Lecture 17: Regularization/Shrinkage Methods- Elastic Net Causal Inference

Big Data and Machine Learning for Applied Economics Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

October 13, 2020

Agenda

1. Recap
2. More predictors than observations ($k > n$)
    - OLS with more predictors than observations
    - Lasso and Ridge with $k > n$
3. Elastic Net
4. Lasso for Causality
    - Application
5. Review & Next Steps
6. Further Readings

Recap: Regularization

- For $\lambda \geq 0$ given, consider minimizing the following objective function
- Lasso:

$$min_\beta L(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s| \tag{1}$$

- Ridge:

$$min_\beta R(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} (\beta_s)^2 \tag{2}$$

Recap: Regularization Demo

```r
#Load the required packages
library("dplyr") #for data wrangling
library("caret") #ML

data(swiss) #loads the data set

set.seed(123) #set the seed for replication purposes
str(swiss) #conmpact display
```

```
## ’data.frame’:    47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

Recap: Regularization Demo

```r
ols <- train(Fertility ~ .,    # model to fit
                     data = swiss,
                     trControl = trainControl(method = "cv", number = 10),     # Method: crossvalidation,
                     method = "lm")                       # specifying regression model
ols
```

```
## Linear Regression
##
## 47 samples
##  5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 42, 42, 42, 42, 42, 44, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   7.424916  0.6922072  6.31218
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Recap: Regularization Demo

```r
lambda <- 10^seq(-2, 3, length = 100)
lasso <- train(
  Fertility ~., data = swiss, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneGrid = expand.grid(alpha = 1, lambda=lambda), preProcess = c("center", "scale")
  )

lasso
```

```
## glmnet
##
## 47 samples
##  5 predictor
##
## Pre-processing: centered (5), scaled (5)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 43, 43, 43, 42, 42, 41, ...
## Resampling results across tuning parameters:
##
##  ...
##
## Tuning parameter ’alpha’ was held constant at a value of 1
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 1 and lambda = 0.02009233.
```

Recap: Regularization Demo

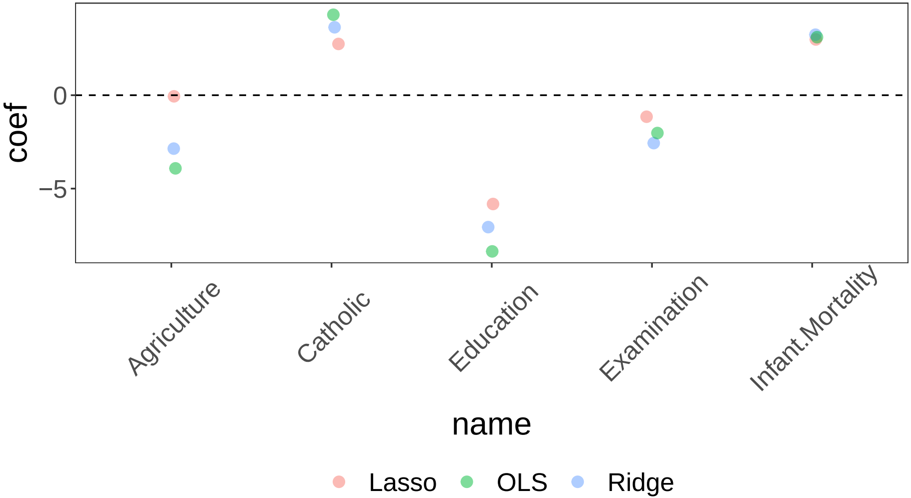
```
ridge <- train(
  Fertility ~., data = swiss, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneGrid = expand.grid(alpha = 0,lambda=lambda), preProcess = c("center", "scale")
  )
ridge
```

```
## glmnet
##
## 47 samples
##  5 predictor
##
## Pre-processing: centered (5), scaled (5)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 42, 42, 43, 44, 42, 42, ...
## Resampling results across tuning parameters:
##
##  ...
##
## Tuning parameter ’alpha’ was held constant at a value of 0
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 0 and lambda = 0.7390722.
```

Recap: Regularization Demo

```
##
## Call:
## summary.resamples(object = ., metric = "RMSE")
##
## Models: ridge, lasso
## Number of resamples: 10
##
## RMSE
##           Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA’s
## ridge 2.615430 4.674108 7.627190 6.923531 8.939798 10.55026    0
## lasso 3.205868 5.553161 5.961622 7.324069 8.587818 13.46074    0
```

Recap: Regularization Demo



More predictors than observations ($k > n$)

- Objective 1: Accuracy
  - Minimize prediction error (in one step) $\rightarrow$ Ridge, Lasso
- Objective 2: Dimensionality
  - Reduce the predictor space $\rightarrow$ Lasso's free lunch
- What happens when we have more predictors than observations ($k > n$)?
  - OLS fails
  - Ridge and Lasso to the rescue?

OLS when $k > n$

- Rank? Max number of rows or columns that are linearly independent
  - Implies $rank(X_{k\times n}) \leq min(k,n)$
- MCO we need $rank(X_{k\times n}) = k \implies k \leq n$
- If $rank(X_{k\times n}) = k$ then $rank(X'X) = k$
- If $k > n$, then $rank(X'X) \leq n < k$ then $(X'X)$ cannot be inverted
- Ridge and Lasso work when $k \geq n$

Ridge when $k > n$

$$min_{\beta} R(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} (\beta_s)^2 \tag{3}$$

- Solution → data augmentation
- Problem set 2 HW
- Intuition: Ridge "adds" k additional points.
- Allows us to "deal" with $k \geq n$

Lasso when $k > n$

- Lasso works fine in this case
- However, there are some issues to keep in mind
  - When $k > n$ chooses at most n variables
  - When we have a group of highly correlated variables,
    - Lasso chooses only one. Makes it unstable for prediction. (Doesn't happen to Ridge)
    - Ridge shrinks the coefficients of correlated variables toward each other. This makes Ridge "work" better than Lasso. "Work" in terms of prediction error

Naive Elastic Net

- Elastic net: happy medium.
  - Good job at prediction and selecting variables

$$min_{\beta}NEL(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda_1\sum_{s=2}^{p}|\beta_s| + \lambda_2\sum_{s=2}^{p}\beta_s^2 \tag{4}$$

- Mixes Ridge and Lasso
- Lasso selects predictors
- Strict convexity part of the penalty (ridge) solves the grouping instability problem
- H.W.: $\beta_{OLS} > 0$ one predictor standarized

$$\hat{\beta}_{naive\,EN} = \frac{\left(\hat{\beta}_{OLS} - \frac{\lambda_1}{2}\right)_+}{1+\lambda_2} \tag{5}$$

Elastic Net

- Elastic Net: reescaled version
- Double Shrinkage introduces "too" much bias, *final* version "corrects" for this

$$\hat{\beta}_{EN} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}_{naive\,EN} \tag{6}$$

- Careful sometimes software asks.
- How to choose (λ_1, λ_2)? $\rightarrow$ Bidimensional Crossvalidation
- Zou, H. & Hastie, T. (2005)

Model Selection When the Goal is Causal Inference

Motivation

- Up to this point we only cared about prediction
- Can we use some of these models to do causal inference?
- We are going to see how we can adapt regularization (lasso) to do inference

Model Selection When the Goal is Causal Inference

Let's start with the following model

$$y_i = \alpha d_i + g(w_i) + \zeta_i \tag{7}$$

were

- d_i is the treatment/policy variable of interest,
- w_i is a set controls
- $E[\theta_i | d_i, w_i] = 0$

Model Selection When the Goal is Causal Inference

- Traditional approach: researcher selects w_i, problem is that mistakes can occur.
- Same if uses an "automatic" model selection approach. It can leave out potentially important variables with small coefficients but non zero coefficients out
- The omission of such variables then generally contaminates estimation and inference results based on the selected set of variables. (e.g. OVB)
- The validity of this approach is delicate because it relies on perfect model selection.
- Because model selection mistakes seem inevitable in realistic settings, it is important to develop inference procedures that are robust to such mistakes.
- Solution here: Use Lasso

Model Selection When the Goal is Causal Inference

- Using Lasso is useful for prediction
- However, naively using Lasso e to draw inferences about model parameters can be problematic.
- Part of the difficulty is that these procedures are designed for prediction, not for inference
- Leeb and Pötscher 2008 show that methods that tend to do a good job at prediction can lead to incorrect conclusions when inference is the main objective
- This observation suggests that more desirable inference properties may be obtained if one focuses on model selection over the predictive parts of the economic problem
  - The reduced forms and first-stages—rather than using model selection in the structural model directly.

Approximate sparse models

- Suppose we are interested in the following model

$$y_i = g(w_i) + \zeta_i \tag{8}$$

with

  - $E(\zeta_i | g(x_i)) = 0$
  - $i = 1, \ldots, n$ are iid
  - To avoid over-fitting and produce good out of sample forecast we will need to restrict or regularize $g(.)$
  - The focus here is on regularization that treats $g(w_i)$ as a high-dimensional, approximately linear model:

$$g(w_i) = \sum_{j=1}^{p} \beta_j x_{ij} + r_{pi} \tag{9}$$

  - where $p >> n$ and r_{pi} is small enough

Approximate sparse models

$$g(w_i) = \sum_{j=1}^{p} \beta_j x_{ij} + r_{pi} \tag{10}$$

- Approximate sparsity of this high-dimensional linear model imposes the restriction that linear combinations of only $s << n$ x_{ij} variables provide a good approximation to $g(w_i)$
- A bonus is that the identity of this s x_{ij} variables are a priori unknown
- And that we can have a nonzero approximation error r_{pi}
- We are going to try to learn the identities of these variables while estimating the coefficients.

Approximate sparse models

- We can use Lasso that is slightly modified

$$L(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|\gamma_s \tag{11}$$

  - where $\lambda > 0$ is the penalty level chosen using Belloni, Chen, Chernozhukov, and Hansen (2012)
  - γ_j are *penalty loadings*
  - *penalty loadings* are chosen to insure equivariance of coefficient estimates to rescaling of x_{ij} and can also be chosen to address heteroskedasticity, clustering, and non-gaussian errors

Inference with Selection among Many Controls

- Consider a linear model where a treatment variable, d_i, is taken as exogenous after conditioning on control variables

$$y_i = \alpha d_i + x_i'\theta_y + r_{yi} + \zeta_i \tag{12}$$

- where $E[\zeta_i | d_i, x_i, r_{yi}] = 0$
- x_i is a p-dimensional vector with $p >> n$
- r_{yi} is an approximation error
- the parameter of interest is α

Inference with Selection among Many Controls

- Naive approach (doesn't work)

$$y_i = \alpha d_i + x_i' \theta_y + r_{yi} + \zeta_i \tag{13}$$

  - Select control variables by applying Lasso, forcing the treatment variable to remain in the model
  - One could then try to estimate and do inference about α by applying ordinary least squares with y_i as the outcome, and d_i and any selected control variables as regressors.
  - The problem is that it target prediction → any variable that is highly correlated to the treatment variable will tend to be dropped
  - Of course, the exclusion of a variable that is highly correlated to the treatment will lead to substantial omitted-variables bias

Inference with Selection among Many Controls

- There are problems with the above naive approach.
- It ignores a key component to understanding omitted-variables bias, the relationship between the treatment variable and the controls.
  - To aid in learning about this relationship, we introduce an additional "reduced form" relation between the treatment and controls:

$$d_i = x_i'\theta_d + r_{di} + v_i \tag{14}$$

where $E[v_i|x_i, r_{di}] = 0$

Inference with Selection among Many Controls

- The naive approach is based on a "structural" model where the target is to learn the treatment effect given controls, not an equation representing a prediction rule for y_i given d_i and x_i.
- It is thus useful to transform the first equation of this section to a reduced form, predictive equation by substituting the equation introduced for d_i into the "structural" equation yielding the reduced form system:

$$\begin{aligned} y_i &= x_i'(\alpha\theta_d + \theta_y) + (\alpha r_{di} + r_{yi}) + r_{di} + (\alpha v_i + \zeta_i) = x_i'\pi + r_{ci} + \epsilon_i \\ d_i &= x_i'\theta_d + r_{di} + v_i \end{aligned} \tag{15}$$

- where $E(\epsilon_i | x_i, r_{ci}] = 0$
- r_{ci} is a composite approximation error
- Both of these equations represent predictive relationships, which may be estimated using high-dimensional methods.

Inference with Selection among Many Controls

- To prevent model selection mistakes, it is important to consider both equations for selection:

- We apply variable selection methods to each of the two reduced form equations and then use all of the selected controls in estimation of α.

- We select
  1. A set of variables that are useful for predicting y_i, say x_{yi}, and
  2. A set of variables that are useful for predicting d_i, say x_{di}.
- We then estimate α by ordinary least squares regression of y_i on d_i and the union of the variables selected for predicting y_i and d_i, contained in x_{yi} and x_{di}.
- We thus make sure we use variables that are important for either of the two predictive relationships to guard against OVB

Inference with Selection among Many Controls

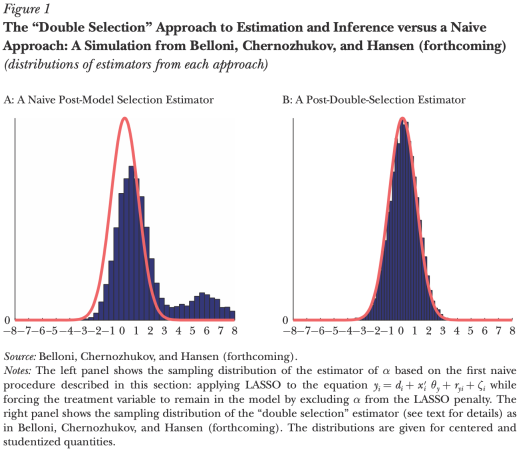


Figure 1

The "Double Selection" Approach to Estimation and Inference versus a Naive Approach: A Simulation from Belloni, Chernozhukov, and Hansen (forthcoming)

(distributions of estimators from each approach)

A: A Naive Post-Model Selection Estimator

B: A Post-Double-Selection Estimator

Source: Belloni, Chernozhukov, and Hansen (forthcoming).

Notes: The left panel shows the sampling distribution of the estimator of α based on the first naive procedure described in this section: applying LASSO to the equation $y_i = d_i + x_i' \theta_y + r_{yi} + \zeta_i$ while forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. The right panel shows the sampling distribution of the "double selection" estimator (see text for details) as in Belloni, Chernozhukov, and Hansen (forthcoming). The distributions are given for centered and studentized quantities.

Application: Estimation of the treatment effect in a linear model with many confounding factors

- What is the effect of an initial (lagged) level of GDP per capita on the growth rates of GDP per capita?
- Solow-Swan-Ramsey growth model predicts convergence
- Poorer countries should typically grow faster and therefore should tend to catch up with the richer countries, conditional on a set of institutional and societal characteristics.
- Covariates that describe such characteristics include variables measuring education and science policies, strength of market institutions, trade openness, savings rates and others.

Application: Estimation of the treatment effect in a linear model with many confounding factors

Thus, we are interested in a specification of the form:

$$y_i = \alpha d_i + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{16}$$

where

- y_i is the growth rate of GDP over a specified decade in country i,
- d_i is the log of the initial level of GDP at the beginning of the specified period,
- x_{ij}'s form a long list of country i's characteristics at the beginning of the specified period.
- We are interested in testing the hypothesis of convergence, $\alpha_1 < 0$.

Application: Estimation of the treatment effect in a linear model with many confounding factors

For this exercise we use the Barro and Lee (1994) data

```
require("hdm") #package
data(GrowthData) #load data
dim(GrowthData)
```

```
## [1] 90 63
```

The number of covariates p is large relative to the sample size n

```
y = GrowthData[,1,drop=F]
d = GrowthData[,3, drop=F]
X = as.matrix(GrowthData)[,-c(1,2,3)]
varnames = colnames(GrowthData)
```

Application: Estimation of the treatment effect in a linear model with many confounding factors

- Now we can estimate the effect of the initial GDP level.
- First, we estimate by OLS:

```
xnames= varnames[-c(1,2,3)] # names of X variables
dandxnames= varnames[-c(1,2)] # names of D and X variables

# create formulas by pasting names (this saves typing times)
fmla= as.formula(paste("Outcome ~ ", paste(dandxnames, collapse= "+")))

# Estimate using OLS
ls.effect= lm(fmla, data=GrowthData)
```

Application: Estimation of the treatment effect in a linear model with many confounding factors

Second, we estimate the effect by the partialling out by Post-Lasso:

```
dX = as.matrix(cbind(d,X))
lasso.effect = rlassoEffect(x=X, y=y, d=d, method="partialling out")
summary(lasso.effect)
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##       Estimate. Std. Error t value Pr(>|t|)
## [1,]  -0.04981     0.01394  -3.574 0.000351 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Application: Estimation of the treatment effect in a linear model with many confounding factors

Third, we estimate the effect by the double selection method:

```
dX = as.matrix(cbind(d,X))
doublesel.effect = rlassoEffect(x=X, y=y, d=d, method="double selection")
summary(doublesel.effect)
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##          Estimate. Std. Error t value Pr(>|t|)
## gdpsh465  -0.05001    0.01579  -3.167  0.00154 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Application: Estimation of the treatment effect in a linear model with many confounding factors

- Collecting the results

	Estimate	Std. Error
full reg via ols	-0.01	0.02989
partial reg via post-lasso	-0.05	0.01394
partial reg via double selection	-0.05	0.01579

Review & Next Steps

- Today:
    - More predictors than observations $(k > n)$
        - OLS doesn't work
        - Lasso and Ridge work with issues
    - Elastic Net
    - Lasso for Causality: Post Lasso Double Selection
- Next class: Classification
- Questions? Questions about software?

Further Readings

- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2), 608-650.

- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2), 29-50.

- Chernozhukov, V., Hansen, C., & Spindler, M (2016). hdm: High-Dimensional Metrics R Journal, 8(2), 185-199. `https://journal.r-project.org/archive/2016/RJ-2016-040/index.html`

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

- Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal StatisticalSociety, Series B.67: pp. 301–320