ML in Aid of Estimation: Part I

Lasso and ATE

Itamar Caspi

April 29, 2019 (updated: 2019-05-02)

Replicating this presentation

Use the pacman package to install and load packages:

```r
if (!require("pacman")) install.packages("pacman")

pacman::p_load(tidyverse, tidymodels, hdm, ggdag, knitr, xaringan, RefManageR)
```

Outline

1. Causal Inference and Treatment Effects
2. Variable Selection Using the Lasso
3. Lasso In Aid of Causal Inference
4. Empirical Illustration using `hdm`

Causal Inference And Treatment Effects

The road not taken

Source: https://mru.org/courses/mastering-econometrics/ceteris-paribus

Treatment and potential outcomes

- Treatment

$$D_i = \begin{cases} 1, & \text{if unit } i \text{ received the treatment} \\ 0, & \text{otherwise.} \end{cases}$$

- Treatment and potential outcomes

$$\begin{aligned} Y_{i0} &\quad \text{is the potential outcome for unit } i \text{ with } D_i = 0 \\ Y_{i1} &\quad \text{is the potential outcome for unit } i \text{ with } D_i = 1 \end{aligned}$$

- Observed outcome: Under the Stable Unit Treatment Value Assumption (SUTVA), The realization of unit i's outcome is

$$Y_i = Y_{1i} D_i + Y_{0i}(1 - D_i)$$

- Fundamental problem of causal inference: We cannot observe *both* Y_{1i} and Y_{0i}.

Treatment effect and observed outcomes

- Individual treatment effect: The difference between unit i's potential outcomes:

$$\tau_i = Y_{1i} - Y_{0i}$$

- *Average treatment effect* (ATE)

$$\mathbb{E}[\tau_i] = \mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$$

- *Average treatment effect for the treatment group* (ATT)

$$\mathbb{E}[\tau_i | D_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1] = \mathbb{E}[Y_{1i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 1]$$

NOTE: The complement of the treatment group is the *control* group.

Selection bias

A naive estimand for ATE is the difference between average outcomes based on treatment status

However, this might be misleading:

$$\mathbb{E}\left[Y_i|D_i=1\right]-\mathbb{E}\left[Y_i|D_i=0\right]=\underbrace{\mathbb{E}\left[Y_{1i}|D_i=1\right]-\mathbb{E}\left[Y_{0i}|D_i=1\right]}_{\text{ATT}}+\underbrace{\mathbb{E}\left[Y_{0i}|D_i=1\right]-\mathbb{E}\left[Y_{0i}|D_i=0\right]}_{\text{selection bias}}$$

> **Causal inference is mostly about eliminating selection-bias**

EXAMPLE: Individuals who go to private universities probably have different characteristics than those who go to public universities.

Randomized control trial (RCT) solves selection bias

In an RCT, the treatments are randomly assigned. This means entails that D_i is *independent* of potential outcomes, namely

$$\{Y_{1i}, Y_{0i}\} \perp D_i$$

RCTs enables us to estimate ATE using the average difference in outcomes by treatment status:

$$\begin{aligned}
\mathbb{E}\left[Y_i | D_i = 1\right] - \mathbb{E}\left[Y_i | D_i = 0\right] &= \mathbb{E}\left[Y_{1i} | D_i = 1\right] - \mathbb{E}\left[Y_0 | D_i = 0\right] \\
&= \mathbb{E}\left[Y_{1i} | D_i = 1\right] - \mathbb{E}\left[Y_{0i} | D_i = 1\right] \\
&= \mathbb{E}\left[Y_{1i} - Y_{0i} | D_i = 1\right] \\
&= \mathbb{E}\left[Y_{1i} - Y_{0i}\right] \\
&= \text{ATE}
\end{aligned}$$

EXAMPLE: In theory, randomly assigning students to private and public universities would allow us to estimate the ATE going to private school have on future earnings. Clearly, RCT in this case is infeasibile.

Estimands and regression

Assume for now that the treatment effect is constant across all individuals, i.e.,

$$\tau = Y_{1i} - Y_{0i}, \quad \forall i.$$

Accordingly, we can express Y_i as

$$\begin{aligned} Y_i &= Y_{1i}D_i + Y_{0i}(1 - D_i) \\ &= Y_{0i} + D_i(Y_{1i} - Y_{0i}), \\ &= Y_{0i} + \tau D_i, & \text{since } \tau = Y_{1i} - Y_{0i} \\ &= \mathbb{E}[Y_{0i}] + \tau D_i + Y_{0i} - \mathbb{E}[Y_{0i}], & \text{add and subtract } \mathbb{E}[Y_{0i}] \end{aligned}$$

Or more conveniently

$$Y_i = \alpha + \tau D_i + u_i,$$

where $\alpha = \mathbb{E}[Y_{0i}]$ and $u_i = Y_{0i} - \mathbb{E}[Y_{0i}]$ is the random component of Y_{0i}.

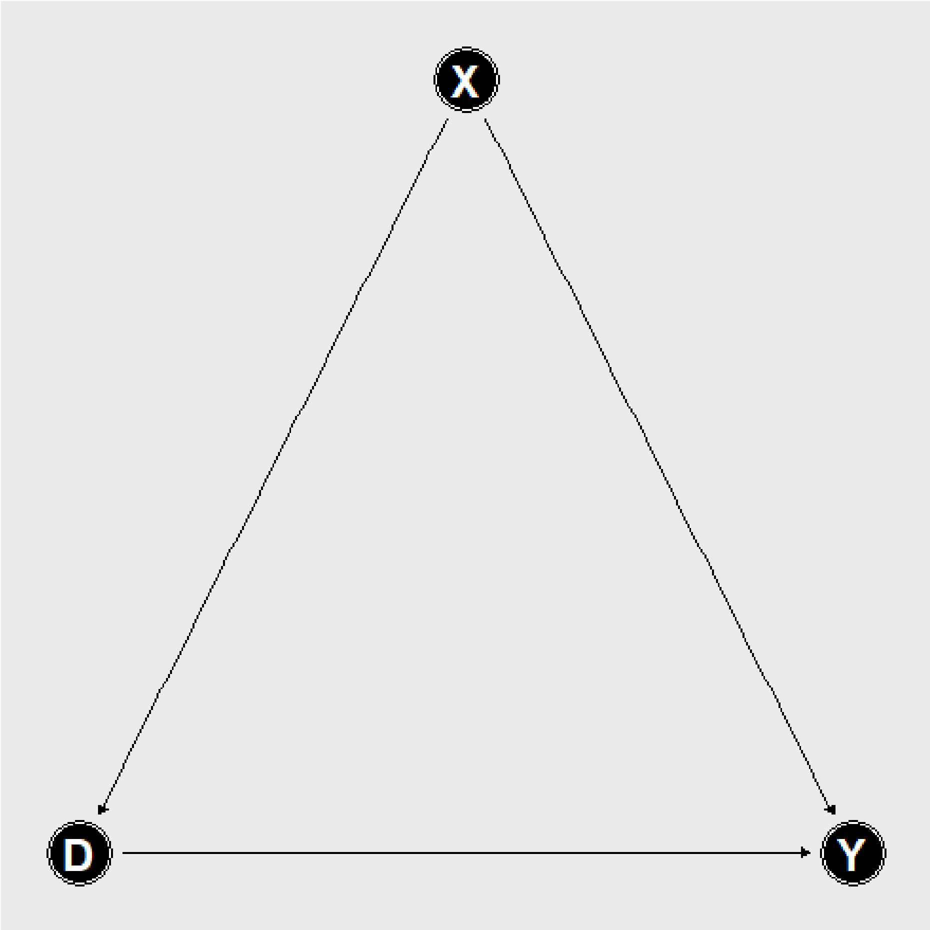
Confoundness

In observational studies, treatments are not randomly assigned. Think of $D_i = \{\text{private}, \text{public}\}$.

In this case, identifying causal effects depended on the *Unconfoundness* assumption (also known as "selection-on-observable"), which is defined as

$$\{Y_{1i}, Y_{0i}\} \perp D_i | \mathbf{X}_i$$

In words: treatment assignment is independent of potential outcomes *conditional* on observed $\mathbf{X}_i$, i.e., selection bias *disappears* when we control for $\mathbf{X}_i$.



Adjusting for confoundness

The most common approach for controlling for $\mathbf{X}_i$ is by adding them to the regression:

$$Y_i = \alpha + \tau D_i + \mathbf{X}_i'\boldsymbol{\beta} + u_i,$$

COMMENTS:

1. Strictly speaking, the above regression model is valid if we actually *believe* that the "true" model is $Y_i = \alpha + \tau D_i + \mathbf{X}_i'\boldsymbol{\beta} + u_i$.

2. If D_i is randomly assigned, adding $\mathbf{X}_i$ to the regression might increases the accuracy of ATE.

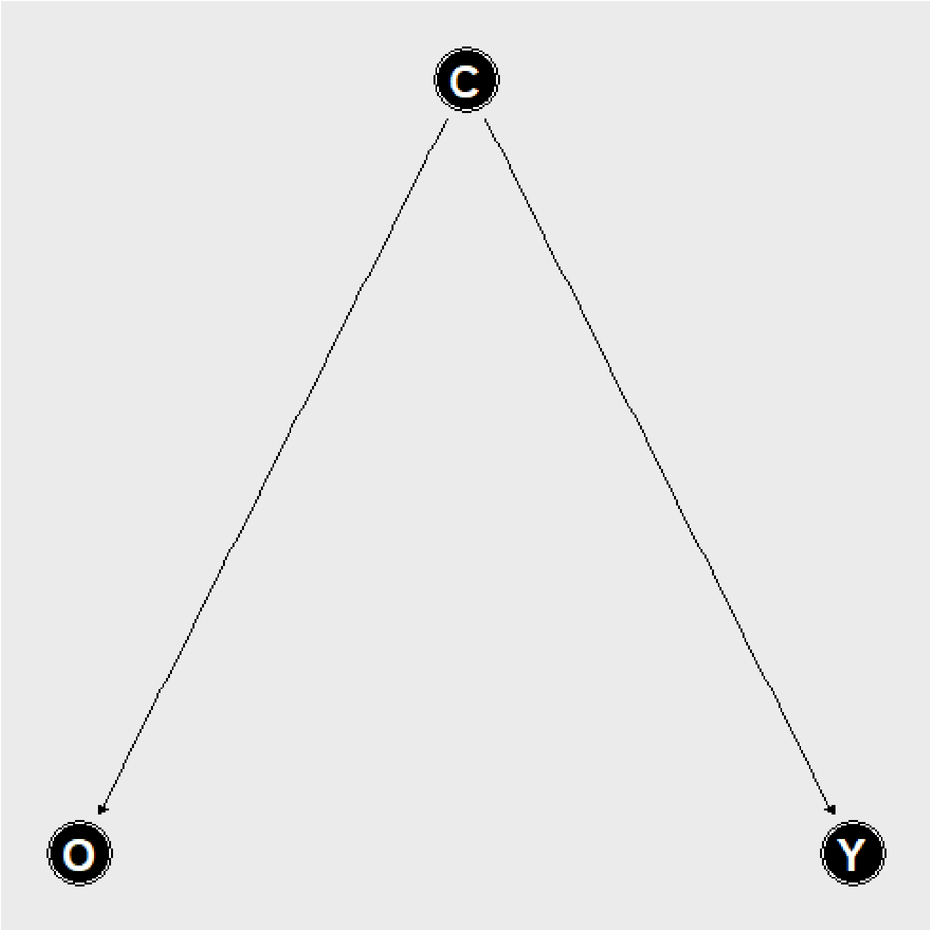
3. If D_i is assigned conditional on $\mathbf{X}_i$ (e.g., in observational settings), adding $\mathbf{X}_i$ to the regression eliminates selection bias.

An aside: Bad controls

Bad controls are variables that are themselves outcome variables.

EXAMPLE: Occupation as control in a return to years of schooling regression.

This distinction becomes important when dealing with high-dimensional data



Causal inference in high-dimensional setting

Consider again the standard "treatment effect regression":

$$Y_i = \alpha + \underbrace{\tau D_i}_{\text{low dimensional}} + \underbrace{\sum_{j=1}^{k} \beta_j X_{ij}}_{\text{high dimensional}} + \varepsilon_i, \quad \text{for } i = 1, \ldots, n$$

Our object of interest is $\hat{\tau}$, the estimated *average treatment effect* (ATE).

In high-dimensional settings $k \gg n$.

Variable Selection Using the Lasso

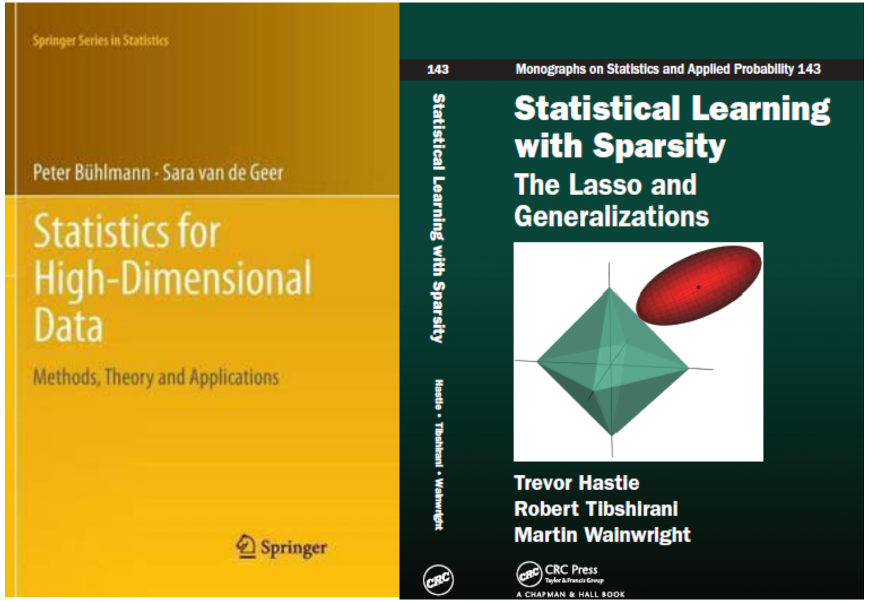
ML: guarantees vs. guidance

- Most (if not all) of what we've done so far is based on *guidance*
    - Choosing the number of folds in CV
    - Size of the holdout set
    - Tuning parameter(s)
    - loss function
    - function class
- In causal inference, we need *guaranties*
    - variable selection
    - Confidence intervals and p-values
- To get guarantees, we typically need
    - Assumptions about a "true" model
    - Asymptotics $n \to \infty$, $k \to ?$

Resources on the theory of Lasso

- *Statistical Learning with Sparsity - The Lasso and Generalizations* (Hastie,Tibshirani, and Wainwright), **Chapter 11: Theoretical Results for the Lasso.** (PDF available online)

- *Statistics for High-Dimensional Data - Methods, Theory and Applications* (Buhlmann and van de Geer), **Chapter 7: Variable Selection with the Lasso.**



Some notation to help you penetrate the Lasso literature

Suppose $\boldsymbol{\beta}$ is a $\boldsymbol{k} \times \mathbf{1}$ vector with typical element β_i.

- The ℓ_0-norm is defined as $||\boldsymbol{\beta}||_0 = \sum_{j=1}^{k} \mathbf{1}_{\{\beta_j \neq 0\}}$, i.e., the number of non-zero elements in $\boldsymbol{\beta}$.
- The ℓ_1-norm is defined as $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{k} |\beta_j|$.
- The ℓ_2-norm is defined as $||\boldsymbol{\beta}||_2 = \left(\sum_{j=1}^{k} |\beta_j|^2\right)^{\frac{1}{2}}$, i.e., Euclidean norm.
- The ℓ_∞-norm is defined as $||\boldsymbol{\beta}||_\infty = \mathbf{sup}_j\, |\beta_j|$, i.e., the maximum entries' magnitude of $\boldsymbol{\beta}$.

Recap: Regularized linear regression

Typically, the regularized linear regression estimator is given by

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^k} n^{-1} \sum_{i=1}^{n} (Y_i - \alpha - \boldsymbol{\beta}' X_i)^2 + \lambda R(\boldsymbol{\beta})$$

where

Method	$R(\boldsymbol{\beta})$
OLS	0
Subset selection	$\|\|\boldsymbol{\beta}\|\|_0$
Lasso	$\|\|\boldsymbol{\beta}\|\|_1$
Ridge	$\|\|\boldsymbol{\beta}\|\|_2^2$
Elastic Net	$\alpha\|\|\boldsymbol{\beta}\|\|_0 + (1-\alpha)\|\|\boldsymbol{\beta}\|\|_2^2$

NOTE: We assume throughout that both Y_i and the elements in X_i have been standardized so that they have mean zero and unit variance.

Lasso: The basic setup

The linear regression model:

$$Y_i = \alpha + \boldsymbol{X}_i' \boldsymbol{\beta}^0 + \varepsilon_i, \quad i = 1, \ldots, n,$$

$$\mathbb{E}\left[\varepsilon_i X_i\right] = 0, \quad \alpha \in \mathbb{R}, \quad \boldsymbol{\beta}^0 \in \mathbb{R}^k.$$

Under the *exact sparsity* assumption, only a subset of variables of size $s \ll k$ is included in the model where $s \equiv \|\boldsymbol{\beta}\|_0$ is the sparsity index.

$$\underbrace{\mathbf{X}_S = \left(X_{(1)}, \ldots, X_{(s)}\right),}_{\text{sparse variables}} \quad \underbrace{\mathbf{X}_{S^c} = \left(X_{(s+1)}, \ldots, X_{(k)}\right)}_{\text{non-sparse variables}}$$

where $S = \operatorname{supp}(\boldsymbol{\beta}^0) = \{j : \beta_j \neq 0\}$ is the subset of active predictors, $\mathbf{X}_S \in \mathbb{R}^{n \times s}$ corresponds to the subset of covariates that are in the sparse set, and $\mathbf{X}_{S^C} \in \mathbb{R}^{n \times k-s}$ is the subset of the "irrelevant" non-sparse variables.

Evaluation of the Lasso

Let β^0 denote the true vector of coefficients and let $\widehat{\beta}$ denote the Lasso estimator.

We can asses the quality of the Lasso in several ways:

I. Prediction quality

$$\text{Loss}_{\text{ pred}}\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}^0\right) = \frac{1}{N}\left\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^0\right\|_2^2$$

II. Parameter consistency

$$\text{Loss}_{\text{param}}\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}^*\right) = \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_2^2$$

III. Support recovery (sparsistency)

$$\text{Loss}_{\text{vs}}\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}^0\right) = \begin{cases} 0 & \text{if } \operatorname{sign}\left(\widehat{\beta}_i\right) = \operatorname{sign}\left(\beta_i^0\right) \text{ for all } i = 1, \ldots, p \\ 1 & \text{otherwise.} \end{cases}$$

Lasso as a variable selection tool

- Variable selection consistency is essential for causal inference.
- Lasso is often used as a variable selection tool.
- Being able to select the "true" support by Lasso relies on strong assumptions about
  - the ability to distinguish between relevant and irrelevant variables.
  - the ability to identify $\boldsymbol{\beta}$.

Critical assumption #1: Distinguishable betas

Lower eigenvalue the min eigenvalue $\lambda_{\min}$ of the sub-matrix $\mathbf{X}_S$ is bounded away from zero.

$$\lambda_{\min}\left(\mathbf{X}_S'\mathbf{X}_S/N\right) \geq C_{\min} > 0$$

Linear dependence between the columns of $\mathbf{X}_s$ would make it impossible to identify the true $\boldsymbol{\beta}$, even if we *knew* which variables are included in $\mathbf{X}_S$.

NOTE: The high-dimension's lower eigenvalue condition replaces the low-dimension's rank condition (i.e., that $\mathbf{X}'\mathbf{X}$ is invertible)

Critical assumption #2: Distinguishable active predictors

Irrepresentability condition (Zou ,2006; Zhao and Yu, 2006): There must exist some $\eta \in [0, 1)$ such that

$$\max_{j \in S^c} \left\| \left(\mathbf{X}_S' \mathbf{X}_S\right)^{-1} \mathbf{X}_S' \mathbf{x}_j \right\|_1 \leq \eta$$

INTUITION: What's inside $\|\cdot\|$ is like regressing $\mathbf{x}_j$ on the variables in $\mathbf{X}_s$.

- When $\eta = 0$, the sparse and non-sparse variables are orthogonal to each other.
- When $\eta = 1$, we can reconstruct (some elements of) $\mathbf{X}_S$ using $\mathbf{X}_{S^C}$.

Thus, the irrepresentability condition roughly states that we can distinguish the sparse variables from the non-sparse ones.

Some words on the optimal tuning parameter

- As we've seen thorough this course, it is also common to choose λ empirically, often by cross-validation, based on its predictive performance

- In causal analysis, inference and not prediction is the end goal. Moreover, these two objectives often contradict each other (bias vs. variance)

- Optimally, the choice of λ should provide guarantees about the performance of the model.

- Roughly speaking, when it comes to satisfying sparsistency, λ is set such that it selects non-zero β's with high probability.

Lasso In Aid of Causal Inference

"Naive" implementation of the Lasso

Run `glmnet`

```
glmnet(Y ~ DX)
```

where `DX` is the feature matrix which includes X_i and D_i.

The estimated coefficients are:

$$\left(\widehat{\alpha}, \hat{\tau}, \widehat{\boldsymbol{\beta}}'\right)' = \underset{\alpha,\tau\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau D_i - \boldsymbol{\beta}' X_i\right)^2 + \lambda \left(|\tau| + \sum_{j=1}^{k} |\beta_j|\right)$$

PROBLEMS:

1. Both $\hat{\tau}$ and $\widehat{\boldsymbol{\beta}}$ are biased towards zero (shrinkage).
2. Lasso might drop D_i, i.e., shrink $\hat{\tau}$ to zero. Can also happen to relevant confounders.
3. How to choose λ?

Toward a solution

OK, lets keep D_i in:

$$\left(\widehat{\alpha}, \hat{\tau}, \widehat{\boldsymbol{\beta}}'\right)' = \underset{\alpha,\tau\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^k}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau D_i - \boldsymbol{\beta}' X_i\right)^2 + \lambda\left(\sum_{j=1}^{k} |\beta_j|\right)$$

Then, *debias* the results using "Post-Lasso", i.e, use Lasso for variable selection and then run OLS with the selected variables.

PROBLEMS:

1. How to choose λ?

2. *Omitted variable bias*: The Lasso might drop features that are correlated with D_i because they are "bad" predictor of Y_i.

Problem solved?

What can go wrong? Distribution of $\sqrt{n}(\hat{\alpha} - \alpha)$ is not what you think

$$y_i = d_i\alpha + x_i\beta + \epsilon_i, \quad d_i = x_i\gamma + v_i$$

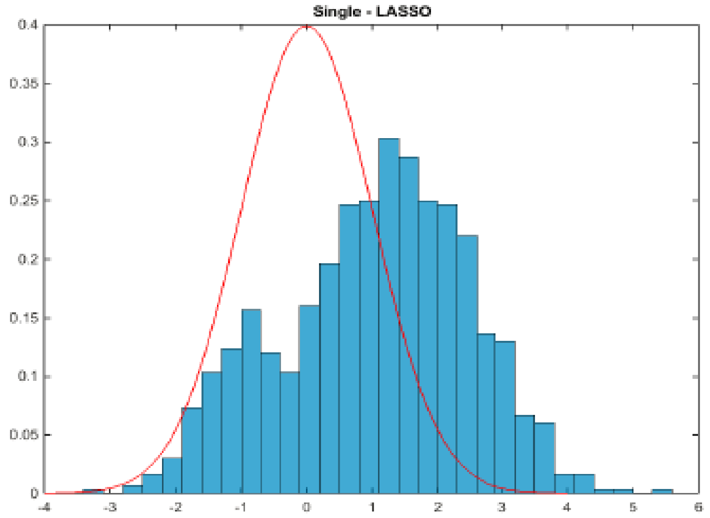
$$\alpha = \mathbf{0}, \quad \beta = \mathbf{.2}, \quad \gamma = .8,$$

$$n = 100$$

$$\epsilon_i \sim N(0,1)$$

$$(d_i, x_i) \sim N\left(0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}\right)$$

- selection done by **Lasso**



Reject $H_0 : \alpha = 0$ (the truth) of no effect about 50% of the time

Solution: Double-selection Lasso (Belloni, et al., REStud 2013)

First step: Regress Y_i on $\mathbf{X}_i$ and D_i on $\mathbf{X}_i$:

$$\widehat{\gamma} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{\gamma}'\mathbf{X}_i\right)^2 + \lambda_\gamma \left(\sum_{j=2}^{p} |\gamma_j|\right)$$
$$\hat{\delta} = \underset{\boldsymbol{\delta}\in\mathbb{R}^{q+1}}{\arg\min} \sum_{i=1}^{n} \left(D_i - \boldsymbol{\delta}'\mathbf{X}_i\right)^2 + \lambda_\delta \left(\sum_{j=2}^{q} |\delta_j|\right)$$

Second step: Refit the model by OLS and include the $\mathbf{X}$'s that are significant predictors of Y_i *and* D_i.

Third step: Proceed to inference using standard confidence intervals.

> The Tuning parameter λ is set such that the non-sparse coefficients are correctly selected with high probability.

Does it work?

Double Selection Works

$$y_i = d_i\alpha + x_i\beta + \epsilon_i, \quad d_i = x_i\gamma + v_i$$

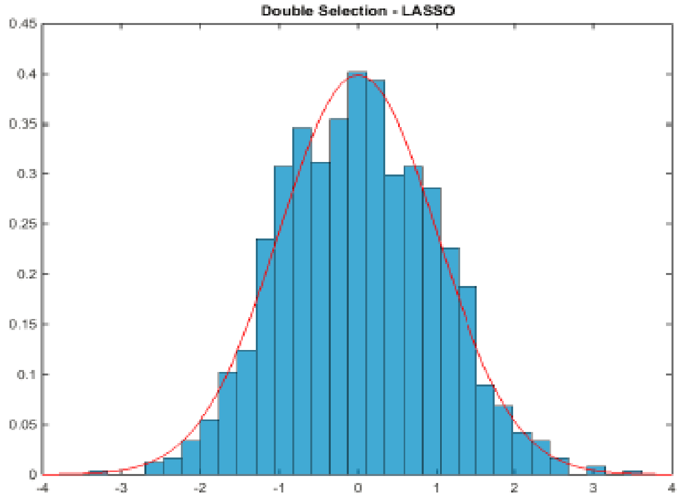
$$\alpha = \mathbf{0}, \quad \beta = \mathbf{.2}, \quad \gamma = .8,$$

$$n = 100$$

$$\epsilon_i \sim N(0,1)$$

$$(d_i, x_i) \sim N\left(0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}\right)$$

- **double selection** done by **Lasso**



Reject $H_0 : \alpha = 0$ (the truth) about 5% of the time (nominal size = 5%)

Intuition: Partialling-out regression

consider the following two alternatives for estimating the effect of X_i on Y_i, adjusting for Z_i:

Alternative 1: Run

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i$$

Alternative 2: First, run Y_i on Z_i and X_i on Z_i and keep the residuals, i.e., run

$$Y_i = \gamma_0 + \gamma_1 Z_i + u_i^Y, \quad \text{and} \quad X_i = \delta_0 + \delta_1 Z_i + u_i^X$$

and keep $\widehat{u}_i^Y$ and $\widehat{u}_i^X$. Next, run

$$\widehat{u}_i^Y = \beta^* \widehat{u}_i^X + v_i$$

According to the Frisch-Waugh-Lovell (FWV) Theorem,

$$\widehat{\beta} = \widehat{\beta}^*.$$

Notes on the guarantees of double-selection Lasso

Approximate Sparsity Consider the following regression model:

$$Y_i = f(\boldsymbol{W}_i) + \varepsilon_i = \boldsymbol{X}_i'\boldsymbol{\beta}_0 + r_i + \varepsilon_i$$

where r_i is the approximation error.

Under *approximate sparsity*, it is assumed that $f(\boldsymbol{W}_i)$ can be approximated sufficiently well (up to r_i) by $\boldsymbol{X}_i'\boldsymbol{\beta}_0$, while using only a small number of non-zero coefficients.

Restricted Sparse Eigenvalue Condition (RSEC) This condition puts bounds on the number of variables outside the support the Lasso can select. Relevant for the post-lasso stage.

Tuning λ

Further extensions of double-selection

1. Chernozhukov et al. (AER 2017): Other function classes ("Double-ML"), e.g., use random forest for $Y_i \sim X_i$ and regularized logit for $D_i \sim X_i$.

2. Instrumental variables (Belloni et al., Ecta 2012, Chernozhukov et al., AER 2015)

3. Heterogeneous treatment effects (Belloni et al., Ecta 2017)

4. Panel data (Belloni, et al., JBES 2016)

Empirical Illustration using `hdm`

The hdm package*

"High-Dimensional Metrics" (`hdm`) by Victor Chernozhukov, Chris Hansen, and Martin Spindler is an R package for estimation and quantification of uncertainty in high-dimensional approximately sparse models.

To install the package:

```
install.packages("hdm")
```

To load the package:

```
library(hdm)
```

[*] There now a new Stata module named Lassopack that includes a rich suite of programs for regularized regression in high-dimensional setting.

Illustration: Testing for growth convergence

The standard growth convergence empirical model:

$$Y_{i,T} = \alpha_0 + \alpha_1 Y_{i,0} + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where

- $Y_{i,T}$ is the growth rate of GDP over a specified decade
- $Y_{i,0}$ is the log of the initial level of GDP at the beginning of the specified decade
- X_{ij} are country characteristics.

The growth convergence hypothesis implies that $\alpha_1 < 0$.

Growth data

To test the growth convergence hypothesis, we will make use of the Barro and Lee (1994) dataset

```
data("GrowthData")
```

The data contain macroeconomic information for large set of countries over several decades. In particular,

- n = 90 countries
- k = 60 country features

Not so big...

Nevertheless, the number of covariates is large relative to the sample size $\Rightarrow$ variable selection is important!

```r
library(tidyverse)

GrowthData %>% as_tibble %>% head(2)
```

```
## # A tibble: 2 x 63
##   Outcome intercept gdpsh465 bmp1l freeop freetar   h65  hm65  hf65   p65
##     <dbl>     <int>    <dbl> <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 -0.0243         1     6.59 0.284  0.153  0.0439 0.007 0.013 0.001 0.290
## 2  0.100          1     6.83 0.614  0.314  0.0618 0.019 0.032 0.007 0.91
## # ... with 53 more variables: pm65 <dbl>, pf65 <dbl>, s65 <dbl>,
## #   sm65 <dbl>, sf65 <dbl>, fert65 <dbl>, mort65 <dbl>, lifee065 <dbl>,
## #   gpop1 <dbl>, fert1 <dbl>, mort1 <dbl>, invsh41 <dbl>, geetot1 <dbl>,
## #   geerec1 <dbl>, gde1 <dbl>, govwb1 <dbl>, govsh41 <dbl>,
## #   gvxdxe41 <dbl>, high65 <dbl>, highm65 <dbl>, highf65 <dbl>,
## #   highc65 <dbl>, highcm65 <dbl>, highcf65 <dbl>, human65 <dbl>,
## #   humanm65 <dbl>, humanf65 <dbl>, hyr65 <dbl>, hyrm65 <dbl>,
## #   hyrf65 <dbl>, no65 <dbl>, nom65 <dbl>, nof65 <dbl>, pinstab1 <dbl>,
## #   pop65 <int>, worker65 <dbl>, pop1565 <dbl>, pop6565 <dbl>,
## #   sec65 <dbl>, secm65 <dbl>, secf65 <dbl>, secc65 <dbl>, seccm65 <dbl>,
## #   seccf65 <dbl>, syr65 <dbl>, syrm65 <dbl>, syrf65 <dbl>,
## #   teapri65 <dbl>, teasec65 <dbl>, ex1 <dbl>, im1 <dbl>, xr65 <dbl>,
## #   tot1 <dbl>
```

Data processing

Rename the response and "treatment" variables:

```
GrowthData <- GrowthData %>%
  rename(YT = Outcome, Y0 = gdpsh465)
```

Transform the data to vectors and matrices (to be used in the `rlassoEffect()` function)

```
YT <- GrowthData %>% select(YT) %>% pull()

Y0 <- GrowthData %>% select(Y0) %>% pull()

X <- GrowthData %>%
  select(-c("Y0", "YT")) %>%
  as.matrix()

Y0_X <- GrowthData %>%
  select(-YT) %>%
  as.matrix()
```

Estimation of the convergence parameter α_1

(1) OLS

```
ols <- lm(YT ~ ., data = GrowthData)
```

(2) Naive (rigorous) Lasso

```
naive_Lasso <- rlasso(x = Y0_X, y = YT)
```

Does the Lasso drop `Y0`?

```
naive_Lasso$beta[2]
```

```
## Y0
##  0
```

Unfortunately, yes...

Estimation of the convergence parameter α_1

(3) Partialling out Lasso

```
part_Lasso <- rlassoEffect(x = X, y = YT, d = Y0,
                           method = "partialling out")
```

(4) Double-selection Lasso

```
double_Lasso <- rlassoEffect(x = X, y = YT, d = Y0,
                           method = "double selection")
```

Tidying the results

```
# OLS
ols_tbl <- tidy(ols) %>%
  filter(term == "Y0") %>%
  mutate(method = "OLS") %>%
  select(method, estimate, std.error)
# Naive Lasso
naive_Lasso_tbl <- tibble(method = "Naive Lasso",
                              estimate = NA,
                              std.error = NA)
# Partialling-out Lasso
results_part_Lasso <- summary(part_Lasso)[[1]][1, 1:2]
part_Lasso_tbl     <- tibble(method = "Partialling-out Lasso",
                          estimate = results_part_Lasso[1],
                          std.error = results_part_Lasso[2])
# Double-sellection Lasso
results_double_Lasso <- summary(double_Lasso)[[1]][1, 1:2]
double_Lasso_tbl <- tibble(method = "Double-selection Lasso",
                           estimate = results_double_Lasso[1],
                           std.error = results_double_Lasso[2])
```

Results of the convergence test

```r
kable(bind_rows(ols_tbl,
          naive_Lasso_tbl,
          part_Lasso_tbl,
          double_Lasso_tbl),
      format = "html")
```

method	estimate	std.error
OLS	-0.0093780	0.0298877
Naive Lasso	NA	NA
Partialling-out Lasso	-0.0498115	0.0139364
Double-selection Lasso	-0.0500059	0.0157914

Double-selection and partialling-out yield much more precise estimates and provide support the conditional convergence hypothesis `slides %>% end()`

Source code