Penalized Least Squares with LASSO and SCAD

San-Teng Huang,Shang-Chien Ho,Hsing-Cheng Pan

National Dong Hwa University

2019/01/04

Outline

Multiple Linear Regression

- Consider multiple linear model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{ip-1} + \varepsilon_i \quad , i = 1, 2, \ldots, n \tag{1}$$

  where $\boldsymbol{\beta}$ is $p \times 1$ vector, $\mathbf{X}$ is $n \times p$ matrix,
  and ε_i *i.i.d* with mean 0, variance σ^2, for $i = 1, \ldots, n$.
- Assume only some parameters are non-zero.
  the true $\boldsymbol{\beta_0} = (\beta_{01}, \beta_{02}, \ldots, \beta_{0p}) = (\boldsymbol{\beta_{01}^T}, \boldsymbol{\beta_{02}^T}) = (\boldsymbol{\beta_{01}^T}, \mathbf{0})$

Goal

- Variable selection
  the true $\boldsymbol{\beta_0} = (\beta_{01}, \beta_{02}, ..., \beta_{0p}) = (\boldsymbol{\beta_{01}^T}, \boldsymbol{\beta_{02}^T}) = (\boldsymbol{\beta_{01}^T}, \mathbf{0})$

The LASSO : ℓ_1 penalty

- Tibshirani (1996) introduced the LASSO : least absolute shrinkage and selection operator.
- LASSO coefficients are the solutions to the ℓ_1 optimization problem :

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad s.t. \, ||\beta||_1 \le t \tag{2}$$

$$\Leftrightarrow \min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

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

λ(or t) as a tuning parameter

- The constraint :

$$\sum_{j=1}^{p} |\beta_j| \leq t \quad or \quad \lambda \sum_{j=1}^{p} |\beta_j|$$

  - If $\lambda = 0$, then it means no shrinkage. (hence is the OLS solutions.)
  - Large enough λ (or small enough t) will set some coefficients exactly equal to 0.

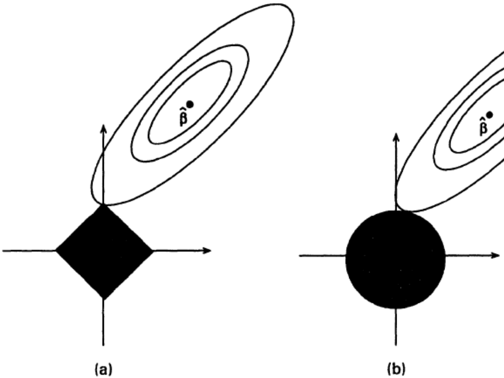
- Ridge regression's coefficients are the solutions to the ℓ_2 optimization problem :

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad s.t. \, ||\beta||_2 \leq t$$

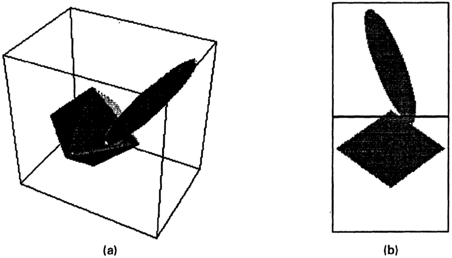
$$\Leftrightarrow \min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

Why the LASSO set coefficients exactly equal to 0 ?



Estimation picture for (a) the lasso and (b) ridge regression



(a) Example in which the lasso estimate falls in an octant different from the overall least squares estimate; (b) overhead view

Example

- Consider model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{10} X_{i10} + \varepsilon_i \quad , i = 1, 2, \ldots, 100$$

  where $\mathbf{X} \sim MVN(0, \Sigma)$, Σ be covariance matrix with
  for $j, k = 1, \ldots, 10$, $j \neq k$, $Var(X_j) = 1$, $Cov(X_j, X_k) = 0.5$
  and $\varepsilon_i \sim N(0, 3^2)$, $i = 1, \ldots, 100$.
  the true $\boldsymbol{\beta_0} = (3, 1.5, 2, -7, 15, 0, 0, 0, 0, 0)$
- With 100 replication, count the number of estimated value greater than 10^{-3}. The parameter λ choosed by 10-fold cross validation.

	OLS		LASSO($\lambda = 0.2636$)	
	bias(sd)	count(%)	bias(sd)	count(%)
$\beta_1 = 3$	0.0400(0.44)	1	-0.1529(0.42)	1
$\beta_2 = 1.5$	0.0307(0.45)	1	-0.1655(0.42)	1
$\beta_3 = 2$	0.0474(0.43)	1	-0.1435(0.42)	1
$\beta_4 = 7$	-0.0495(0.40)	1	0.3294(0.39)	1
$\beta_5 = 15$	0.0623(0.47)	1	-0.1395(0.45)	1
$\beta_6 = 0$	-0.0102(0.45)	1	0.0272(0.25)	0.59
$\beta_7 = 0$	0.0028(0.39)	0.99	0.0547(0.22)	0.54
$\beta_8 = 0$	-0.0636(0.43)	1	0.0215(0.25)	0.55
$\beta_9 = 0$	0.0180(0.44)	1	0.0744(0.25)	0.61
$\beta_{10} = 0$	-0.0678(0.39)	1	0.0045(0.21)	0.58

Penalized Least Squares

- To minimize

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + n\sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{4}$$

where $p_\lambda(|\beta|)$ is a penalty function.

- The penalized least squares estimator is

$$\hat{\beta} = arg\min_{\beta} Q(\boldsymbol{\beta})$$

Penalty function

- LASSO(L_1-penalty):

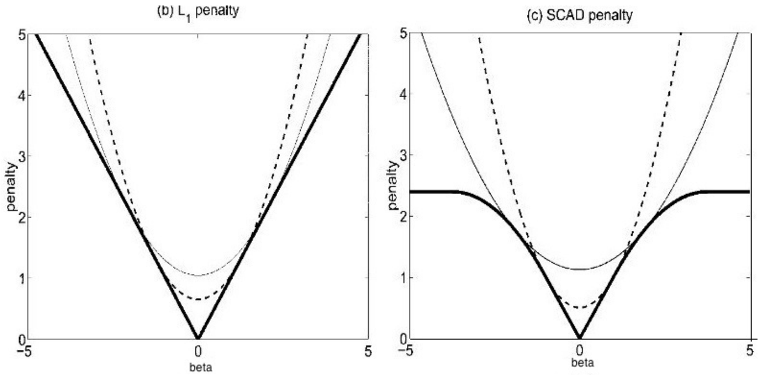
$$p_\lambda(|\beta|) = \lambda|\beta|$$

- Smoothly Clipped Absolute Deviation (SCAD):

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & , |\beta| \le \lambda \\ -\left(\dfrac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}\right), & \lambda < |\beta| \le a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & , |\beta| > a\lambda \end{cases}$$

where $a = 3.7$

Plot of $p_\lambda(|\beta|)$:



What is good penalty function?

- **Unbiasedness**: The estimator is nearly unbiased when the true unknown parameter is large to avoid modeling bias.
- **Sparsity**: The estimator is a thresholding rule, which sets small estimated coefficients to zero.
- **Continuity**: The estimator is continuous in data $\boldsymbol{x}$ to avoid instability in model prediction.

	Unbiasedness	Sparsity	Continuity
L_1-penalty(LASSO)	no	yes	yes
L_2-penalty(ridge regression)	no	no	yes
SCAD	yes	yes	yes

Conti. Example

- Consider model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{10} X_{i10} + \varepsilon_i \quad , i = 1, 2, \ldots, 100$$

  where $\mathbf{X} \sim MVN(0, \Sigma)$, Σ be covariance matrix with for $j, k = 1, \ldots, 10,\ j \neq k,\ Var(X_j) = 1,\ Cov(X_j, X_k) = 0.5$ and $\varepsilon_i \sim N(0, 3^2),\ i = 1, \ldots, 100.$
  the true $\boldsymbol{\beta_0} = (3, 1.5, 2, -7, 15, 0, 0, 0, 0, 0)$
- With 100 replication, count the number of estimated value greater than 10^{-3}. The parameter λ choosed by 10-fold cross validation.

	SCAD($\lambda = 0.3792$)		LASSO($\lambda = 0.2636$)	
	bias(sd)	count(%)	bias(sd)	count(%)
$\beta_1 = 3$	0.0415(0.42)	1	-0.1529(0.42)	1
$\beta_2 = 1.5$	-0.0394(0.52)	1	-0.1655(0.42)	1
$\beta_3 = 2$	0.0390(0.43)	1	-0.1435(0.42)	1
$\beta_4 = 7$	-0.0571(0.38)	1	0.3294(0.39)	1
$\beta_5 = 15$	0.0740(0.46)	1	-0.1395(0.45)	1
$\beta_6 = 0$	-0.0230(0.23)	0.46	0.0272(0.25)	0.59
$\beta_7 = 0$	0.0053(0.17)	0.36	0.0547(0.22)	0.54
$\beta_8 = 0$	-0.0303(0.21)	0.44	0.0215(0.25)	0.55
$\beta_9 = 0$	0.0218(0.19)	0.48	0.0744(0.25)	0.61
$\beta_{10} = 0$	-0.0279(0.16)	0.39	0.0045(0.21)	0.58

	SCAD		LASSO	
$\lambda = 0.3792$	bias(sd)	count(%)	bias(sd)	count(%)
$\beta_1 = 3$	0.0415(0.42)	1	-0.2232(0.42)	1
$\beta_2 = 1.5$	-0.0394(0.52)	1	-0.2341(0.41)	1
$\beta_3 = 2$	0.0390(0.43)	1	-0.2100(0.42)	1
$\beta_4 = 7$	-0.0571(0.38)	1	0.5049(0.39)	1
$\beta_5 = 15$	0.0740(0.46)	1	-0.2126(0.45)	1
$\beta_6 = 0$	-0.0230(0.23)	0.46	0.0420(0.19)	0.47
$\beta_7 = 0$	0.0053(0.17)	0.36	0.0578(0.18)	0.42
$\beta_8 = 0$	-0.0303(0.21)	0.44	0.0317(0.20)	0.43
$\beta_9 = 0$	0.0218(0.19)	0.48	0.0802(0.20)	0.41
$\beta_{10} = 0$	-0.0279(0.16)	0.39	0.0175(0.15)	0.39

	SCAD		LASSO	
$\lambda = 0.7585$	bias(sd)	count(%)	bias(sd)	count(%)
$\beta_1 = 3$	0.1655(0.50)	1	-0.4177(0.44)	1
$\beta_2 = 1.5$	-0.4698(0.57)	0.97	-0.4319(0.42)	1
$\beta_3 = 2$	-0.2262(0.64)	1	-0.4000(0.44)	1
$\beta_4 = 7$	0.0947(0.38)	1	1.1121(0.40)	1
$\beta_5 = 15$	0.2376(0.46)	1	-0.4214(0.46)	1
$\beta_6 = 0$	0.0018(0.05)	0.18	0.0357(0.10)	0.26
$\beta_7 = 0$	0.0148(0.06)	0.14	0.0462(0.12)	0.23
$\beta_8 = 0$	0.0083(0.06)	0.12	0.0396(0.12)	0.19
$\beta_9 = 0$	0.0186(0.07)	0.15	0.0616(0.15)	0.27
$\beta_{10} = 0$	0.0035(0.04)	0.09	0.0163(0.07)	0.18

Conclusion

- Variable selection via penalized least squares.

	Unbiasedness	Sparsity	Continuity
L_1-penalty(LASSO)	no	yes	yes
L_2-penalty(ridge regression)	no	no	yes
SCAD	yes	yes	yes