Variable Selection in Regression using Maximal Correlation and Distance Correlation

Deniz Yenigün [1] Maria Rizzo [2]

[1]Istanbul Bilgi University, Department of Industrial Engineering

[2]Bowling Green State University, Department of Mathematics and Statistics

13 October 2017, Sabancı University

Table of Contents

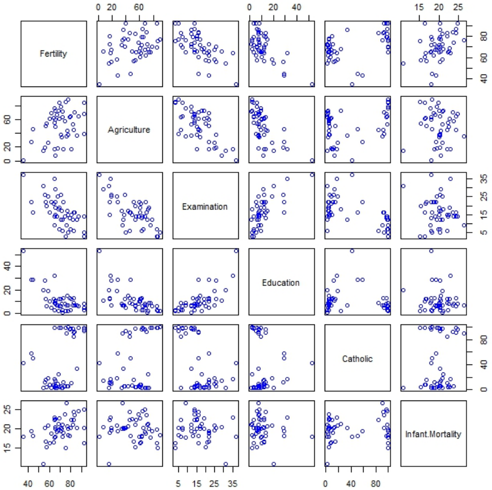
Variable Selection

- Recent improvements in data collection technologies give rise to complex regression problems where the number of candidate predictor variables explaining the response variable may be very large.
- In most of these regression problems the main task is to select the most influential predictors explaining the response, and removing the others from the model.
- These problems are usually referred to as variable selection problems in the statistical literature.

Example: Swiss Fertility Data

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

- Y - Common standardized fertility measure (Fertility)
- X_1 - Percentage of males involved in agriculture as occupation (Agriculture)
- X_2 - Percentage of draftees receiving highest mark on army examination (Examination)
- X_3 - Percentage of education beyond primary school for draftees (Education)
- X_4 - Percentage of Catholic (Catholic)
- X_5 - Live births who live less than 1 year (Infant Mortality)


Fertility
Agriculture
Examination
Education
Catholic
Infant.Mortality

Subset Selection

Consider the linear regression model

$$Y = X\beta + \epsilon, \tag{1}$$

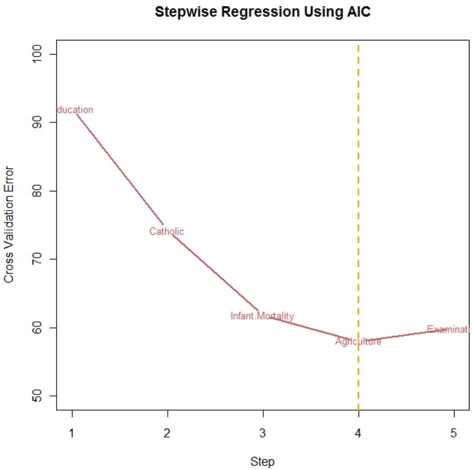
where Y is a vector of length n representing the response variable, X is an n by p matrix representing the predictor variables, β is a vector of length p containing regression coefficients, and ϵ is a vector of length n containing independent normal noise terms.

The essential goal in variable selection is to divide X into the set of active terms X_A and the set of inactive terms X_I.

Issues:

- Comparison Criterion for two candidates of X_A.
  - Akaike Information Criterion: $AIC = n \log(RSS/n) + 2p$
  - Bayesian Information Criterion: $BIC = n \log(RSS/n) + p \log n$
  - Computationally Intensive Comparison Criteria: k-Fold Cross-Validation, etc.
- Computational Method. If there are p candidate predictors, there are $2^p - 1$ possible candidates for X_A. Ex: When $p = 20$ $\rightarrow$ 1,048,575 possible models to check.
  - Stepwise Methods (Forward and Backward).
  - Branch-and Bounds, Leaps-and-Bounds.
  - Stagewise Methods.

Stepwise AIC Example: Swiss Fertility Data



Shrinkage Methods

The discrete nature of subset selection methods may lead to high variance in some situations.

Due to their continuous nature, *shrinkage methods* may provide an alternative to the subset selection methods.

- Ridge Regression (Hoerl and Kennard, 1970a,b)
- Lasso (Tibshirani, 1996)
- LARS (Efron *et. al.*, 2004)

Lasso

Tibshirani (1996) proposed *lasso*, which minimizes the residual sum of squares

$$\|Y - X\beta\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq \theta. \tag{2}$$

Here $\theta \geq 0$ is a tuning parameter that shrinks the coefficients. When θ is large enough, this becomes the least squares method. The shrinkage reduces some of the coefficients to zero and yields a natural variable selection.

Rényi (1959) Postulates for Measures of Dependence

A) $\delta(X, Y)$ is defined for every pair X, Y neither of which is constant with probability 1.
B) $\delta(X, Y) = \delta(Y, X)$.
C) $0 \leq \delta(X, Y) \leq 1$.
D) $\delta(X, Y) = 0$ if and only if X and Y are independent.
E) $\delta(X, Y) = 1$ if either $X = g(Y)$ or $Y = f(X)$, where $g(\cdot)$ and $f(\cdot)$ are Borel-measurable functions.
F) If the Borel-measurable functions $g(\cdot)$ and $f(\cdot)$ map the real axis in a one-to-one way to itself, then $\delta(f(X), g(Y)) = \delta(X, Y)$.
G) If the joint distribution of X and Y is normal, then $\delta(X, Y) = |R(X, Y)|$, where $R(X, Y)$ is the correlation coefficient of X and Y.

Maximal Correlation

The maximal correlation S between two random variables (X, Y) is defined as

$$S(X, Y) = \sup_{f,g} \rho(f(X), g(Y)),$$

where ρ denotes the classical correlation coefficient, and the supremum is taken over all functions of X and Y with finite and positive non-zero variance.

Maximal Correlation satisfies all 7 postulates listed by Rényi.

Product Moment Correlation satisfies B, C, and G only.

Gebelein (1941)
Rényi (1959)
Csáki and Fisher (1963)
Breiman and Friedman (1985)
Koyak (1987)
Sethuraman (1990)
Dembo et. al. (2001)
Bryc et. al. (2005)
Yenigun et. al. (2011)

Distance Correlation

Consider random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$. The characteristic functions of X and Y are denoted by f_X and f_Y, respectively, and the joint characteristic function of X and Y is $f_{X,Y}$.

The distance covariance between X and Y is

$$V^2(X, Y) = \|f_{X,Y}(t, s) - f_X(t)f_Y(s)\|^2. \tag{3}$$

See Szekely, Rizzo, Bakirov (2007) for the norm $\| \cdot \|$.

Similarly, the distance variance of X is

$$V^2(X) = \|f_{X,X}(t,s) - f_X(t)f_X(s)\|^2, \tag{4}$$

and the distance correlation between X and Y is

$$R^2(X,Y) = \begin{cases} \frac{V^2(X,Y)}{\sqrt{V^2(X)V^2(Y)}}, & V^2(X)V^2(Y) > 0 \\ 0, & V^2(X)V^2(Y) = 0 \end{cases}. \tag{5}$$

Distance correlation satisfies the Rényi postulates A, B, C, D. The rest is partly satisfied.

Proposed Methods

We propose two model selection methods based on the dependence measures distance correlation and maximal correlation.

- Stepwise regression using distance correlation
- Stepwise regression using maximal correlation

We begin with defining partial distance (/maximal) correlation.

Partial Distance (/Maximal) Correlation

Consider random variables X, Y, and a possibly vector valued random variable Z.

Given Z, the partial distance (/maximal) correlation between X and Y is computed as follows:

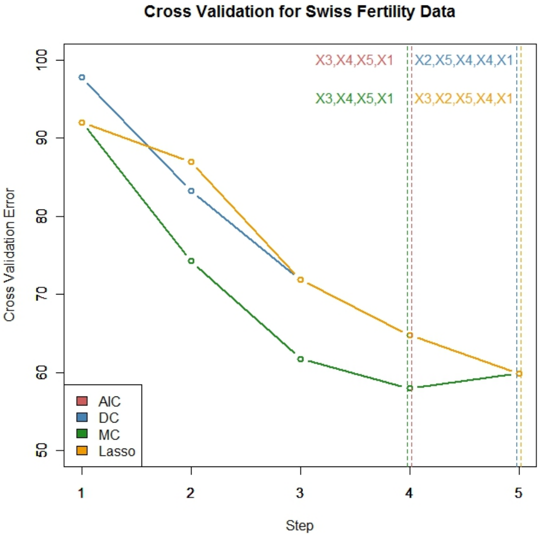
- Regress X on Z, denote the error terms by R_X.
- Regress Y on Z, denote the error terms by R_Y.
- The distance (/maximal) correlation between R_X and R_Y is the partial distance correlation between X and Y, given Z.

Stepwise Regression Using Distance (/Maximal) Correlation

Then we can define a stepwise regression procedure, using distance (/maximal) correlation as follows:

1. Consider all candidate predictor variables individually and find the one which has the largest distance (/maximal) correlation with the dependent variable.
2. For the remaining steps, add one more term such that the partial distance (/maximal) correlation with the dependent variable, given the previously entered variable(s), is largest.
3. Stop when all terms have entered the model. The step with the smallest cross-validation error is the selected model.

Illustration on Swiss Fertility Data



Simulation Study

We consider 6 cases.

- Case 1: Linear Relations
- Case 2: Non-Linear Relations
- Case 3: Dependent but Uncorrelated Variables
- Case 4: Constant Collinearity Among Predictors
- Case 5: Toeplitz Collinearity Among Predictors
- Case 6: A Generalized Linear Model: Gamma Regression

For each case we considered $N = 100$ samples of size $n = 100$.

Case 1: Linear Relations

We consider a total of $p = 8$ candidate predictors having independent standard normal distributions, $q = 3$ of which are related with the dependent variable via:

$$Y = X\beta + \epsilon,$$

where $\beta = [1, 1, 1, 0, 0, 0, 0, 0]$ and $\epsilon \sim N(0, \sigma = 2)$.

Case 2: Non-Linear Relations

We consider a total of $p = 8$ candidate predictors from the following distributions: $X_1 \sim N(0, 1)$, $X_2 \sim N(0, 2)$, $X_3 \sim U(-1.5, 1.5)$, $X_4, ..., X_8 \sim U(-1, 1)$. The first $q = 4$ are related with the dependent variable via:

$$Y = \log[4 + \sin(3X_1) + \sin(X_2) + X_3^2 + X_4 + 0.1\epsilon],$$

where $\epsilon \sim N(0, \sigma = 1)$.

Case 3: Dependent but Uncorrelated Variables

We consider a total of $p = 8$ candidate predictors from the following distributions: $X_1 \sim N(0, 1.4)$, $X_2 \sim U(-1.7, 1.7)$, $X_3 \sim N(0, 0.8)$, $X_4, ..., X_8 \sim N(0, 1)$. Let us define $Y_1, ..., Y_3$ as follows:

$$Y_1 = |X_1|, \quad Y_2 = X_2^2, \quad Y_3 = X_3^2.$$

It can be shown that the pairs (X_i, Y_i), $i = 1, 2, 3$, are uncorrelated. We define the dependent variable as

$$Y = |X_1| + X_2^2 + X_3^2.$$

Case 4: Constant Collinearity Among Predictors

We consider a total of $p = 8$ candidate predictors from a multivariate normal distribution, $\mathbf{X} \sim N_P(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \cdots & \theta \\ \vdots & \vdots & \ddots & \vdots \\ \theta & \theta & \cdots & 1 \end{bmatrix}.$$

We set $\theta = 0.6$. The first $q = 3$ of these variables are related with the dependent variable via:

$$Y = X\beta + \epsilon,$$

where $\beta = [1, 1, 1, 0, 0, 0, 0, 0]$ and $\epsilon \sim N(0, \sigma = 2)$.

Case 5: Toeplitz Type Collinearity Among Predictors

This is the same as Case 4, but

$$\Sigma = \begin{bmatrix} 1 & \theta & \theta^2 & \cdots & \theta^{p-1} \\ \theta & 1 & \theta & \cdots & \theta^{p-2} \\ \theta^2 & \theta & 1 & \cdots & \theta^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta^{p-1} & \theta^{p-2} & \theta^{p-3} & \cdots & 1 \end{bmatrix}.$$

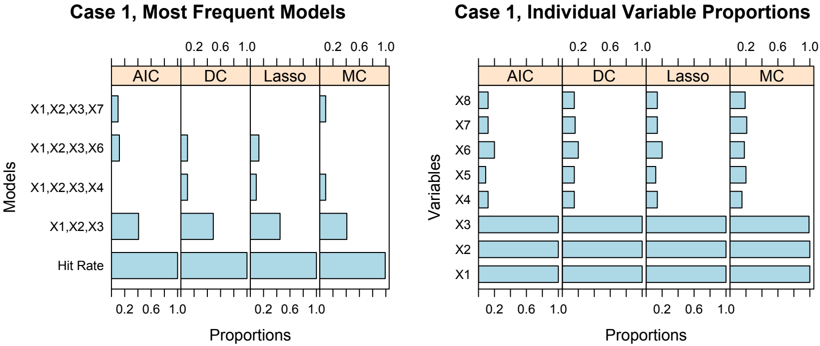
Case 6: A Generalized Linear Model (Gamma Regression)

We consider $p = 8$ candidate predictors following standard normal distribution, $q = 3$ of which are related with the response via:

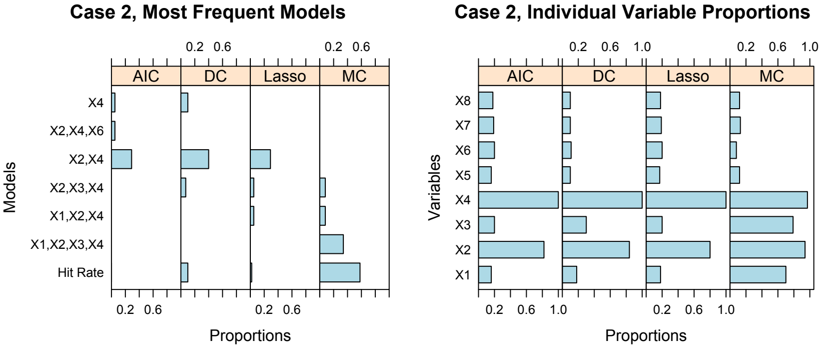
$$L = X\beta,$$

with $\beta = [0.25, 0.25, 0.25, 0, 0, 0, 0, 0]$. The link function is the log function, thus the mean vector of the responses are $\hat{\mu} = e^L$. Responses are generated from gamma distribution with mean $\hat{\mu}$ and unit variance

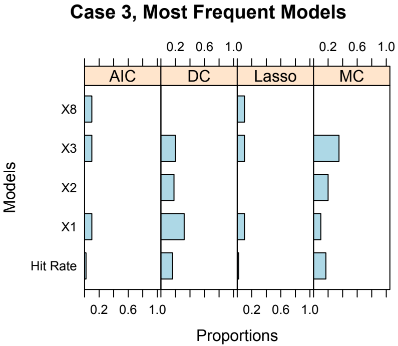
Case 1: Linear Relations

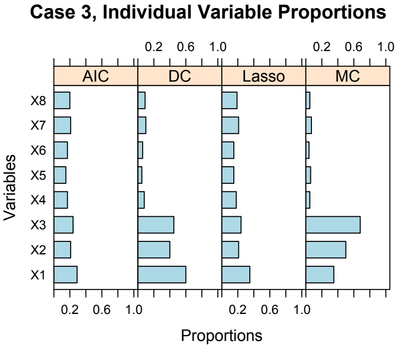


Case 2: Non-Linear Relations

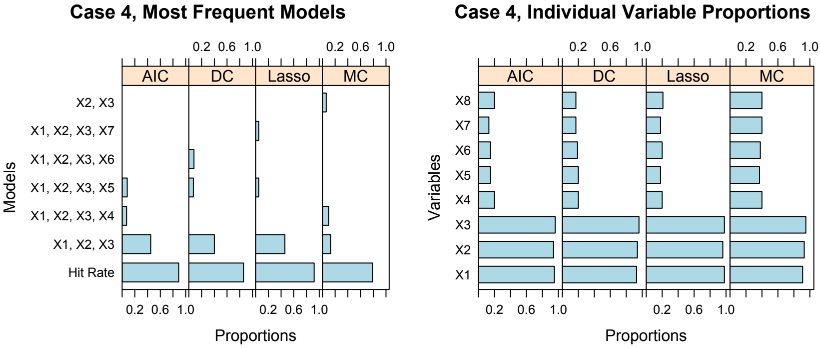


Case 3: Dependent but Uncorrelated Variables

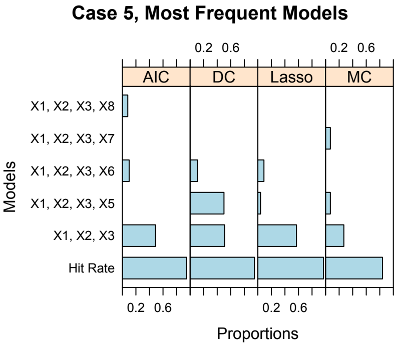


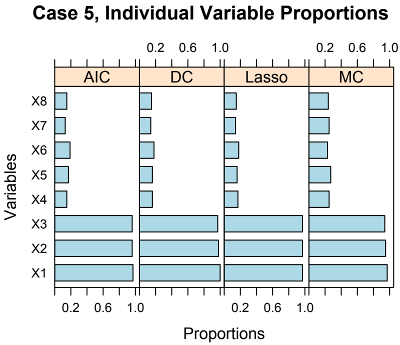


Case 4: Constant Collinearity Among Predictors

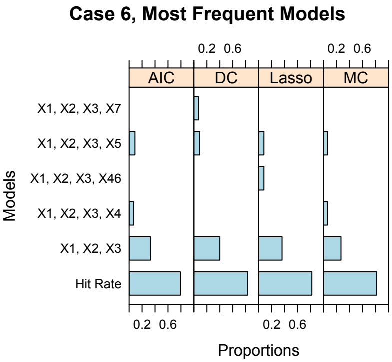


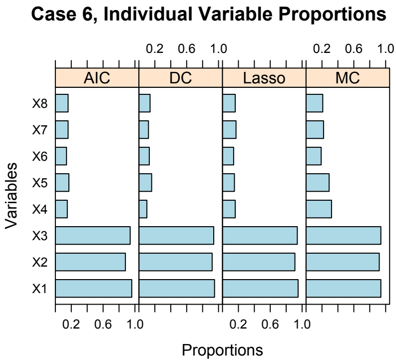
Case 5: Toeplitz Type Collinearity Among Predictors





Case 6: A Generalized Linear Model (Gamma Regression)





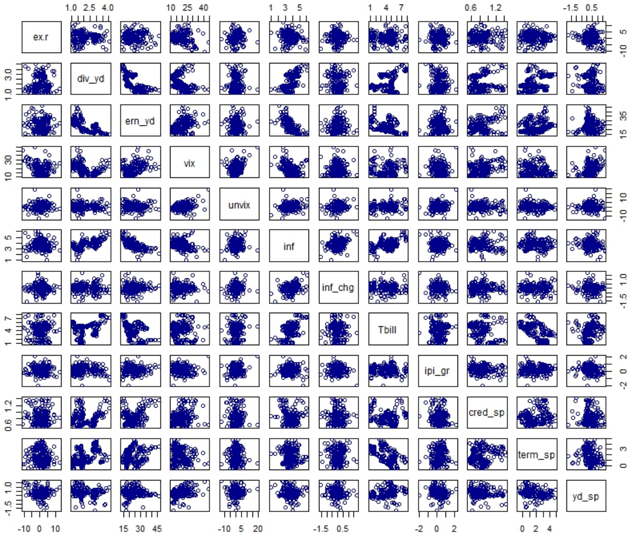
Application: S&P 500 Monthly Returns Data

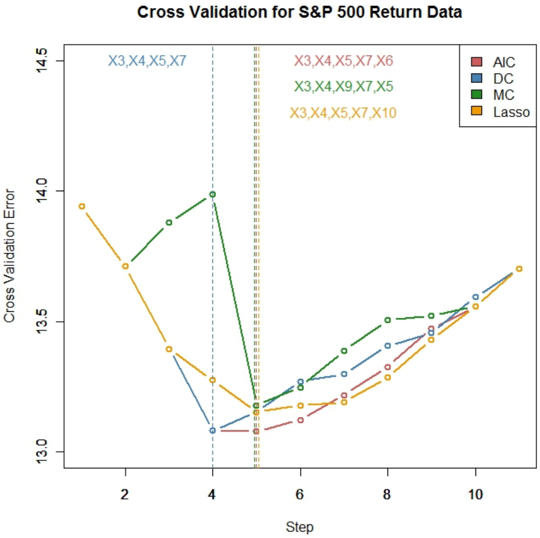
S&P 500 is an index portfolio defined by Standard & Poor's rating agency.

Monthly returns of S&P 500 index and the values of 11 candidate predictors between January 1989 and December 2007 (n=216) were analyzed using the four methods discussed above.

- Stepwise AIC
- Stepwise DC
- Stepwise MC
- Lasso

- Y - Monthly expected return of S&P 500 index (ex.r)
- X_1 - Dividend yield (div_yd)
- X_2 - Earnings yield (ern_yd)
- X_3 - Volatility index (vix)
- X_4 - Unexpected volatility (unvix)
- X_5 - Inflation rate (inf)
- X_6 - Change in inflation rate (inf_chg)
- X_7 - 90-day treasury bill (Tbill)
- X_8 - Industrial production index growth (ipi_gr)
- X_9 - Credit spread (cred_sp)
- X_{10} - Term spread (term_sp)
- X_{11} - Yield spread (yd_sp)


ex.r
div_yd
ern_yd
vix
unvix
inf
inf_chg
Tbill
ipi_gr
cred_sp
term_sp
yd_sp


Cross Validation for S&P 500 Return Data
X3,X4,X5,X7
X3,X4,X5,X7,X6
X3,X4,X9,X7,X5
X3,X4,X5,X7,X10
AIC
DC
MC
Lasso
Cross Validation Error
13.0
13.5
14.0
14.5
2
4
6
8
10
Step

Conclusions

- Maximal Correlation and Distance Correlation were employed as comparison criteria in stepwise regression
- The methods are easy to implement
- The performances of the methods are comparable with commonly used methods
- In the presence of nonlinear or uncorrelated dependencies, our methods may be favorable

Selected References

Breiman, L., Friedman, J., 1985. Estimating optimal transformations for multiple regression and correlation (with discussion). J. Amer. Statist. Assoc., 80, 580-619.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression, The Annals of Statistics, 32, 407-499.

Miller, A., 2002. Subset Selection in Regression, CRC Press.

Szekely, G.J., Rizzo, M.L., Bakirov, N.K., 2007. Measuring and testing dependence by correlation of distances, The Annals of Statistics, 35, 2769-2794.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso, J. R. Statist. Soc. B, 58, 267-288.

Yenigün, C.D., Szekely, G.J., Rizzo, M.L., 2011. A test of independence in two way contingency tables based on maximal correlation, Communications in Statistics: Theory and Methods, 40, 2225-2242.

Yenigün, C.D., Rizzo, M.L., 2015. Variable selection in regression using maximal correlation and distance correlation, Journal of Statistical Computation and Simulation, 85, 1692-1705.