Model selection and estimation in regression with grouped variables

Ming Yuan and Yi Lin

Haozhiheng，Nankai University

关于nn-garrotte,lasso和LARS

- TIBSHIRANI.png

history of lasso

- the lasso is just regression with an l_1-norm penalty,and l_1-norms have been around for a long time.
- Breiman,1995 proposed non-negative garrotte
- his idea was to minimize ,with respect to $c = c_j$, $\sum_{i=1}^{N}(y_i - \sum_j c_j x_{ij}\hat{\beta}_j)^2$ subject to $c_j \geq 0, \sum_{j=1}^{p} c_j \leq t$ $\hat{\beta}_j$are usual least square estimates.
- Tibshirani combined the two stages into one and named it lasso.
- the idea of lasso is to minimize $\sum_{i=1}^{N}(y_i - \sum_j c_j x_{ij}\hat{\beta}_j)^2$ subject to $,\sum_{j=1}^{p} \beta_j \leq t$

- 

- Efron 2004 proposed LARS algorithm, and make a connection between lasso and LARS.

问题的提出

- $$Y = \sum_{j=1}^{J} X_j \beta_j + \varepsilon \cdots\cdots (1.1)$$
- Y is an $n \times 1$ vector,ε $N_n(0, \sigma^2 I)$,X_j is an $n \times p_j$ matrix corresponding to the jth factor and β_j is a coefficient vector of size p_j,$j = 1, \ldots, J$
- What the concept **Grouped Variables** means
- goal of this paper is to **select important factors** for accurate estimation in equation(1)

模型选择文献回顾

- Tibshirani(1996) proposed lasso
- Efron et al.(2004) proposed least angle regression selection
- lasso and LARS are designed for selecting individual input variables,not for general factor selection.
- grouped lasso and grouped LARS,also consider a group version of the non-negative garrotte.
- to select the final models on the solution paths of group selection methods,we use C_p criterion.

lasso

- Robert Tibshirani proposed the popular Lasso in paper Regeression Shrinkage and Selection via the Lasso
- the full name of lasso is least absolute shrinkage and selection operator
- lasso arises from constrained form of ordinary Least square regression where the sum of the absolute value of the regression coefficients is constrained to be smaller than a specified parameter.
- minimize $\|y - X\beta\|^2$ subject to $\sum_{j=1}^{J} |\beta_j| \leq t$
- provide the lasso parameter t is small enough,some of the regression coefficients will be exactly zero.
- by increasing the lasso parameter in discrete steps, we obtain a sequence of regression coefficient where the nonzero coefficients at each step correspond to selected parameters.

group lasso

- given positive definite matrices $K_1, \ldots, K_J$ the group lasso estimate is defined as the solution to

$$\frac{1}{2}\|Y - \sum_{j=1}^{J} X_j\beta_j\|^2 + \lambda \sum_{j=1}^{J} \|\beta_j\|_{K_j}$$

- for a vector $\eta \in R^d$, $d \geq 1$,and a symmetric $d \times d$ positive definite matrix K,denote$\|\eta\|_K = (\eta' K\eta)^{\frac{1}{2}}$
$\|\eta\| = \|\eta\|_{I_d}$

Lasso 的性质

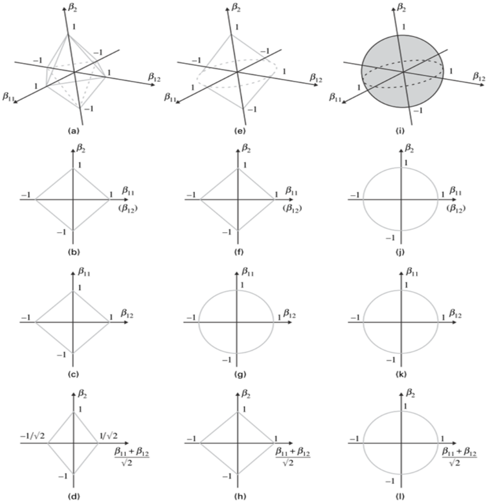

Fig. 1. (a)–(d) l_1-penalty, (e)–(h) group lasso penalty and (i)–(l) l_2-penalty

Lasso的性质

- The $l_1$penalty treats the three co-ordinate directions differently from other directions,and this encourages sparsity in individual coefficients.
- The $l_2$penalty treats all directions equally and does not encourage sparsity.
- the group lasso encourages sparsity at the factor level.

lasso 的算法

- extension of the shooting algorithm(Fu,1999)for the lasso.
- Karush-Kuhn-Tucker condition
- 这部分内容先放过，回头再来补过。先看后面的LARS

least angle regression algorithm

- Efron et al.2004 propose LAR algorithm for variable selection
- start with all coefficients equal to zero
- find the predictor most correlated with the response say X_j1
- take the largest **step** possible in the direction of this predictor until some other predictor,say X_j2,has as much correlation with the current residual.(at this point LARS parts company with Forward Selection)
- proceeds in a direction equiangular between the two predictors until a third variable X_j3 earns its way into the most correlated set
- proceeds equiangularly between X_j1, X_j2, X_j3,until a fourth variable enters and so on

geometry explanation for LAR algorithm

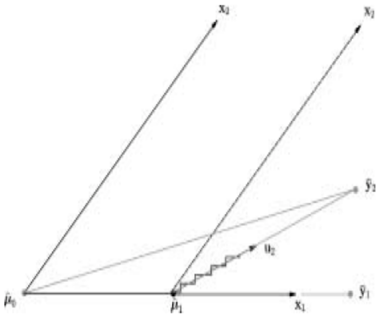

FIG. 2. *The LARS algorithm in the case of $m = 2$ covariates; $\bar{\mathbf{y}}_2$ is the projection of $\mathbf{y}$ into $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$. Beginning at $\hat{\boldsymbol{\mu}}_0 = \mathbf{0}$, the residual vector $\bar{\mathbf{y}}_2 - \hat{\boldsymbol{\mu}}_0$ has greater correlation with $\mathbf{x}_1$ than $\mathbf{x}_2$; the next LARS estimate is $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0 + \hat{\gamma}_1 \mathbf{x}_1$, where $\hat{\gamma}_1$ is chosen such that $\bar{\mathbf{y}}_2 - \hat{\boldsymbol{\mu}}_1$ bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$; then $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1 + \hat{\gamma}_2 \mathbf{u}_2$, where $\mathbf{u}_2$ is the unit bisector; $\hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{y}}_2$ in the case $m = 2$, but not for the case $m > 2$; see Figure 4. The staircase indicates a typical Stagewise path. Here LARS gives the Stagewise track as $\varepsilon \to 0$, but a modification is necessary to guarantee agreement in higher dimensions; see Section 3.2.*

-

group LARS

- Define the angle $\theta(r, X_j)$ between an n-vector r and a factor X_j
- $$cos^2\theta(r, X_j) = \frac{\|X_j' r\|^2}{\|r\|^2}$$
- find the solution path which is the projection of the current residual on the space spanned by the current factor
- proceed in the direction until find another factor X_{j2} s.t $\frac{\|X_{j1}' r\|^2}{p_{j1}} = \frac{\|X_{j2}' r\|^2}{p_{j2}}$

non-negative garrotte

- Breiman(1995) propose nn-garrotte
- estimate of $\beta_j$is the least square estimate $\hat{\beta}_j^{LS}$ scaled by a constant $d_j(\lambda)$ given by

$$d(\lambda) = argmin_d(\frac{1}{2}\|Y - Zd\|^2 + \lambda \sum_{j=1}^{J} d_j)$$

subject to $d_j \geq 0, \forall j$
where $Z = (Z_1, \ldots, Z_J)$ and $Z_j = X_j \hat{\beta}_j^{LS}$

group nn-garrotte

- $$d(\lambda) = argmin_d(\frac{1}{2}\|Y - Zd\|^2 + \lambda \sum_{j=1}^{J} p_j d_j)$$

  subject to $d_j \geq 0, \forall j$

- Theorem 1.The solution path of the group lasso is piecewise linear if and only if any group lasso solution $\hat{\beta}$ can be written as $\hat{\beta}_j = c_j \beta_j^{LS}$ for some scalars $c_1, \ldots, c_J$
- simple example
  for group lasso:$fraction(\beta) = \frac{\sum_j \|\beta_j\| \sqrt{p_j}}{\sum_j \|\beta_j^{LS}\| \sqrt{p_j}}$
  for the group nn-garrotte:$fraction(d) = \frac{\sum_j p_j d_j}{\sum_j p_j}$
  for the group LARS:

- $C_p typecriterion$:

$$C_p)\hat{\mu} = \frac{\|Y-\hat{\mu}\|^2}{\sigma^2} - n + 2df_{\mu,\sigma^2}, \text{where } df_{\mu,\sigma^2} = \sum_{i=1}^{n} \frac{cov(\hat{\mu}_i, Y_i)}{\sigma^2}$$

- positive cone condition
- bootstrap
- Theorem 2.Consider the model with the design matrix X being orthonormal.For any estimate on the solution path of the group lasso,group LARS,or the group nn-garrotte, we have $df = E(\tilde{df})$