High Dimensional Data and Statistical Learning

Lecture 7: Variants of Lasso

(Adaptive Lasso, Elastic Net and Group Lasso)

Weixing Song

Department of Statistics
Kansas State University

Outline

Outline

Model:

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i, \quad 1 \leq i \leq n$$

Notations:

- Response: $\mathbf{y} = (y_1, \ldots, y_n)^T$.
- Predictors: $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, 2, \ldots, p$.
- Design Matrix: $\mathbf{X}_{n\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$.
- Residuals: $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^T$.
- Regression Coefficients: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.
- True Regression Coefficients: $\boldsymbol{\beta}^o = (\beta_1^o, \ldots, \beta_p^o)^T$.
- Oracle Set: $\mathcal{O} = \{j : \beta_j^o \neq 0\}$.
- Underlying Model Dimension: $d^0 = \|\mathcal{O}\| = \#\{j : \beta_j^o \neq 0\}$.

Centering and Standardization

WLOG, we assume that the response and predictors are centered and the predictors are standardized as follows

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad 1 \leq j \leq p.$$

Then there is no intercept in the model.

Each predictor is standardized to have the same magnitude in L_2. So the corresponding regression coefficients are "comparable".

After model fitting, the results can be readily transformed back to the original scale.

Lasso: Review

The Lasso estimator of $\boldsymbol{\beta}^o$ is

$$\hat{\boldsymbol{\beta}}(\lambda) = \text{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}.$$

Consider the set of estimated variables using Lasso

$$\hat{\mathcal{O}}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0, j = 1, 2, \ldots, p\}.$$

From the analysis of the Lars algorithm (Efron et al., 2004), we know that

$$|\hat{\mathcal{O}}(\lambda)| \leq \min(n, p) \quad \text{for all } \lambda.$$

Lasso tends to select a superset of the relevant covariates from $\mathcal{O}$. Here is one result that captures this property:

For some $C > 0$, define

$$\mathcal{O}^{(C)} = \{j : |\beta_j^o| \geq C, j = 1, 2, \ldots, p\}, \quad \mathcal{O} = \{j : |\beta_j^o| \neq 0, j = 1, 2, \ldots, p\},$$

where $\boldsymbol{\beta}^o$ is the true underlying parameter. One can show that for any fixed $0 < C < \infty$, as $n \to \infty$,

$$P(\hat{\mathcal{O}}(\lambda) \supset \mathcal{O}^{(C)}) \to 1.$$

So we do not miss any relevant covriates.

The question that follows is that when and how $\lim \hat{\mathcal{O}}(\lambda) = \mathcal{O}$.

Ideally we want our model to exactly select the set of true covariates as $n \to \infty$. This property is called consistency and it is the first requirement of an oracle procedure.

Definition: Oracle Procedure

Denote $\hat{\boldsymbol{\beta}}(\delta)$ the coefficient estimator for fitting procedure δ. We call δ an oracle procedure if $\hat{\boldsymbol{\beta}}(\delta)$ (asymptotically) has the following properties:

- **Consistency:** Identifies right subset model: $\{j : \beta_j(\delta) \neq 0\} = \mathcal{O}$,
- **Asymptotic Normality:** $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{O}}(\delta) - \boldsymbol{\beta}^o_{\mathcal{O}}) \xrightarrow{\mathscr{L}} N(0, \Sigma_o)$, where Σ_o is the covariance matrix knowing the true subset model.

For the sign consistency of Lasso, we introduced the irrepresentable condition.

Outline

An approach to obtaining a convex objective function which yields oracle estimators is using a weighted L_1 penalty with weights determined by an initial estimator. See Zou (2006) for detail.

Adaptive Lasso estimates $\hat{\boldsymbol{\beta}}$ is defined by

$$\hat{\boldsymbol{\beta}}(\lambda) = \text{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\},$$

where the weights can be constructed as

$$w_j = \begin{cases} |\tilde{\beta}_j|^{-\gamma}, & \text{if } \tilde{\beta}_j \neq 0; \\ \infty, & \text{if } \tilde{\beta}_j = 0. \end{cases}$$

We can use OLS (small p case) or Lasso estimator (large p case) as $\tilde{\boldsymbol{\beta}}$.

If $\tilde{\beta}_j = 0$, then $\hat{\beta}_j = 0$. If $|\tilde{\beta}_j|$ is large, the penalty is small for the j-th coefficient, and vice versa.

Adaptive Lasso: Computation

We can use the same algorithms for solving Lasso problems to solve adaptive Lasso.

Without loss of generality, assume that all $\tilde{\beta}_j \neq 0$ or all $w_j \neq \infty$.

Define $W = \text{diag}(w_j)_{p\times p}$, and denote $\beta_j^* = w_j\beta_j$, and $\boldsymbol{\beta}^* = W\boldsymbol{\beta}$. Then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}W^{-1}W\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \text{ where } \mathbf{X}^* = \mathbf{X}W^{-1},$$

and

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j| = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_1.$$

If $\hat{\boldsymbol{\beta}}^*(\lambda)$ is the solution of the above Lasso problem, then

$$\hat{\boldsymbol{\beta}}(\lambda) = W^{-1}\hat{\boldsymbol{\beta}}^*(\lambda)$$

will be the adaptive Lasso solution.

With a proper choice of λ, the adaptive Lasso enjoys the oracle properties. The following result is from Zou (2006) for fixed p case.

Theorem

Suppose that $\sqrt{n}\lambda \to 0$ and $\lambda n^{(\gamma+1)/2} \to \infty$. Then the adaptive Lasso estimates must satisfy the following:

- Consistency in Variable Selection: $\lim_{n\to\infty} P(\hat{\mathcal{O}}(\lambda) = \mathcal{O}) = 1$ as $n \to \infty$;
- Asymptotic Normality: $\sqrt{n}(\hat{\beta}_{\mathcal{O}}(\delta) - \beta^o_{\mathcal{O}}) \Longrightarrow N(0, \sigma^2 \Sigma_{11}^{-1})$.

Therefore, the adaptive Lasso is consistent without requiring the irrepresentable condition.

Extensions to $p = p(n)$ is made by Huang et al. (2008).

Sing Consistency in Large p Case

Consider the same setting as we derive the sign consistency for Lasso.

Let c_1 and c_2 be the smallest and largest eigenvalues of Σ_{11} respectively. Huang et al. (2008) showed the sign consistency of adaptive Lasso under the following conditions, with appropriately choosing tuning sequence:

- $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$.
- There exist w_{n*} and w_n^* such that

$$P\left(\max_{j\in\mathcal{O}} |w_{nj}| \leq w_{n*}, \text{ and } \min_{j\notin\mathcal{O}} |w_{nj}| > w_{n^*}\right) = 1 - o(1).$$

- $\beta_* > 2w_{n*}\lambda\sqrt{d^o}/c_1$, $w_n^* > w_{n*}\sqrt{c_2\, d^o}/c_1$.

Therefore, the irrepresentable condition is not required, if the initial estimator is "good enough".

Outline

There are a number of limitations of the Lasso estimator, which make the Lasso inappropriate for variable selection in some situations.

In the $p > n$ case, the Lasso selects at most n variables before it saturates. This could be a limiting feature for a variable selection method.

Lasso has no grouping property, it tends to only select one variable among a group of highly correlated variables.

For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the Lasso is dominated by ridge regression.

For any fixed non-negative $\lambda = (\lambda_1, \lambda_2)$, the elastic net (eNet) criterion is defined as

$$L(\lambda_1, \lambda_2, \beta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1\|\beta\|_1 + \frac{1}{2}\lambda_2\|\beta\|^2.$$

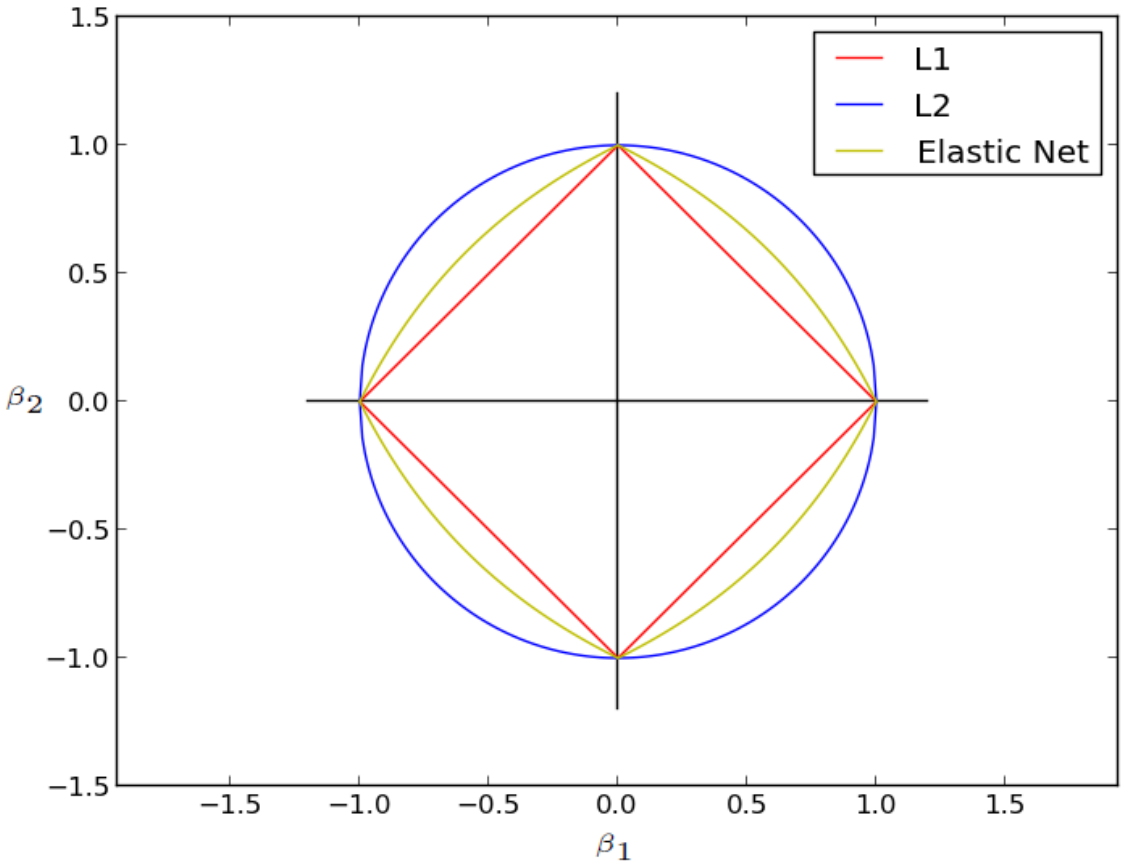
It is equivalent to a constraint LS method. Let $\alpha = \lambda_2/(2\lambda_1 + \lambda_2)$. Then an equivalent optimization problem is

$$\hat{\beta} = \text{argmin}_{\beta}\|\mathbf{y} - \mathbf{X}\beta\|^2, \text{ subject to } (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|^2 \leq t.$$

The eNet penalty is a convex combination of the Lasso and ridge penalty. For all $\alpha \in [0, 1)$, the penalty function is singular at 0 and it is strict convex for all $\alpha > 0$.

Penalty Function in eNet

Take $p = 2$ as an example.



eNet: Solution

The eNet objective function $L(\lambda_1, \lambda_2, \boldsymbol{\beta})$ is equivalent to a Lasso problem on augmented data $(\mathbf{X}^*, \mathbf{y})$, where

$$\mathbf{X}^*_{(n+p)\times p} = (\mathbf{X}^T, \sqrt{n\lambda_2} I)^T, \quad \mathbf{y}^* = (\mathbf{y}^T, \mathbf{0}^T)^T.$$

Then the naive eNet criterion becomes

$$L(\lambda, \boldsymbol{\beta}) = \frac{1}{2n} \|\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1.$$

Since $\text{rank}(\mathbf{X}^*) = p$, the elastic net can potentially select all p predictors.

The above formulation also says that the Naive eNet enjoys the computational advantages of Lasso.

eNet: Orthogonal Design

In the orthogonal design case, that is, $\mathbf{x}_j^T \mathbf{x}_k = \delta_{jk}$, where $\delta_{jk} = I(j = k)$, the j-th eNet estimator is

$$\hat{\beta}_j = \frac{\text{sgn}(\hat{\beta}_j^{\text{LS}})(|\hat{\beta}_j^{\text{LS}}| - \lambda_1)_+}{1 + \lambda_2} = \frac{\mathcal{S}(\hat{\beta}_j^{\text{LS}}; \lambda_1)}{1 + \lambda_2},$$

where $\mathcal{S}(\cdot; \lambda_1)$ is the soft-thresholding operator.

To compare, recall the estimates for ridge regression and Lasso under orthogonal design:

$$\hat{\beta}_j^{\text{ridge}} = \frac{\hat{\beta}_j^{\text{LS}}}{1 + \lambda_2}, \quad \hat{\beta}_j^{\text{Lasso}} = \mathcal{S}(\hat{\beta}_j^{\text{LS}}; \lambda_1).$$

eNet: Solution in Orthogonal Design

Note that

$$\begin{aligned} L(\lambda_1, \lambda_2, \boldsymbol{\beta}) &= \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda_2\|\boldsymbol{\beta}\|^2 \\ &= \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{LS}}\|^2 + \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{LS}}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda_2\|\boldsymbol{\beta}\|^2. \end{aligned}$$

Note that $\mathbf{X}^T\mathbf{X}/n = I$, then minimizing $L(\lambda_1, \lambda_2, \boldsymbol{\beta})$ is equivalent to minimizing

$$\frac{1}{2}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\mathrm{LS}}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda_2\|\boldsymbol{\beta}\|^2.$$

Easy to see the minimization problem is separable. That is, for each j, the eNet solution is

$$\begin{aligned} \hat{\beta}_j &= \mathrm{argmin}_{\beta_j}\left\{\frac{1}{2}(\beta_j - \hat{\beta}_j^{\mathrm{LS}})^2 + \lambda_1|\beta_j| + \frac{1}{2}\lambda_2\beta_j^2\right\} \\ &= \mathrm{argmin}_{\beta_j}\left\{\frac{1}{2}\left(\beta_j - \frac{\hat{\beta}_j^{\mathrm{LS}}}{\lambda_2+1}\right)^2 + \frac{\lambda_1}{\lambda_2+1}|\beta_j|\right\} = \mathcal{S}\left(\frac{\hat{\beta}_j^{\mathrm{LS}}}{\lambda_2+1}; \frac{\lambda_1}{\lambda_2+1}\right). \end{aligned}$$

Grouping Effect

Sometimes, identification of "grouped variables" is desired.

Consider a general penalized regression approach

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda J(\boldsymbol{\beta}) \right\},$$

where $J(\cdot)$ is the penalty function.

A regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal.

Question: Does OLS or Lasso have the grouping effect?

Lemma

Assume $\mathbf{x}_i = \mathbf{x}_j,\ i, j \in \{1, 2, \ldots, p\}$.

- If $J(\cdot)$ is strictly convex, then $\hat{\beta}_i = \hat{\beta}_j$, for $\forall \lambda > 0$.
- If $J(\cdot) = \|\boldsymbol{\beta}\|_1$, and $\hat{\beta}_1, \hat{\beta}_2 \geq 0$, then $\hat{\boldsymbol{\beta}}^*$ is another minimizer of the generic penalization function, where

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k & \text{if } k \neq i, j; \\ s(\hat{\beta}_i + \hat{\beta}_j) & \text{if } k = i; \\ (1-s)(\hat{\beta}_i + \hat{\beta}_j) & \text{if } k = j \end{cases}$$

for any $s \in [0, 1]$.

Lasso is convex but not strictly convex; the eNet with $\lambda_2 > 0$ is strictly convex.

Theorem

Let $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \hat{\boldsymbol{\beta}}$ be the naive eNet estimate. Let $\rho_{jk} = \text{corr}(\mathbf{x}_j, \mathbf{x}_k)$. If $\hat{\beta}_j \hat{\beta}_k > 0$, we have

$$\frac{\sqrt{n}|\hat{\beta}_j - \hat{\beta}_k|}{\|\mathbf{y}\|} \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho_{jk})}.$$

Proof: Take derivative with β_j, β_k of the eNet objective function, respectively, we get

$$-n^{-1}\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda_1 \text{sgn}(\hat{\beta}_j) + \lambda_2 \hat{\beta}_j = 0$$
$$-n^{-1}\mathbf{x}_k^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda_1 \text{sgn}(\hat{\beta}_k) + \lambda_2 \hat{\beta}_k = 0.$$

Hence, $\lambda_2(\hat{\beta}_j - \hat{\beta}_k) = n^{-1}(\mathbf{x}_j - \mathbf{x}_k)^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. Note that $\rho_{jk} = \mathbf{x}_j^T\mathbf{x}_k/n$, and $\|\mathbf{x}_j - \mathbf{x}_k\|^2 = 2n(1-\rho_{jk})$, we have

$$n^{-1}\left|(\mathbf{x}_j - \mathbf{x}_k)^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right| \leq n^{-1}\|\mathbf{x}_j - \mathbf{x}_k\| \cdot \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|.$$

The theorem follows the fact that $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\| \leq \|\mathbf{y}\|$.

Remarks on eNet

The LHS of the inequality in the theorem is unitless, which describes the difference between the coefficient path of predictors j and k.

The naive eNet does not perform "satisfactorily".

The ridge penalty introduce an extra bias factor $1/(1+\lambda_2)$. This ridge shrinkage on top of the Lasso shrinkage is the double shrinkage effect discussed in Zou and Hastie (2005).

Zou and Hastie (2005) proposed to removes the ridge shrinkage factor by multiplying the naive eNet by $(1+\lambda_2)$ to obtain the eNet estimator

$$\hat{\boldsymbol{\beta}}_{\text{eNet}} = (1+\lambda_2)\hat{\boldsymbol{\beta}}_{\text{naive}-\text{eNet}}.$$

Scaling preserves the variable selection property but solves the shrinkage problem.

The eNet estimator can be considered as Lasso estimator based on augmented data.

The theoretical analysis can also take advantage of this fact. Prediction bound and sign consistency can be similarly established.

We can consider a more general criterion

$$\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda_1\sum_{j=1}^{p}w_j|\beta_j|+\frac{1}{2}\lambda_2\boldsymbol{\beta}^T Q\boldsymbol{\beta},$$

where w_j's are adaptive weights and Q is a given positive semi-definite matrix.

Outline

Model Set-up

Consider

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon},$$

where

- $\mathbf{y} = (y_1, \ldots, y_n)^T$
- $\mathbf{X}_j = (\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jd_j})$ represents the design matrix of the j-th group of variables of size d_j with $\mathbf{x}_{jk} = (x_{1jjk}, \ldots, x_{njk})^T$
- $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_j})^T$
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$

We are interested in selecting groups of variables.

Examples include analysis of high-dimensional genomic data in order to find functional groups, pathways, groups consisting of co-expressed genes, SNPs in the same haplotype block, and many others.

Example (Impact Study): The impact study was part of a three-year project designed to measure the impact of nutritional policies and environmental change on obesity in the high school students enrolled in Seattle Public Schools. The primary goal of this study is to determine the effects of different risk factor on body mass index (BMI).

Table 1 provides the definitions of the variables included in the study. The 25 covariates can be naturally classified into eight different groups, measuring different aspects such as food sources and demographics. The response variable is the logarithm of the body mass index.

There are 799 subjects with complete records.

Table. *Impact study. Dictionary of covariates*

Group	Variable	Type	Definition
Age	V1	C	Age
	V2	C	Age^2
Gender	V3	B	Female gender
Ethnicity	V4	B	Ethnic (American Indian/Alaska native)
	V5	B	Ethnic (Hispanic/Latino)
	V6	B	Ethnic (Asian)
	V7	B	Ethnic (Native Hawaiian/Pacific Islander)
	V8	B	Ethnic (White)
	V9	B	Ethnic (Do not know)
	V10	B	No answer
	V11	B	Bi/multi-racial
	V12	B	Speaking other language
Food source	V13	B	Breakfast/lunch from cafeteria more than 3 times per week
	V14	B	Food from à la carte more than 3 times per week
	V15	B	Fast food
	V16	B	Food from home more than 3 times per week
Consumption	V17	C	Fizzy drinks
(Unhealthy)	V18	B	Sweets
	V19	B	Crisps
	V20	B	Cake
	V21	B	Ice cream
Consumption	V22	C	Milk
(Healthy)	V23	C	Fruit and vegetable
School	V24	B	School A or B
Physical activity	V25	C	Mild physical activity
	V26	C	Hard physical activity

Type, type of variable; C, continuous; B, binary.

Group Lasso: Formulation

Let $\Sigma_j = \mathbf{X}_j^T \mathbf{X}_j / n$. The group Lasso criterion is defined as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \lambda), \tag{1}$$

where

$$L(\boldsymbol{\beta}; \lambda) = \frac{1}{2n} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{X}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{p} \sqrt{d_j} \|\boldsymbol{\beta}_j\|_{\Sigma_j},$$

and $\|\boldsymbol{\beta}_j\|_{\Sigma_j} = (\boldsymbol{\beta}_j^T \Sigma_j \boldsymbol{\beta}_j)^{1/2}$.

Write $\Sigma_j = R_j^T R_j$ for a $d_j \times d_j$ upper triangular matrix R_j via Cholesky decomposition.

Assume that Σ_j is invertible. Let $\tilde{\mathbf{X}}_j = \mathbf{X}_j R_j^{-1}$ and $\mathbf{b}_j = R_j \boldsymbol{\beta}_j$, and denote

$$\hat{\mathbf{b}} = \text{argmin}_{\mathbf{b}} L(\mathbf{b}; \lambda) = \text{argmin}_{\mathbf{b}} \left\{ \frac{1}{2n} \left\| \mathbf{y} - \sum_{j=1}^{p} \tilde{\mathbf{X}}_j \mathbf{b}_j \right\|^2 + \lambda \sum_{j=1}^{p} \sqrt{d_j} \|\mathbf{b}_j\| \right\}.$$

Then $\hat{\boldsymbol{\beta}}_j = R_j^{-1} \hat{\mathbf{b}}_j$.

Note that

$$n^{-1}\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j = (R_j^{-1})^T(n^{-1}\mathbf{X}_j^T\mathbf{X}_j)R_j^{-1} = I_{d_j}.$$

Thus using the $\|\cdot\|_{\Sigma_j}$ norm in (1) amounts to standardizing the desing matrices $\mathbf{X}_j$'s.

WLOG, we assume that $\mathbf{X}_j$ is orthonormalized with $n^{-1}\mathbf{X}_j^T\mathbf{X}_j = I_{d_j}$.

The implementation of the group Lasso is an extension of the shooting algorithm (Fu, 1999) for the Lasso. It is motivated by the following proposition, which is a direct consequence of Karush-Kuhn-Tucker (KKT) conditions.

Necessary and Sufficient Conditions for Group Lasso Solution

A necessary and sufficient condition for $\hat{\mathbf{b}}$ to be a solution to $\operatorname{argmin}_{\mathbf{b}} L(\mathbf{b};\lambda)$ is

$$-n^{-1}\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}) + \frac{\lambda\hat{\mathbf{b}}_j\sqrt{d_j}}{\|\hat{\mathbf{b}}_j\|} = 0, \qquad \forall\,\hat{\mathbf{b}}_j \neq 0,$$

$$\|n^{-1}\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})\| \leq \lambda\sqrt{d_j} \qquad \forall\,\hat{\mathbf{b}}_j = 0, \quad j = 1,2,\ldots,p.$$

Recall that $n^{-1}\mathbf{X}_j^T\mathbf{X}_j = I_{d_j}$. It can be verified that the solution to the equations in proposition is

$$\hat{\mathbf{b}}_j = \left(1 - \frac{\lambda\sqrt{d_j}}{\|n^{-1}\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}^T\hat{\mathbf{b}}_{-j})\|}\right)_+ n^{-1}\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}^T\hat{\mathbf{b}}_{-j}),$$

where $\mathbf{b}_{-j} = (\hat{\mathbf{b}}_1^T, \ldots, \hat{\mathbf{b}}_{j-1}, \mathbf{0}^T, \hat{\mathbf{b}}_{j+1}^T, \ldots, \hat{\mathbf{b}}_p^T)^T$.

The Group Lasso solution can therefore be obtained by iteratively applying the above equation.

In particular, if we further assume that $\mathbf{X}_j^T\mathbf{X}_k = 0, j \neq k$, then it is easy to see that The problem simplifies to that of estimation in p single-group models of the form $\mathbf{y} = \mathbf{X}_j\mathbf{b}_j + \boldsymbol{\varepsilon}$.

Let $\mathbf{z}_j = (\mathbf{X}_j^T\mathbf{X}_j)^{-1}\mathbf{X}_j^T\mathbf{y} = n^{-1}\mathbf{X}_j^T\mathbf{y}$ be the LSE of $\mathbf{b}_j$. Then based on the previous proposition, we can show that the group Lasso estimator for the j-th group is

$$\hat{\mathbf{b}}_j(\lambda) = \text{argmin}_{\mathbf{b}_j}\left\{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}_j\mathbf{b}_j\|^2 + \lambda\sqrt{d_j}\|\mathbf{b}_j\|\right\} = \left(1 - \frac{\lambda\sqrt{d_j}}{\|\mathbf{z}_j\|}\right)_+ \mathbf{z}_j. \quad (2)$$

Group Lasso: Group Coordinate Descent

The group coordinate descent (GCD) algorithm is a straightforward extension of the coordinate descent algorithm we have discussed.

Suppose the current values for the group coefficients $\tilde{\boldsymbol{\beta}}_k^{(s)}$, $k \neq j$ are given. We want to minimize L with respect to $\boldsymbol{\beta}_j$. Define

$$L_j(\boldsymbol{\beta}_j; \lambda) = \frac{1}{2n} \left\| \mathbf{y} - \sum_{k \neq j} \mathbf{X}_k \tilde{\boldsymbol{\beta}}_k^{(s)} - \mathbf{X}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sqrt{d_j} \|\boldsymbol{\beta}_j\|.$$

Denote $\tilde{\mathbf{y}}_j = \sum_{k \neq j} \mathbf{X}_k \tilde{\boldsymbol{\beta}}_k^{(s)}$ and $\tilde{\mathbf{z}}_j = n^{-1} \mathbf{X}_j^T (\mathbf{y} - \tilde{\mathbf{y}}_j)$. Let $\tilde{\boldsymbol{\beta}}_j$ denote the minimizer of $L_j(\boldsymbol{\beta}_j; \lambda)$. We have

$$\tilde{\boldsymbol{\beta}}_j = \mathcal{S}(\tilde{\mathbf{z}}_j; \sqrt{d_j}\lambda),$$

where $\mathcal{S}$ is the multivariate soft-threshold operator defined by

$$\mathcal{S}(\mathbf{z}; t) = \left(1 - \frac{t}{\|\mathbf{z}\|}\right)_+ \mathbf{z}.$$

Coordinate Descent Algorithm

For any fixed λ,

1. Start with an initial value for $\tilde{\boldsymbol{\beta}}^{(0)} = \left(\tilde{\boldsymbol{\beta}}_1^{(0)\,T}, \ldots, \tilde{\boldsymbol{\beta}}_p^{(0)\,T}\right)^T$;
2. In the $s+1$-th iteration,
   (1). Let $j = 1$;
   (2). Calculate

$$\tilde{\mathbf{z}}_j = n^{-1}\mathbf{X}_j^T(\mathbf{y} - \tilde{\mathbf{y}}_j) = n^{-1}\mathbf{X}_j^T(\mathbf{y} - \tilde{\mathbf{y}} + \mathbf{X}_j\tilde{\boldsymbol{\beta}}_j^{(s)}) = n^{-1}\mathbf{X}_j r + \tilde{\boldsymbol{\beta}}_j^{(s)},$$

   where $\tilde{\mathbf{y}} = \sum_{j=1}^{p} \mathbf{X}_j\tilde{\boldsymbol{\beta}}_j^{(s)}$ is the vector of current fitted values and $r = \mathbf{y} - \tilde{\mathbf{y}}$ is the current residual.
   (3). Update $\tilde{\boldsymbol{\beta}}_j^{(s+1)}$ using $\mathcal{S}(\tilde{\mathbf{z}}_j; \sqrt{d_j}\lambda)$. If $j = p$, then exit step 2.
   (4). Update r using $r - \mathbf{X}_j(\tilde{\boldsymbol{\beta}}_j^{(s+1)} - \tilde{\boldsymbol{\beta}}_j^{(s)})$.
   (5). Let $j \Leftarrow j + 1$, repeat (2)-(4).
3. Repeat step 2 for $s+1$ until convergence.

NOTE: The above algorithm is designed for the cases in which the predictors are standardized to have L_2-norm n.

The last step ensures that r always holds the current values of the residuals.

The GCD algorithm has the potential to be extremely efficient, in that the above three steps require only $O(2nd_j)$ operations.

One full iteration can be completed at a computational cost of $O(nd)$ operations.

The algorithm described above can be used repeatedly to compute $\hat{\boldsymbol{\beta}}(\lambda)$ on a grid of values of λ.

Let $\lambda_{\max}$ be the smallest value for which all coefficients are 0 and $\lambda_{\min}$ be the minimum value of λ. From (2), one can take $\lambda_{\max} = \max_{1 \leq j \leq p} \|n^{-1}\mathbf{X}_j\mathbf{y}\|$. If $\sum_{j=1}^{p} d_j < n$ and the design matrix if full rank, $\lambda_{\min}$ can be 0. In other settings, we use $\lambda_{\min} = 0.001\lambda_{\max}$.

Let $\lambda_{\max} > \lambda_1 > \cdots > \lambda_K > \lambda_{\min}$ be a grid of decreasing λ-values. We start at $\lambda_{\max}$ for which $\tilde{\boldsymbol{\beta}}$ has the solution 0, and proceed along the grid using the value of $\hat{\boldsymbol{\beta}}$ at the previous point of λ in the grid as the initial value for the current point in the algorithm.

Outline

References

- Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Annals of Statistics 36(2), 587-613.
- Huang, J., S. Ma, H. Xie, and C.-H. Zhang (2009). A group bridge approach for variable selection. Biometrika 96(2), 339-355.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49-67.
- Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418-1429.
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B 67(2), 301-320.
- Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics 37, 1733-1751.