

Literature Review

J. R. Statist. Soc. B (1996)
58, No. 1, pp. 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

Regression Shrinkage and Selection via the Lasso - jstor

<https://www.jstor.org/stable/2346178> 翻譯這個網頁

由 R Tibshirani 著作 - 1996 - 被引用 31361 次 - 相關文章

PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY

Phil. Trans. R. Soc. A (2009) 367, 4237–4253
doi:10.1098/rsta.2009.0159

INTRODUCTION

Statistical challenges of high-dimensional data

By IAIN M. JOHNSTONE¹ AND D. MICHAEL TITTERINGTON^{2,*}

¹*Department of Statistics, Stanford University, Stanford, CA 94305, USA*

²*Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK*

Open Access Medical Statistics

Dovepress

open access to scientific and medical research

Open Access Full Text Article

REVIEW

High-dimensional data and linear models: a review

This article was published in the following Dove Press journal:
Open Access Medical Statistics
6 August 2014
Number of times this article has been viewed

M Brimacombe

Department of Biostatistics,
University of Kansas Medical Center,
Kansas City, KS, USA

Abstract: The need to understand large database structures is an important issue in biology and medical science. This review paper is aimed at quantitative researchers and statisticians for guidance in modeling large numbers of variables in medical research. It discusses standard linear models and the geometry that underlies their analysis.

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 21, Number 7, May 2016

ISSN 1531-7714

Regularization Methods for Fitting Linear Models with Small Sample Sizes: Fitting the Lasso Estimator using R

W. Holmes Finch, *Ball State University*
Maria E. Hernandez Finch, *Ball State University*

Linear model

The linear model

The standard linear regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_j x_{ji} + \varepsilon_i \quad (1)$$

Where

y_i = Dependent variable value for subject i

x_{ji} = Independent variable j value for subject i

β_0 = Intercept

β_j = Coefficient for independent variable j

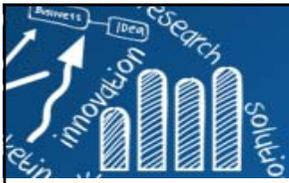
ε_i = Error for subject i

least squares (LS) estimator is typically used.

$$e^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

What happen if $n \ll p$?

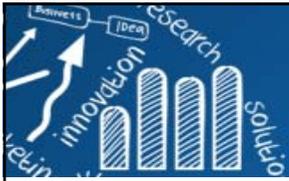
$y \sim x_1 + x_2 + \dots + x_{100}$, if $n=10, 100, 1000$



- when $n \ll p$:
 - parameter estimates: can not converge, high variance.
 - **reduce power**: conclude that one or more of x variables are not related to the y , when in fact they are.
 - **collinearity**, or very strong relationship among x variables, leading to biased parameter estimators.
- not possible to obtain LS estimators.

How to deal with HD data

- **Variable selection methods** (e.g. stepwise regression, best subsets regression)
 - the variable selection methods can produce estimates with inflated standard errors for the coefficients (Hastie, Tibshirani, & Friedman, 2009)
- **Data reduction techniques** (e.g. principal components regression, supervised principal components regression, and partial least squares regression).
 - Data reduction models combine the independent variables into a small number of linear combinations, making **interpretation** of results for individual variables somewhat **more difficult**, and creating an extra layer of complexity in the model as a whole (Finch, Hernandez Finch, & Moss, 2014).



- **Variable selection methods:** assign an inclusion weight of either 1 (include the variable in the model) or 0 (exclude the variable from the model), and then separately estimate the value of β_j for the included variables.
- **Regularization methods** identify optimal values of the β_j such that the most important independent variables receive higher values, and the least important are assigned coefficients at or near 0.

The lasso

2. THE LASSO

2.1. Definition

Suppose that we have data (\mathbf{x}^i, y_i) , $i = 1, 2, \dots, N$, where $\mathbf{x}^i = (x_{i1}, \dots, x_{ip})^T$ are the predictor variables and y_i are the responses. As in the usual regression set-up, we assume either that the observations are independent or that the y_i s are conditionally independent given the x_{ij} s. We assume that the x_{ij} are standardized so that $\sum_i x_{ij}/N = 0$, $\sum_i x_{ij}^2/N = 1$.

Letting $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$, the lasso estimate $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is defined by

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg \min \left\{ \sum_{i=1}^N \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to } \sum_j |\beta_j| \leq t. \quad (1)$$

Here $t \geq 0$ is a tuning parameter. Now, for all t , the solution for α is $\hat{\alpha} = \bar{y}$. We can assume without loss of generality that $\bar{y} = 0$ and hence omit α .

Computation of the solution to equation (1) is a **quadratic programming problem with linear inequality constraints**. We describe some efficient and stable algorithms for this problem in Section 6.

The lasso

$$e^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\hat{\beta}_j| \quad (3)$$

- **Regularization** methods have in common the application of a **penalty** to the LS estimator in regression model.
- Tuning parameter λ : control the amount of shrinkage (i.e. the degree to which the relationship of the independent variables to the dependent variable are down weighted or removed from the model).
 - Larger λ values: greater shrinkage of the model; i.e. a greater reduction in the number of independent variables that are likely to be included in the final model.
 - A λ of 0 leads to the LS estimator.

the optimal λ is the one that minimizes the leave-one-out MSE value calculated using equation (4).

$$MSE_{k\lambda} = \frac{\sum_{i=1}^N (y_{ik} - \hat{y}_{ik\lambda})^2}{N_k} \quad (4)$$

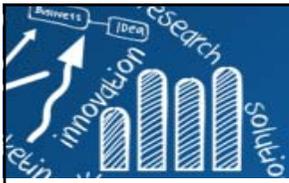
Where

y_i = Dependent variable value for subject i in test set k

$\hat{y}_{ik\lambda}$ = Model predicted dependent variable value for subject i in test set k using λ

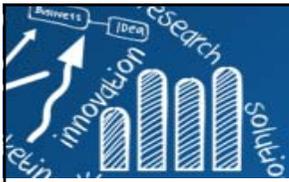
The lasso

- The least squares estimator is known to have **low bias** in many situations, but can also have relatively **large variance**, particularly in the context of high dimensional data; i.e. relatively many predictors and few observations (Loh & Wainwright, 2012).
- The **lasso** has been found to have somewhat **greater bias** than the standard least squares estimator, but with **lower variance**, particularly in the high dimensional case (Hastie, Tibshirani, & Wainwright, 2015).



lasso Approach for Fitting Linear Models^{10/11}

- The data were collected on 10 adults with autism (自閉症) who were clients of an autism research and service provision center at a large Midwestern university.
- Adults identified with autism represent a particularly difficult population from which to sample, meaning that quite frequently **sample sizes are small**.
- **Sample:** 10 adults (9 males), with a mean age of 20 years, 2 months (SD=1 year, 9.6 months).
- **Interest:** the relationship between **executive functioning (16 independent variables)** as measured by the Delis-Kaplan Executive Functioning System (DKEFS; Delis, Kaplan, & Kramer, 2001) and the **full scale intelligence score (FSIQ) (dependent variable)** on the Wechsler Adult Intelligence Scale, 4th edition (WAIS-IV; Wechsler, 2008).



Example

Independent Variables

Variable

Visual scanning
Number sequencing
Letter sequencing
Number-letter sequencing
Motor speed
Letter fluency
Category fluency
Category switching
Category switching accuracy
Filled dots
Empty dots
Dots switching
Color naming
Word reading
Inhibition
Inhibition/switching

Appendix

R code and output for fitting the lasso and elastic net models for example data

Example data file

Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
3.741530	0.573051	-0.175230	-1.339954	-0.368095	1.090042	-0.115272	-0.577052	0.425472	0.179867	1.088520
0.100441	1.183853	-0.694153	-0.766538	0.455033	-0.017487	-1.367410	-0.050084	-0.817974	-1.559255	0.579605
2.553595	1.007614	1.543381	0.463916	-0.898300	-0.053513	1.533398	0.180512	0.113829	-0.096545	-0.352276
0.153503	-1.341994	-1.445909	1.730850	1.027419	0.677408	-0.001175	-0.138712	-0.759287	-0.447889	0.483444
1.465349	1.104495	-0.507631	-0.517296	0.242078	0.761720	-1.901134	-2.223851	-0.736562	2.318569	-2.272791
-0.012732	0.111837	-0.846025	0.155868	-0.897112	-1.184396	-0.295120	0.881524	0.966334	-1.903001	1.055233
0.726743	-0.320148	0.297111	0.508650	0.206923	-0.527616	-0.030750	-0.805411	0.766234	0.496932	-0.120334
2.170335	0.068812	1.809094	-0.761952	-2.154671	-0.286850	-0.860617	-0.102291	2.345841	0.284032	0.253651
-1.669843	-1.172163	-1.161900	0.935259	0.858773	-0.271187	-1.231314	-0.238721	-1.086486	0.989511	2.269332
1.009178	1.741643	1.454726	-2.975978	2.920440	-0.798064	0.156104	1.350790	-1.084402	-0.943684	-0.180285

```
#Read the data from a .dat file, print the data to be sure that it# #was read in
correctly, and create matrices of the independent and#
#dependent variables.#
demo<-read.table("c:/research/lasso demonstration/demo.dat", header=F)
demo
demo.iv<-as.matrix(demo[,2:11])
demo.dv<-as.matrix(demo[,1])
#Cross-validation to determine optimal value of lambda#
demo.lasso.cv<-cv.glmnet(demo.iv, demo.dv, type.measure="mse", nfolds=10)
plot(demo.lasso.cv)
```