Extensions to High Dimensions

27th February 2023

What happens when $p > n$?

- Recall the multiple linear regression model

$$Y_i = \beta_0 + \beta_1^T X_i + \epsilon_i, i = 1, \cdots, N.$$

- Let $\mathbf{y} = (Y_1, \cdots, Y_N)$ and $\mathbf{X} \in \mathbb{R}^{N \times p}$ be the sample matrix.
- $\hat{\beta}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
- The matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ is not invertible when $p > n$. (Why?)
- It is difficult to interpret a p-dimensional coefficient β.
- A popular assumption for solving the high-dimensional issues is the sparsity assumption.
- Only some elements of β is non-zero. Namely, $|\{j : \beta_j \neq 0\}|_0 = s \ll p$.

What happens when $p > n$?

- Recall the multiple linear regression model

$$Y_i = \beta_0 + \beta_1^T X_i + \epsilon_i, i = 1, \cdots, N.$$

- Let $\mathbf{y} = (Y_1, \cdots, Y_N)$ and $\mathbf{X} \in \mathbb{R}^{N \times p}$ be the sample matrix.
- $\hat{\beta}^{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.
- The matrix $(\mathbf{X}^T \mathbf{X})^{-1}$ is not invertible when $p > n$. (Why?)
- It is difficult to interpret a p-dimensional coefficient β.
- A popular assumption for solving the high-dimensional issues is the sparsity assumption.
- Only some elements of β are non-zero. Namely, $|\{j : \beta_j \neq 0\}|_0 = s \ll p$.

Possible Solutions

- Subset selection i.e. select a small number of predictors from p predictors.
- Increase the sample size n.
- Pseudo inverse for $(\mathbf{X}^T\mathbf{X})^{-1}$.
- Shrinkage: ridge, lasso, elastic net, group lasso...
- Other methods

Ridge Regression ***

- The ridge regression is defined as

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 / N + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

- Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
- The larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero.
- When $\lambda = 0$, the ridge estimator reduces to the OLS.
- Can the elements of $\hat{\beta}^{\text{ridge}}$ be exactly zero?

Ridge Regression

- An equivalent way to write the ridge problem is

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

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,$$

- There is a one to-one correspondence between the parameters λ and t.
- When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance.

Ridge Regression

- A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin.
- Ridge regularization makes this problem alleviated.
- The ridge solutions are not equivalent under scaling of the inputs, and so one normally standardizes the inputs.

Solution for ridge regression

- Ridge regression has a closed form solution!
- Writing the objective function in matrix form,

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta,$$

  the ridge regression solutions are easily seen to be

$$\boxed{\hat{\beta}^{\text{ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}}$$

  .

- Ridge estimator is biased.
- The larger the λ, the larger the bias.

Can Ridge do variable selection?

Solution for ridge regression

- Ridge can handle the case where $p > n$ numerically.
- λ can be chosen by Information Criterion or cross-validation.
- Can the ridge estimator be consistent?
- How large is the bias?
- Is there any theoretical choice for λ?
- Unfortunately, the current theoretical results requires a **sample complexity** $n > Cp$ for a large enough constant $C > 1$.
- Under some technical conditions, $\|\hat{\beta}^{\text{ridge}} - \beta\|_2 = O(p/n)$.

Example: Prostate Cancer

- Response: prostate-specific antigen.
- The predictors are log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

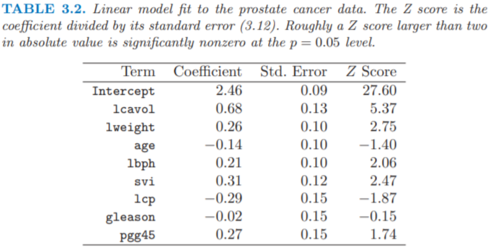
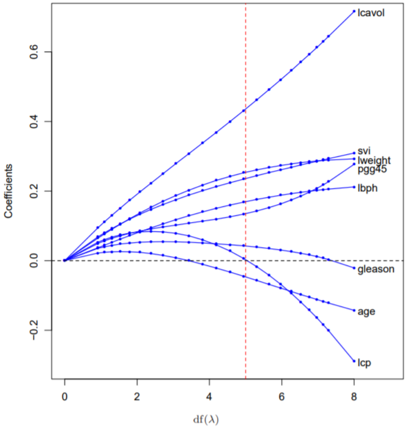


TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the* $p = 0.05$ *level.*

Term	Coefficient	Std. Error	Z Score
`Intercept`	2.46	0.09	27.60
`lcavol`	0.68	0.13	5.37
`lweight`	0.26	0.10	2.75
`age`	−0.14	0.10	−1.40
`lbph`	0.21	0.10	2.06
`svi`	0.31	0.12	2.47
`lcp`	−0.29	0.15	−1.87
`gleason`	−0.02	0.15	−0.15
`pgg45`	0.27	0.15	1.74

Example: Prostate Cancer (Con't)



Note: df(λ) decreases with λ.

Discussion for ridge

- Can handle high-dimensional case numerically.
- Easy to obtain
- Biased estimator
- Cannot select variables

LASSO * **

- Short for least absolute shrinkage and selection operator.
- LASSO estimator is defined as

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- Or equivalently

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

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

LASSO Explanation

- Consider an orthonormal input matrix $\mathbf{X}$, namely $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$.

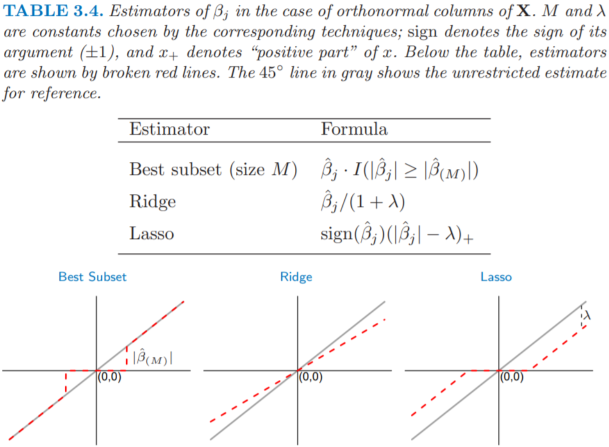


TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of $\mathbf{X}$. M and λ are constants chosen by the corresponding techniques;* sign *denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

LASSO Explanation

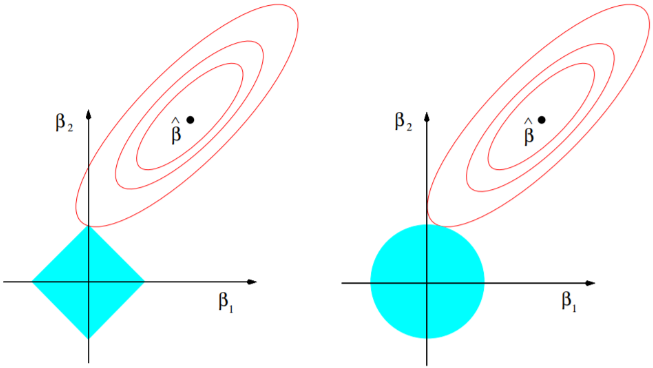


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

LASSO Estimation

- Unlike the ridge regression, the lasso objective function does not have a closed-form solution.
- Case becomes more challenging since the lasso penalty (ℓ_1 penalty) $\sum_{j=1}^{p} |\beta_j|$ is not always differentiable.
- We introduce two popular algorithms: Least Angle Regression (LAR) and coordinate descent algorithm.

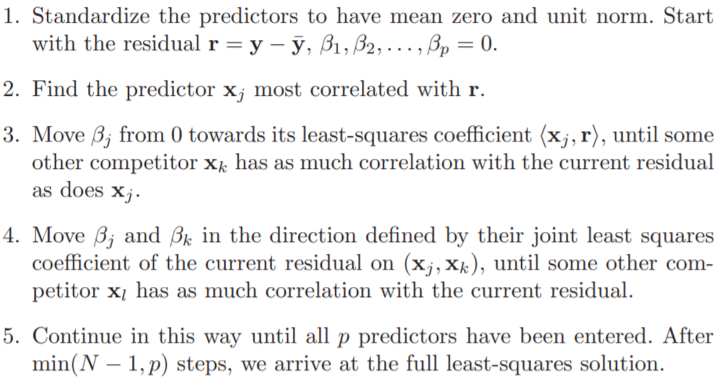
Least Angle Regression

- Forward stepwise regression builds a model sequentially, adding one variable at a time. At each step, it identifies the best variable to include in the active set, and then updates the least squares fit to include all the active variables.
- Least angle regression uses a similar strategy, but only enters "as much" of a predictor as it deserves.
- At the first step it identifies the variable most correlated with the response. Rather than fit this variable completely, LAR moves the coefficient of this variable continuously toward its least squares value (causing its correlation with the evolving residual to decrease in absolute value).

Least Angle Regression

- As soon as another variable “catches up” in terms of correlation with the residual, the process is paused.
- The second variable then joins the active set, and their coefficients are moved together in a way that keeps their correlations tied and decreasing.
- This process is continued until all the variables are in the model, and ends at the full least-squares fit.

Least Angle Regression



1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.

2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.

3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.

4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.

5. Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full least-squares solution.

Least Angle Regression

Suppose $\mathcal{A}_k$ is the active set of variables at the beginning of the k th step, and let $\beta_{\mathcal{A}_k}$ be the coefficient vector for these variables at this step; there will be $k-1$ nonzero values, and the one just entered will be zero. If $\mathbf{r}_k = \mathbf{y} - \mathbf{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$ is the current residual, then the direction for this step is

$$\delta_k = \left(\mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k}\right)^{-1} \mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k.$$

The coefficient profile then evolves as $\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha \cdot \delta_k$.

Least Angle Regression

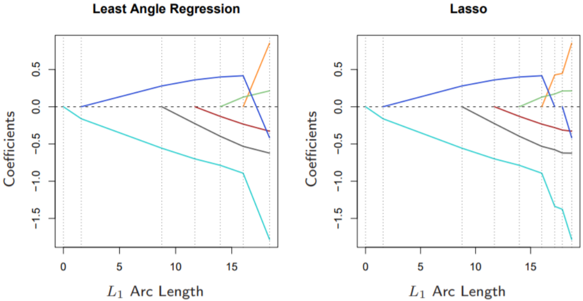


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about* 18.

Least Angle Regression

- The right panel of Figure 3.15 shows the lasso coefficient profiles on the same data. They are almost identical to those in the left panel, and differ for the first time when the blue coefficient passes back through zero.
- LAR (lasso) If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.
- The solution given by LAR is very similar to lasso. However it may not be the global solution for the lasso objective function

Coordinate Descent Algorithm **

- Suppose the predictors are all standardized to have mean zero and unit norm. Denote by $\tilde{\beta}_k(\lambda)$ the current estimate for β_k at penalty parameter λ. We can rearrange (3.52) to isolate β_j,

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

- This can be viewed as a univariate lasso problem with response variable the partial residual $y_i - \tilde{y}_i^{(j)} = y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k(\lambda)$.
- Explicit solution,

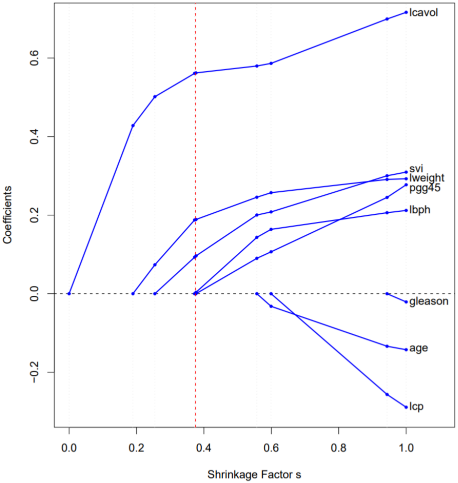
$$\tilde{\beta}_j(\lambda) \leftarrow S\left(\sum_{i=1}^{N} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right), \lambda\right).$$

- $S(t, \lambda) = \operatorname{sign}(t)(|t| - \lambda)_+$is the soft-thresholding operator.

Coordinate Descent Algorithm

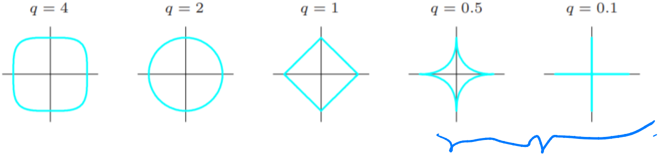
- We can also use this simple algorithm to efficiently compute the lasso solutions at a grid of values of λ.
- We start with the smallest value $\lambda_{\max}$ for which $\hat{\beta}\left(\lambda_{\max}\right) = 0$
- Then decrease it a little and cycle through the variables until convergence.
- Then λ is decreased again and the process is repeated, using the previous solution as a "warm start" for the new value of λ.
- This can be faster than the LARS algorithm, especially in large p, small s problems.

Return to Prostate Cancer



Generalization of LASSO and Ridge

- Recall that the lasso penalty is $\lambda \sum_{j=1}^{p} |\beta_j|$, the ridge penalty is $\lambda \sum_{j=1}^{p} \beta_j^2$. They are corresponding to the ℓ_1 and ℓ_2 norms of β, respectively
- This motivates us to consider $\lambda \sum_{j=1}^{p} |\beta_j|^q$



- when $q > 1$, we do not have variable selection property.

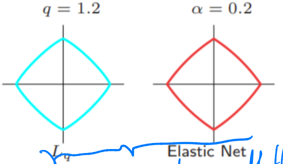
why not use thres two penalties.

Generalization of LASSO and Ridge

- Another direct generalization is elastic net

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha) |\beta_j| \right)$$

- Weighted sum of lasso and ridge



what is the difference between them?

- Elastic net can select variables!

Some theoretical results for LASSO

- Note that LASSO aims to solve the high-dimensional problems. The dimension p can be much greater than n.
- Convergence you may have learned: Converge in Probability; Almost everywhere convergence; Converge in distribution...
- They have a prerequisite $n \to \infty$ and p is a fixed constant.
- Let us consider a simple case. Suppose that $X_i \sim N(\mu, I_p)$, for $i = 1, \cdots, n$ independently. For $j = 1, \cdots, p$, we have $Pr(|\bar{X}_j - \mu| > \epsilon) \leq c_1 \exp(-c_2 n \epsilon^2)$ (why?). Hence, we have $|\bar{X}_j - \mu| \leq M\sqrt{1/n}$ with probability at least $1 - c_1 \exp(-c_2 M)$.
- We usually say $|\bar{X}_j - \mu| = O(1/\sqrt{n})$, with high probability.
- Now we consider the convergence of $\bar{X}$ in terms of the ℓ_2 norm. Since p may also go to ∞, the traditional convergence is usually not meaningful.

Some theoretical results for LASSO (Preparation)

- $\bar{X} - \mu \sim N(0, I_p/n)$. As such, $n\|\bar{X} - \mu\|_2^2 \sim \chi_p^2$.
- (Probability Tail Bound for Non-central χ^2 Distribution) If W satisfies the non-central Chi-square distribution $\chi_m^2(\lambda)$ with degrees of freedom m and non-centrality parameter λ, so that $W = \sum_{i=1}^m X_i^2$, where $X_i \sim N\left(\mu_i, 1\right)$ with $\sum_i \mu_i^2 = \lambda$. Then for any $x > 0$,

$$P\{X \geq m + \lambda + 2\sqrt{(m + 2\lambda)x} + 2x\} \leq e^{-x},$$
$$P\{X \leq m + \lambda - 2\sqrt{(m + 2\lambda)x}\} \leq e^{-x}.$$

- Using this tail bound, we can show that $\|\bar{X} - \mu\|_2 = O(\sqrt{p/n})$ with high probability.

Some theoretical results for LASSO

- Suppose $\epsilon_i \sim N(0, \sigma^2)$ independently, $|\beta|_0 \leq s$. The design matrix $\mathbf{X}$ satisfies restricted eigenvalue (RE) condition.
- **Definition (Restricted Eigenvalue (RE) condition)** For some $\alpha \geq 1$, and subset $\mathcal{S} \subseteq \{1, \ldots, d\}, \mathcal{S} \neq \emptyset$, let

$$\mathcal{C}_\alpha(\mathcal{S}) = \left\{\Delta \in \mathbb{R}^p : \left\|\Delta_{\mathcal{S}^c}\right\|_1 \leq \alpha \left\|\Delta_{\mathcal{S}}\right\|_1\right\},$$

where $\mathcal{S}^{\mathcal{C}} = \{1, \ldots, p\} \backslash \mathcal{S}$ and $\Delta_{\mathcal{S}} = \{\Delta_j, j \in \mathcal{S}\}$ We say that an $n \times p$ matrix $\mathbf{X}$ satisfies the $RE(\alpha, \kappa)$ condition w.r.t. $\mathcal{S}$ if

$$\frac{1}{n}\|\mathbf{X}\Delta\|^2 \geq \kappa\|\Delta\|^2 \quad \forall \Delta \in \mathcal{C}_\alpha(\mathcal{S})$$

where $\kappa > 0$.

Some theoretical results for LASSO

- If $\lambda/n = C\sqrt{\log p/n}$ for a large enough constant C, then

  $$\|\hat{\beta}^{lasso} - \beta\| = O(\sigma\sqrt{\frac{s\log p}{n}}),$$

  with probability at least $1 - cp^{-1}$.
- If we know the true sparse set $\mathcal{S}$, then the convergence rate is $\sqrt{s/n}$.
- $\log p$ can be viewed as a sacrifice of variable selection.
- With a more careful choice of λ, we may show an upper bound $O(\sigma\sqrt{\frac{s\log(p/s)}{n}})$, which achieves the minimax lower bound.

Some theoretical results for LASSO

- Recall that LASSO can select variables.
- **Question**: Does LASSO have variable selection consistency results?
- Unfortunately, without additional assumptions on the design matrix $\mathbf{X}$, LASSO does not have variable selection consistency.
- Zou, Hui. "The adaptive lasso and its oracle properties." Journal of the American statistical association 101.476 (2006): 1418-1429.
- Zhao, Peng, and Bin Yu. "On model selection consistency of Lasso." The Journal of Machine Learning Research 7 (2006): 2541-2563.

Modifications to LASSO in order to achieve variable selection consistency

- Adaptive lasso: Suppose that $\hat{\beta}$ is a root-n consistent estimator for β. The adaptive lasso penalty is defined as $\lambda \sum_{j=1}^{p} \omega_j |\beta_j|$, where $\omega_j = 1/|\hat{\beta}_j|^\gamma$ for a constant $\gamma > 0$.
- SCAD penalty: the smoothly clipped absolute deviation (SCAD) penalty replaces $\lambda|\beta|$ by $J_a(\beta, \lambda)$, where

$$\frac{dJ_a(\beta,\lambda)}{d\beta} = \lambda \cdot \mathrm{sign}(\beta) \left[I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right]$$

.

- Log Sum penalty, Minimax Concave Penalty...

Other generalizations of LASSO penalty

- Fused LASSO
- Group LASSO
- Graphical LASSO
- ...

LASSO for logistic regression *

- The L_1 penalty used in the lasso (Section 3.4.2) can be used for variable selection and shrinkage with any linear regression model.
- For logistic regression, we would maximize a penalized:

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i \left(\beta_0 + \beta^T x_i \right) - \log \left(1 + e^{\beta_0 + \beta^T x_i} \right) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- As with the lasso, we typically do not penalize the intercept term, and standardize the predictors for the penalty to be meaningful.
- We can use the same quadratic approximations in the Newton algorithm to solve this objective function.

High-dimensional Reduced Rank Regression

- Recall that, to obtain the reduced rank estimator, we need to first obtain the OLS estimator.
- When $p \gg n$, the OLS estimator is ill-conditioned.
- Note that $\mathbf{B} = \mathbf{A}\mathbf{C}^T$, where $\mathbf{A} \in^{p \times r}$ and $\mathbf{C} \in^{q \times r}$.
- Like LASSO, if we assume only some rows of $\mathbf{A}$ are non-zero, then only a small number of predictors are associated with the response.
- Consider the following objective function

$$\min_{\mathbf{A},\mathbf{C}} \left\| \mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{C}^T \right\|^2 + \sum_{i=1}^{p} \lambda_i \left\| \mathbf{A}^i \right\| \quad \text{such that} \quad \mathbf{C}^T\mathbf{C} = \mathbf{I},$$

where the superscript denotes a row of the matrix so that $\mathbf{B}^i$ is a row vector.

High-dimensional Reduced Rank Regression

For fixed $\mathbf{A}$, the optimization problem reduces to

$$\min_{\mathbf{C}} \left\|\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{C}^T\right\| \quad \text{such that} \quad \mathbf{C}^T\mathbf{C} = \mathbf{I},$$

The solution is $\widehat{\mathbf{C}} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are obtained from singular value decomposition $\mathbf{Y}^T\mathbf{X}\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Now we consider optimization over $\mathbf{A}$ for fixed $\mathbf{C}$. Since $\mathbf{C}$ has orthonormal columns, there is a matrix $\mathbf{C}^\perp$ with orthonormal columns such that $(\mathbf{C}, \mathbf{C}^\perp)$ is an orthogonal matrix. Then we have

$$\begin{aligned} \left\|\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{C}^T\right\|^2 &= \left\|\left(\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{C}^T\right)\left(\mathbf{C}, \mathbf{C}^\perp\right)\right\|^2 \\ &= \left\|\mathbf{Y}\mathbf{C} - \mathbf{X}\mathbf{A}\right\|^2 + \left\|\mathbf{Y}\mathbf{C}^\perp\right\|^2. \end{aligned}$$

High-dimensional Reduced Rank Regression

The second term does not involve $\mathbf{A}$. Therefore for fixed A, the optimization problem (8) reduces to

$$\min_{\mathbf{A}} \|\mathbf{YC} - \mathbf{XA}\|^2 + \sum_{i=1}^{p} \lambda_i \left\|\mathbf{A}^i\right\|.$$

- This is a convex problem, which can be solved using similar algorithm as the coordinate descent.
- The algorithm update $\mathbf{A}$ and $\mathbf{C}$ cyclically.
- For more details, see Chen, L., Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. Journal of the American Statistical Association, 107(500), 1533-1545.

Sparse CCA

- In CCA, we need to obtain the $1/2$-inverse of the sample covariance matrices S_{XX} and S_{YY}.
- They are ill-conditioned when $p \gg n$.
- Is there any way that can avoid usage of the $1/2$-inverse matrices?
- We breifly introduce three methods:
  - PMD from "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis".
  - COLAR from "Sparse cca: Adaptive estimation and computational barriers".
  - SCCA from " An iterative penalized least squares approach to sparse canonical correlation analysis".

PMD

- Replace $\widehat{\Sigma}_{\mathbf{YY}}$ and $\widehat{\Sigma}_{\mathbf{XX}}$ with identity matrices to avoid the singularity in these matrices.
- PMD estimates $(\boldsymbol{\alpha}_1^*, \boldsymbol{\beta}_1^*)$ by

$$\left(\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}_1\right) = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}} \boldsymbol{\beta}, \text{ s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} \leqslant 1, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \leqslant 1,$$
$$P_{\mathbf{Y}}(\boldsymbol{\alpha}) \leqslant \tau_1, P_{\mathbf{X}}(\boldsymbol{\beta}) \leqslant \tau_2$$

- $P_{\mathbf{Y}}(\cdot), P_{\mathbf{X}}(\cdot)$ are sparsity-inducing penalty functions such as the ℓ_1 penalty.
- However, the sparse CCA directions from PMD may be inconsistent when $\boldsymbol{\Sigma}_{\mathbf{XX}}$ and $\boldsymbol{\Sigma}_{\mathbf{YY}}$ are far from diagonal.

COLAR

- Given the number of desired pairs $K \geqslant 1$, define $\mathbf{A}_K = \left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K\right), \mathbf{B}_K = \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\right)$ and $\mathbf{F}_K = \mathbf{B}_K \mathbf{A}_K^{\mathrm{T}} \in \mathbb{R}^{p\times q}$.
- Split the data into threebatches with equal sizes and compute $\widehat{\boldsymbol{\Sigma}}^{(j)}_{\mathbf{XX}}, \widehat{\Sigma}^{(j)}_{\mathbf{YY}}, \widehat{\boldsymbol{\Sigma}}^{(j)}_{\mathbf{XY}}$ as sample covariances of the j th batch.
- COLAR then carries out a two-stage analysis.
- First, it finds a sparse estimate of $\mathbf{F}_K$:

$$\widehat{\mathbf{F}}_K = \arg \max_{\mathbf{F}_K \in \mathbb{R}^{p\times q}} \left\{ \mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}^{(1)}_{\mathbf{XY}}, \mathbf{F}_K\right) - \lambda \left\|\mathbf{F}_K\right\|_1 \right\}, \text{ s.t. } \|\mathbf{N}\|_* \leqslant K, \|$$

where $\lambda > 0, \mathbf{N} = \left(\widehat{\boldsymbol{\Sigma}}^{(1)}_{\mathbf{XX}}\right)^{1/2} \mathbf{F}_K \left(\widehat{\boldsymbol{\Sigma}}^{(1)}_{\mathbf{YY}}\right)^{1/2}$, and $\|\mathbf{N}\|_*$ and $\|\mathbf{N}\|_{op}$ are the summation and the maximum of the singular values of $\mathbf{N}$.

- Second, one decomposes $\widehat{F}_K$ on the second batch of data into

COLAR (Con't)

- Second, decompose $\widehat{F}_K$ on the second batch of data into $\left\{\widehat{\mathbf{A}}_K, \widehat{\mathbf{B}}_K\right\}$, which are further rescaled on the third batch of data.
- It can be computationally demanding when $p,\ q$ are both large.

SCCA

(1) Given $\mathbf{A}_{k-1}, \mathbf{B}_{k-1}$, compute $\mathbf{\Omega}_k$; Initialize $\left\{\widehat{\boldsymbol{\alpha}}_k^{(0)}, \widehat{\boldsymbol{\beta}}_k^{(0)}\right\}$.

(2) Repeat the following two steps until convergence:

(a) Set $\tilde{\mathbf{Y}}_k^{(m)} = \mathbf{\Omega}_k^{\mathrm{T}} \mathbf{Y} \widehat{\boldsymbol{\alpha}}_k^{(m)}$. Compute

$$\breve{\boldsymbol{\beta}}_k^{(m)} = \arg\min_{\boldsymbol{\beta}_k} \left\{ \frac{1}{2n} \left\| \tilde{\mathbf{Y}}_k^{(m)} - \mathbf{X}\boldsymbol{\beta}_k \right\|_2^2 + \lambda_{\boldsymbol{\beta}_k} \left\| \boldsymbol{\beta}_k \right\|_1 \right\},$$

and then set $\widehat{\boldsymbol{\beta}}_k^{(m)} = \left[\left\{ \breve{\boldsymbol{\beta}}_k^{(m)} \right\}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \breve{\boldsymbol{\beta}}_k^{(m)} \right]^{-1/2} \cdot \breve{\boldsymbol{\beta}}_k^{(m)}$.

(b) Set $\tilde{\mathbf{X}}_k^{(m)} = \mathbf{\Omega}_k \mathbf{X} \widehat{\boldsymbol{\beta}}_k^{(m)}$. Compute

$$\breve{\boldsymbol{\alpha}}_k^{(m)} = \arg\min_{\boldsymbol{\alpha}_k} \left\{ \frac{1}{2n} \left\| \widetilde{\mathbf{X}}_k^{(m)} - \mathbf{Y}\boldsymbol{\alpha}_k \right\|_2^2 + \lambda_{\boldsymbol{\alpha}_k} \left\| \boldsymbol{\alpha}_k \right\|_1 \right\},$$

and then set $\widehat{\boldsymbol{\alpha}}_k^{(m)} = \left[\left\{ \breve{\boldsymbol{\alpha}}_k^{(m)} \right\}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{\mathbf{YY}} \breve{\boldsymbol{\alpha}}_k^{(m)} \right]^{-1/2} \cdot \breve{\boldsymbol{\alpha}}_k^{(m)}$.

Extension to penalized M-estimator

- After LASSO came out, there has been an explosive growth in sparse method.
  - Direct Truncation
  - Penalized M-estimator
- Negahban, Sahand N., et al. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers." (2012): 538-557.

M-estimator

- Let $Z_1^n := \{Z_1, \ldots, Z_n\}$ denote n identically distributed observations with marginal distribution $\mathbb{P}$.
- Let θ be the parameter of interest.
- $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \to \mathbb{R}$ be a convex and differentiable loss function that, for a given set of observations Z_1^n, assigns a cost $\mathcal{L}\left(\theta; Z_1^n\right)$ to any parameter $\theta \in \mathbb{R}^p$.
- $\theta^* \in \arg\min_{\theta \in \mathbb{R}^p} \overline{\mathcal{L}}(\theta)$ be any minimizer of the population risk $\overline{\mathcal{L}}(\theta) := \mathbb{E}_{Z_1^n}\left[\mathcal{L}\left(\theta; Z_1^n\right)\right]$.
- Examples: lease square estimator; least absolute estimator ;maximum likelihood estimator.

M-estimator

- Under some technical conditions: $\hat{\theta} = \mathrm{argmin}_{\theta}\mathcal{L}\left(\theta; Z_1^n\right)$ satisfies $\hat{\theta} \rightarrow_p \theta^*$.
- Furthermore, $\sqrt{n}(\hat{\theta} - \theta)$ follows normal distribution asymptotically.
- Examples: the OLS estimator for the simple linear regression converge to the true β in probability. The MLE is also consistent.
- Those asymptotic results have a limitation: p much be a fixed constant!
- As lasso, in high-dimensional problems, we consider the case where $p \gg n$.

Penalized M-estimator

- Consider the convex optimization problem (1) $\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \{\mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta)\}$
- $\lambda_n > 0$ is a user-defined regularization penalty and $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}_+$is a norm.
- For LASSO, $\mathcal{R}(\theta) = \sum_{j=1}^{p} |\theta_j|$.
- Goal: provide general techniques for deriving bounds (convergence rate in terms of n and p) on the difference between any solution $\widehat{\theta}_{\lambda_n}$ and the unknown vector θ^* in terms of the ℓ_2 norm.

Penalized M-estimator

- The origin paper considered a wide range of the penalties, defined a concept called “decomposable”, and used the dual norm concept frequently.
- To make it easier, we only consider the LASSO penalty, namely $\mathcal{R}(\theta) = \sum_{j=1}^{p} |\theta_j|$.
- The dual norm for $\mathcal{R}(\theta) = \sum_{j=1}^{p} |\theta_j|$ is $R^*(\theta) = \max_j |\theta_j|$.
- **Lemma**: Suppose that $\mathcal{L}$ is a convex and differentiable function, and consider any optimal solution $\widehat{\theta}$ to the optimization problem with a strictly positive regularization parameter satisfying $\lambda_n \geq 2\mathcal{R}^*\left(\nabla\mathcal{L}\left(\theta^*; Z_1^n\right)\right)$. Then the error $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ belongs to the set

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^p \mid \mathcal{R}\left(\Delta_{S^c}\right) \leq 3\mathcal{R}(\Delta_S)\}.$$

Penalized M-estimator

- S is the non-sparse set for θ and $S^c = \{1, \cdots .p\}/S$ is its complement.
- The set $\mathbb{C}(S)$ can be viewed as a "sparse" set.
- Although the coefficient of Δ corresponding to S^c may not be exactly zero. Its ℓ_1 norm is small.
- Intuitively, $\mathbb{C}(S)$ restricts Δ in a small subspace of $\mathbb{R}^p$. Hence, the penalized estimation is sparse.
- Theoretically, we can prove $\mathbb{C}(S)$ is covered by a sparse set, whose sparse level is higher than S.

Requirement for $\mathcal{L}$

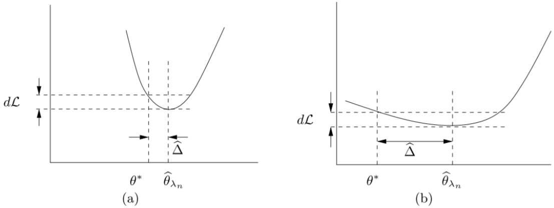


FIG. 2. *Role of curvature in distinguishing parameters.* (a) *Loss function has high curvature around* $\widehat{\Delta}$. *A small excess loss* $d\mathcal{L} = |\mathcal{L}(\widehat{\theta}_{\lambda_n}) - \mathcal{L}(\theta^*)|$ *guarantees that the parameter error* $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ *is also small.* (b) *A less desirable setting, in which the loss function has relatively low curvature around the optimum.*

The loss function satisfies a restricted strong convexity (RSC) condition with curvature $\kappa_{\mathcal{L}} > 0$

$$\delta\mathcal{L}\left(\Delta, \theta^*\right) \geq \kappa_{\mathcal{L}} \|\Delta\|^2 - \tau_{\mathcal{L}}^2\left(\theta^*\right) \text{ for all } \Delta \in \mathbb{C}(S).$$

with high probability.

Convergence results for penalized M-estimator

THEOREM Under RSC condition, any optimal solution $\widehat{\theta}_{\lambda_n}$ satisfies the bound

$$\left\|\widehat{\theta}_{\lambda_n} - \theta^*\right\|^2 \leq 9\frac{{\lambda_n}^2}{\kappa_{\mathcal{L}^2}}S$$

- As a result, under some mild conditions for the design matrix $\mathbf{X}$, the LASSO penalized logistic regression has convergence result $\sqrt{s \log p/n}$.
- This is a theoretical basis for many penalized estimators!