

Augmented Lagrangian method

Augmented Lagrangian method modifies the problem, for $\rho > 0$,

$$\begin{array}{ll} \min_x & f(x) + \frac{\rho}{2}\|Ax - b\|_2^2 \\ \text{subject to} & Ax = b \end{array}$$

uses a modified Lagrangian

$$L_\rho(x, u) = f(x) + u^T(Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2$$

and repeats, for $k = 1, 2, 3, \ldots$

$$\begin{aligned} x^{(k)} &= \underset{x}{\operatorname{argmin}}\ L_\rho\left(x, u^{(k-1)}\right), \\ u^{(k)} &= u^{(k-1)} + \rho\left(Ax^{(k)} - b\right) \end{aligned}$$

Advantage: better convergence properties. Disadvantages: lose decomposability.

ADMM

Alternating direction method of multipliers or ADMM tries for the best of both methods. Consider a problem of the form:

$$\min_{x,z} f(x) + g(z) \text{ subject to } Ax + Bz = c$$

We define augmented Lagrangian, for a parameter $\rho > 0$,

$$L_\rho(x, z, u) = f(x) + g(z) + u^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

We repeat, for $k = 1, 2, 3, \ldots$

$$\begin{aligned}
x^{(k)} &= \operatorname*{argmin}_x L_\rho\left(x, z^{(k-1)}, u^{(k-1)}\right), \\
z^{(k)} &= \operatorname*{argmin}_z L_\rho\left(x^{(k)}, z, u^{(k-1)}\right), \\
u^{(k)} &= u^{(k-1)} + \rho\left(Ax^{(k)} + Bz^{(k)} - c\right).
\end{aligned}$$

ADMM for Lasso

Given $\mathrm{y} \in \mathbb{R}^{\mathrm{n}}, \mathrm{X} \in \mathbb{R}^{\mathrm{n}\times\mathrm{p}}$, we can rewrite the lasso problem:[15]

$$\min_{\beta,\alpha} \frac{1}{2}\|\mathrm{y} - \mathrm{X}\beta\|_2^2 + \lambda\|\alpha\|_1 \text{ subject to } \beta - \alpha = 0,$$

which objective function of ADMM is

$$\mathrm{L}_\rho(\mathrm{x}, \mathrm{z}, \mathrm{u}) = \frac{1}{2}\|\mathrm{y} - \mathrm{X}\beta\|_2^2 + \mathrm{u}^{\mathrm{T}}(\beta - \alpha) + \frac{\rho}{2}\|\beta - \alpha\|_2^2 + \lambda\|\alpha\|_1.$$

ADMM steps:

$$\begin{aligned}
\beta^{(\mathrm{k})} &= \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \rho\mathrm{I}\right)^{-1}\left(\mathrm{X}^{\mathrm{T}}\mathrm{y} + \rho\alpha^{(\mathrm{k}-1)} - \mathrm{u}^{(\mathrm{k}-1)}\right), \\
\alpha^{(\mathrm{k})} &= \mathrm{S}_{\lambda/\rho}\left(\beta^{(\mathrm{k})} + \alpha^{(\mathrm{k}-1)}/\rho\right), \\
\mathrm{u}^{(\mathrm{k})} &= \mathrm{u}^{(\mathrm{k}-1)} + \rho\left(\beta^{(\mathrm{k})} - \alpha^{(\mathrm{k})}\right).
\end{aligned}$$

Proximal Gradient Methods

If f were differentiable, then gradient descent update would be:

$$x^+ = x - t \cdot \nabla f(x),$$

which is equivalent to minimizing the quadratic approximation to f around x, replace $\nabla^2 f(x)$ by $\frac{1}{t}I$

$$x^+ = \underset{z}{\operatorname{argmin}} \; \underbrace{f(x) + \nabla f(x)^T (z - x) + \frac{1}{2t}\|z - x\|_2^2}_{\bar{f}_t(z)}$$

Proximal Gradient Methods

Suppose

$$f(x) = g(x) + h(x).$$

We borrow the idea mentioned above:

- g is convex, differentiable, $\mathrm{dom}(g) = \mathbb{R}^n$.
- h is convex, not necessarily differentiable.

$$\begin{aligned} x^+ &= \operatorname*{argmin}_z \bar{g}_t(z) + h(z) \\ &= \operatorname*{argmin}_z g(x) + \nabla g(x)^T (z - x) + \frac{1}{2t} \|z - x\|_2^2 + h(z) \\ &= \operatorname*{argmin}_z \frac{1}{2t} \|z - (x - t\nabla g(x))\|_2^2 + h(z). \end{aligned}$$

- $\frac{1}{2t}\|z - (x - t\nabla g(x))\|_2^2$: force z to be close to gradient update for g.
- $h(z)$: simultaneously, z would be penalized by h.
- It can be viewed as one type of Majorization-Minimization (MM) algorithm under certain conditions.

Majorization-Minimization (MM) algorithm

- A function $\Psi : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}^1$ majorizes the function f at a point $\beta \in \mathbb{R}^p$ if

$$f(\beta) \leq \Psi(\beta, \theta) \quad \text{for all } \theta \in \mathbb{R}^p$$

with equality holding when $\beta = \theta$.[16]

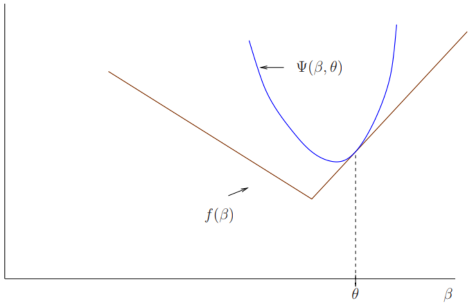


Figure: Illustration of a majorizing function for use in an MM algorithm.

Proximal Gradient Methods

Define proximal mapping:

$$\mathrm{prox}_h(x) = \underset{z}{\mathrm{argmin}}\ \|x - z\|_2^2 + h(z).$$

Noting that $\mathrm{prox}_{th}(x) = \underset{z}{\mathrm{argmin}}\ \frac{1}{2t}\|x - z\|_2^2 + h(z)$.

Proximal gradient descent: choose initialize $x^{(0)}$, repeat:

$$x^{(k)} = \mathrm{prox}_{t_k h}\left(x^{(k-1)} - t_k \nabla g\left(x^{(k-1)}\right)\right), \quad k = 1, 2, 3, \ldots.$$

Proximal Gradient Methods

To make this update step look familiar, we can rewrite it as

$$x^{(k)} = x^{(k-1)} - t_k \cdot G_{t_k}\left(x^{(k-1)}\right),$$

where G_t is the generalized gradient of f as defined:

$$G_t(x) = \frac{x - \text{prox}_{th}(x - t\nabla g(x))}{t}.$$

Goodness

You have a right to be suspicious ... may look like we just swapped one minimization problem for another.
Key point is that $\text{prox}_{th}(\cdot)$ has many efficient algorithms for many important functions h. Note:

- Mapping $\text{prox}_{th}(\cdot)$ doesn't depend on g at all, only on h.
- Smooth part g can be complicated, we only need to compute its gradients. Therefore, it can be easily to generalize to the GLM case.

Convergence analysis: will be in terms of the number of iterations, and each iteration evaluates $\text{prox}_{th}(\cdot)$ once (this can be cheap or expensive, depending on h).

PGD for Lasso

Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$, recall the lasso criterion:

$$f(\beta) = \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{g(\beta)} + \underbrace{\lambda\|\beta\|_1}_{h(\beta)}$$

Recall $\nabla g(\beta) = -X^T(y - X\beta)$, hence the pre-update β is:

$$\beta^+ = \beta + tX^T(y - X\beta).$$

Proximal mapping is now

$$\text{prox}_{th}(\beta^+) = \underset{z}{\text{argmin}} \frac{1}{2t}\|\beta^+ - z\|_2^2 + \lambda\|z\|_1,$$

which could be efficiently solved in parallel for each dimension by using the soft-thresholding operator.

Least Angle regression

Least Angle Regression (LAR) Algorithm[17]

1. Start with $r = y, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_p = 0$. Assume x_j standardized.
2. Find predictor x_j most correlated with r.
3. Increase β_j in the direction of $\mathrm{sign}(\mathrm{corr}(r, x_j))$ until other competitor x_k has as much correlation with current residual as does x_j.
4. Move $(\hat{\beta}_j, \hat{\beta}_k)$ in the joint least squares direction for (x_j, x_k) until some other competitor x_ℓ has as much correlation with the current residual.
5. Continue in this way until all predictors have been entered. Stop when $\mathrm{corr}(r, x_j) = 0, \forall j$, i.e. OLS solution.

LARS

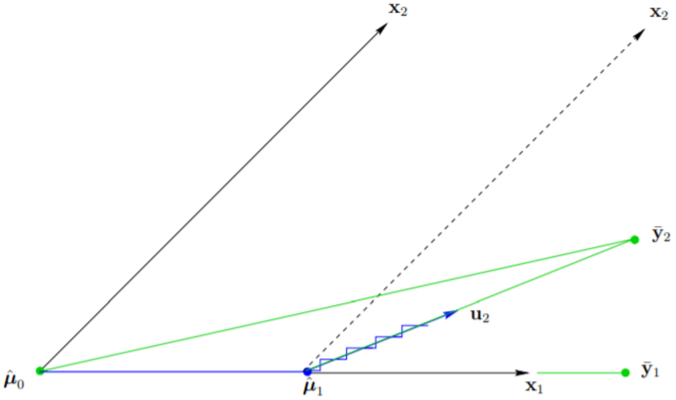


Figure: The LARS algorithm in the case of p = 2 covariates

LARS

For LARS updating, we should choose equiangular direction and step size.

1. Equiangular direction
   - Let $\mathcal{A}$ be active set, $\mathrm{X}_{\mathcal{A}}$ be the predictors in active set.
   - Equiangular vector $\mathrm{u}_{\mathcal{A}}$ should satisfy $\mathrm{X}_{\mathcal{A}}^{'}\mathrm{u}_{\mathcal{A}} = \mathrm{w}1_{\mathcal{A}}$, where w is a unit parameter.
   - A choice of $\mathrm{u}_{\mathcal{A}}$ is $\mathrm{u}_{\mathcal{A}} = \mathrm{w}\mathrm{X}_{\mathcal{A}}(\mathrm{X}_{\mathcal{A}}^{'}\mathrm{X}_{\mathcal{A}})^{-1}1_{\mathcal{A}}$.
2. Step size
   - Let $\widehat{\boldsymbol{\mu}}_{\mathcal{A}}$ be the current LARS estimate, $\widehat{\mathrm{c}} = \mathrm{X}'\left(\mathrm{y} - \widehat{\boldsymbol{\mu}}_{\mathcal{A}}\right)$ is the vector of current correlations. The active set $\mathcal{A}$ satisfy

$$\mathcal{A} = \left\{\mathrm{j} : |\widehat{\mathrm{c}}_{\mathrm{j}}| = \widehat{\mathrm{C}}\right\},$$

     where $\widehat{\mathrm{C}} = \max_{\mathrm{j}}\left\{|\widehat{\mathrm{c}}_{\mathrm{j}}|\right\}$. let $\mathrm{a} \equiv \mathrm{X}'\mathrm{u}_{\mathcal{A}}$.
   - Then, step size is

$$\widehat{\gamma} = \min_{\mathrm{j}\in\mathcal{A}^c}\left\{\frac{\widehat{\mathrm{C}} - \widehat{\mathrm{c}}_{\mathrm{j}}}{\mathrm{w} - \mathrm{a}_{\mathrm{j}}}, \frac{\widehat{\mathrm{C}} + \widehat{\mathrm{c}}_{\mathrm{j}}}{\mathrm{w} + \mathrm{a}_{\mathrm{j}}}\right\}.$$

1. Set $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$ and $k = 0$.
2. **repeat**
3. Calculate $\hat{\mathbf{c}} = X'(\mathbf{y} - \hat{\boldsymbol{\mu}}_k)$ and set $\hat{C} = \max_j\{|\hat{c}_j|\}$.
4. Let $\mathcal{A} = \{j : |\hat{c}_j| = \hat{C}\}$.
5. Set $X_{\mathcal{A}} = (\cdots \mathbf{x}_j \cdots)_{j \in \mathcal{A}}$ for calculating $\bar{\mathbf{y}}_{k+1} = (X'_{\mathcal{A}} X_{\mathcal{A}})^{-1} X'_{\mathcal{A}} \mathbf{y}$ and $\mathbf{a} = X'_{\mathcal{A}}(\bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k)$.
6. Set

$$\hat{\boldsymbol{\mu}}_{k+1} = \hat{\boldsymbol{\mu}}_k + \hat{\gamma}(\bar{\mathbf{y}}_{k+1} - \hat{\boldsymbol{\mu}}_k),$$

   where, if $\mathcal{A}^c \neq \varnothing$,

$$\hat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{\hat{C} - a_j}, \frac{\hat{C} + \hat{c}_j}{\hat{C} + a_j} \right\},$$

   otherwise set $\hat{\gamma} = 1$.
7. $k \leftarrow k + 1$.
8. **until** $\mathcal{A}^c = \varnothing$.

Figure: LARS Algorithm

Relation between LARS and Lasso

- Lasso problem can be written as :

$$\min_{\beta_j^+,\beta_j^-} \sum_{i=1}^n \left(y_i - \left[\sum_{j=1}^p x_{ij}\beta_j^+ - \sum_{j=1}^p x_{ij}\beta_j^-\right]\right)^2$$
$$\text{st. } \beta_j^+ \geq 0, \beta_j^- \geq 0 \ \forall j \text{ and } \sum_{j=1}^p \beta_j^+ + \beta_j^- \leq s.$$

- The Lagrangian is
$\sum_{i=1}^n \left(y_i - \left[\sum_{j=1}^p x_{ij}\beta_j^+ - \sum_{j=1}^p x_{ij}\beta_j^-\right]\right)^2 + \lambda \sum_{j=1}^p \left(\beta_j^+ + \beta_j^-\right) - \sum_{j=1}^p \lambda_j^+ \beta_j^+ - \sum_{j=1}^p \lambda_j^- \beta_j^-$.
- KKT Conditions:

$$\begin{aligned} -x_j^T r + \lambda - \lambda_j^+ &= 0 \\ x_j^T r + \lambda - \lambda_j^- &= 0 \\ \lambda_j^+ \beta_j^+ &= 0 \\ \lambda_j^- \beta_j^- &= 0 \end{aligned}$$

Lasso Path

Lasso Path

- Lasso path is given by $\boldsymbol{\beta}(\lambda)$, where $\boldsymbol{\beta}(\lambda)$ satisfies the KKT conditions.
- $\beta(\lambda_0)$ and $\beta(\lambda_1)$ are two closest points on the lasso path for the same active set $\mathcal{A}$, i.e. $\lambda_1 - \lambda_0 = \delta$, where δ is the smallest number.

We are going to show $\boldsymbol{\beta}(\lambda_1) - \boldsymbol{\beta}(\lambda_0)$ lies on the direction

$$\left(\mathrm{X}_{\mathcal{A}}^{\mathrm{T}} \mathrm{X}_{\mathcal{A}}\right)^{-1} \mathrm{X}_{\mathcal{A}}^{\mathrm{T}} \mathrm{r},$$

where $\mathrm{r} = \mathrm{y} - \mathrm{X}_{\mathcal{A}} \boldsymbol{\beta}(\lambda_0)$.

Lasso Path

- Define $\boldsymbol{\beta}_{\mathcal{A}}(\lambda)$ to be the corresponding coefficients at λ, where $\lambda \in [\lambda_0, \lambda_1]$. Deduction KKT conditions

$$\begin{array}{rl} \Longrightarrow & \mathrm{X}_{\mathcal{A}}^{\mathrm{T}}\left(\mathrm{y}-\mathrm{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}(\lambda)\right)=\lambda 1 \\ \Longrightarrow & \mathrm{X}_{\mathcal{A}}^{\mathrm{T}}\mathrm{X}_{\mathcal{A}}\left(\boldsymbol{\beta}_{\mathcal{A}}\left(\lambda_1\right)-\boldsymbol{\beta}_{\mathcal{A}}\left(\lambda_0\right)\right)=\delta 1 \\ \Longleftrightarrow & \boldsymbol{\beta}_{\mathcal{A}}\left(\lambda_1\right)-\boldsymbol{\beta}_{\mathcal{A}}\left(\lambda_0\right)=\delta\left(\mathrm{X}_{\mathcal{A}}^{\mathrm{T}}\mathrm{X}_{\mathcal{A}}\right)^{-1} 1 \end{array}.$$

- According to the KKT conditions, $\mathrm{X}_{\mathcal{A}}^{\mathrm{T}}\mathrm{r}=\lambda_0 1$:

$$\beta_{\mathcal{A}}\left(\lambda_1\right)-\beta_{\mathcal{A}}\left(\lambda_0\right)=\frac{\delta}{\lambda_0}\left(\mathrm{X}_{\mathcal{A}}^{\mathrm{T}}\mathrm{X}_{\mathcal{A}}\right)^{-1}\mathrm{X}_{\mathcal{A}}^{\mathrm{T}}\mathrm{r}.$$

LARS vs Lasso

Lasso can be thought of as restricted versions of LAR.[18]

- KKT: If $\beta_j^+ > 0$, then $x_j^T r = \lambda$ or if $\beta_j^- > 0$, then $-x_j^T r = \lambda$. (Lasso has this constrain while LARS does not.)
- LARS - uses least squares directions in the active set of variables.
- Lasso - uses least square directions; if a variable crosses zero, it is removed from the active set.

LARS: Lasso Modification

If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.

Summary

We introduce some optimization algorithms and use them to solve Lasso problem.

- Coordinate Descent.
- ADMM.
- Proximal Gradient Methods.
- LARS.

References I

- W. J. Fu, “Penalized regressions: the bridge versus the lasso,” Journal of computational and graphical statistics, vol. 7, no. 3, pp. 397–416, 1998.
- J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” Journal of the American statistical Association, vol. 96, no. 456, pp. 1348–1360, 2001.
- A. Antoniadis and J. Fan, “Regularization of wavelet approximations,” Journal of the American Statistical Association, vol. 96, no. 455, pp. 939–967, 2001.
- C.-H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” The Annals of statistics, vol. 38, no. 2, pp. 894–942, 2010.

J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," Statistica Sinica, vol. 20, no. 1, p. 101, 2010.

R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.

H. Zou, "The adaptive lasso and its oracle properties," Journal of the American statistical association, vol. 101, no. 476, pp. 1418–1429, 2006.

M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès, "Slope—adaptive variable selection via convex optimization," The annals of applied statistics, vol. 9, no. 3, p. 1103, 2015.

References III

H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the royal statistical society: series B (statistical methodology), vol. 67, no. 2, pp. 301–320, 2005.

M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 68, no. 1, pp. 49–67, 2006.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," Journal of computational and graphical statistics, vol. 22, no. 2, pp. 231–245, 2013.

L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in Proceedings of the 26th annual international conference on machine learning, pp. 433–440, 2009.

References IV

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 1, pp. 91–108, 2005.

K. P. Murphy, Machine learning: a probabilistic perspective. MIT press, 2012.

J. Fan, R. Li, C.-H. Zhang, and H. Zou, Statistical foundations of data science.
Chapman and Hall/CRC, 2020.

T. Hastie, R. Tibshirani, and M. Wainwright, Statistical learning with sparsity: the lasso and generalizations.
Chapman and Hall/CRC, 2019.

References V

- B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," The Annals of statistics, vol. 32, no. 2, pp. 407–499, 2004.
- J. H. Friedman, The elements of statistical learning: Data mining, inference, and prediction.
  springer open, 2017.