Variable Inclusion and Shrinkage Algorithm in High Dimension

A.Mkhadri and M.Ouhourane

Faculty of Sciences-Semlalia, Marrakech

19th International Conference on Computational Statistics on August 22nd-27th 2010.

Table of contents

Introduction and motivation

We consider the standard linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

where

- $\mathbf{y} \in I\!R^n$ is the response
- $\mathbf{X}$ is the nxp model matrix, with $\mathbf{x}_j \in I\!R^n, j = 1, ..., p$, are the predictors
- β is a p-vector of unknown parameters which are to be estimated
- ε is a n-vector of (i.i.d.) random errors with mean 0 and variance σ^2

Introduction and motivation

OLS :

$$\widehat{\beta}_{OLS} = \text{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2.$$

Two alternatives class of methods :

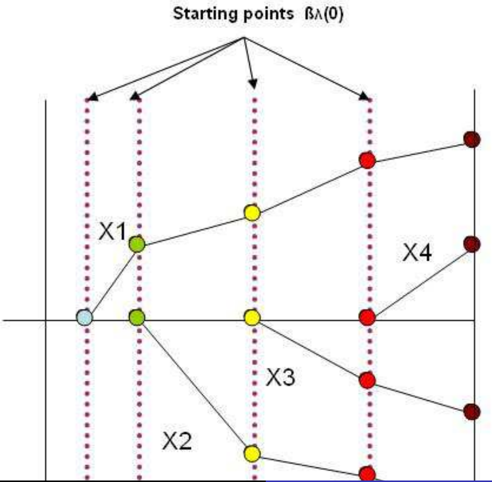
- Classical variable selection
  - Stepwise regression
  - Information criterion AIC, BIC
- Regularization methods

LASSO

Definition

$$\widehat{\beta}_{Lasso} = \text{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1.$$

LASSO



LASSO

Advantages

- Reduce the variability of the estimates by shrinking the coefficients
- Produces interpretable models by shrinking some coefficients to exactly zero

Disadvantages

- In high dimension, the Lasso selects at most n variables
- It's tends to select only some variable from the high correlated group of variables.
- The some tuning parameter is used for both variable selection and shrinkage.

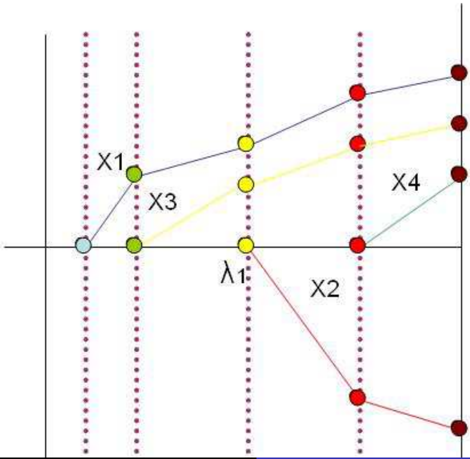
ELASTIC NET

Definition

$$\widehat{\beta}_{NaiveEnet} = \text{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2.$$

$$\widehat{\beta}_{Enet} = (1 + \lambda_2) * \widehat{\beta}_{Naive-Enet}.$$

ELASTIC NET



ELASTIC NET

Advantages

- Encourage a grouping effect
- No limitation on the number of variables that may be selected for the model

Disadvantages

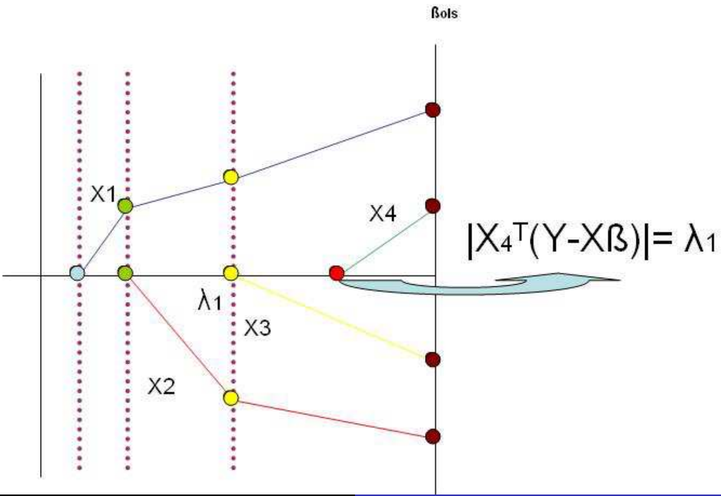
- It must be chosen between over shrink the correct variables and select a number of noise variables
- If some significative variables are ignored, It is not possible to restor

VISA

Definition

- Select the first set of variables using LASSO (starting point $\beta_\lambda(0)$)
- Eliminate the over shrinkage to this set and detects another set of significative variables Simultaneously.
- Eliminates the over shrinkage of the latter set of variables.

VISA



VISA

Advantages

- Select sparse models while avoiding over shrinkage problems

Disadvantages

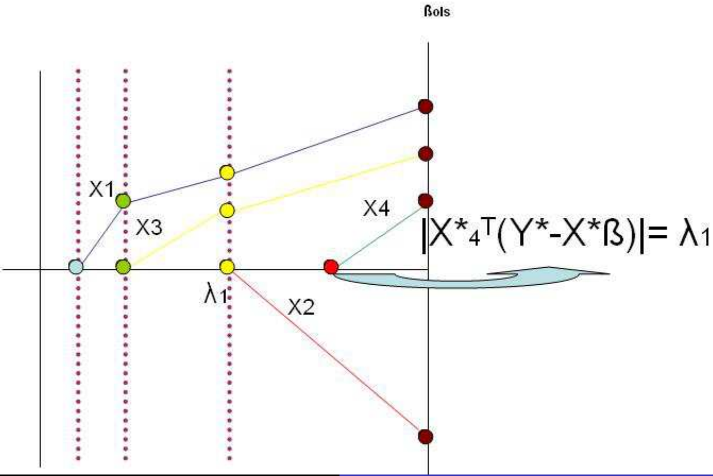
- It does not ensure the grouping effect
- The number of variables in the starting point is limited by number of observations n

VISA NET algorithm

Definition

- Select the first set of variables using Naive-Enet (starting point $\beta_{\lambda_1,\lambda_2}(0)$)
- Eliminate the over shrinkage to this set and detects another set of significative variables Simultaneously.
- Eliminate the over shrinkage of the lather set of variables.

VISA NET



VISA NET algorithm

Lemma1 :Given data set (y, X) and $(\lambda_1, \lambda_2, \phi)$, define an artificial data set by

$$\mathbf{X}^*_{(n+p)\times n} = \left(\frac{\mathbf{X}}{\sqrt{\lambda_2}\mathbf{I}}\right), \mathbf{y}^*_{(n+p)} = \left(\frac{\mathbf{y}}{0}\right)$$

then the $VISA_{ENET}$ is equivalent to a $VISA_{Lars}$ problem on the augmented data set

VISA-Net

Advantages

- ensure that we can select more than *n* variables In the starting set
- it can select groups of high correlated variables
- the over shrinkage of the coefficients and the number of noise variables can be decreased.

Theoretical Results

we show that $VISA_{ENET}$ has non-asymptotic bounds on its estimation errors. Given an index set $j \subset \{1, ..., p\}$ and X_j. Let $\psi(k)$ denote the smallest eigenvalue of the matrix $\{X_j^{*T} X_j^* , |j| \leq k\}$.

Theorem 1. Suppose that $\beta \in \mathbb{R}^p$ is an S-sparse coefficient vector. Consider an $a > 0$, and define $\tau_p = \sigma\sqrt{2(1+a)logp}$. If $\widehat{\beta}$ is a VISA estimator with k non-zero $\widehat{\beta}_j$ coefficients for which $\beta_j = 0$, and $\lambda_\infty = \|X^T(Y - X\widehat{\beta})\|_\infty$,then

$$P(\|\widehat{\beta} - \beta\|_2 > \frac{\lambda_\infty + \tau_p}{(S+k)^{-1/2}\psi(S+k) - \lambda_2}) \leq (p^a\sqrt{4\pi logp})^{-1}$$

The grouping effect and selecting others variables

We generate one data set of 50 observations and 40 predictors. We chose $\beta = (\underbrace{5,..,5}_{5}, \underbrace{3,..,3}_{5}, \underbrace{1,..,1}_{5}, \underbrace{0,...,0}_{25})$.

The predictors X were generated as follows :

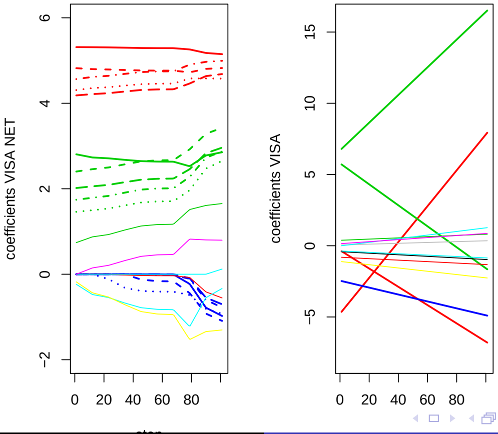
- $Z \sim N(0,5)$
- $Z_i = Z + \varsigma_i, \varsigma_i \sim N(0,1), i = 1,..,3$
- $x_i = Z_1 + \varepsilon_i^x, i = 1,..,5, \varepsilon_i^x \sim N(0,0.1)$
- $x_i = Z_2 + \varepsilon_i^x, i = 6,..,10, \varepsilon_i^x \sim N(0,0.1)$
- $x_i = Z_2 + \varepsilon_i^x, i = 11,..,15, \varepsilon_i^x \sim N(0,0.1)$
- $xi \sim N(0,5), i = 16,..,40$

The response y is generated as :

$$y = X\beta + \epsilon, \varepsilon \sim N(0,5)$$

. Intra-group correlations are high and Inter-groups are average

The grouping effect and selecting others variables



High-dimensional experiments

Exemple	Statistics	LASSO	ENET	VISA	VNET
50 var 100 obs cor 0	$MSE\,\beta$	3.21	3.08	2.77	2.63
	$False-Pos$	14.18	16.81	4.36	4.18
	$False-Neg$	3.11	2.21	3.64	2.9
100 var 50 obs cor 0.5	$MSE\,\beta$	8.39	7.73	10.23	8.18
	$False-Pos$	18.0	25.5	12.62	17.62
	$False-Neg$	3.25	2.12	3.750	3
50 var 100 obs cor 0.95	$MSE\,\beta$	15.79	6.92	15.69	7.04
	$False-Pos$	8.45	33.09	6.36	19.54
	$False-Neg$	4.45	0.27	4.72	1

Table 1 : the simulated examples of four methods based on 100 replications..

bibliography

- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) : Least angle regression. *Annals of Statistics, 32, 407 - 499*.
- Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis, 52, 374-393.*
- Radchenko, P. and James, G. M.(2008) : Variable inclusion and shrinkage algorithms. *Journal of the American statistical association*, vol 103, n 483, 1304-1315.
- Tibshirani, R.(1996) : Regression shrinkage and selection via the Lasso. *journal of the Royal statistical Society, B. 58, 267-288*.
- Zou, H. and Hastie, T. (2005) : Regularization and variable selection via the elasticnet. *Journal of Classification 17 (1), 3-28*.