09 - High-Dimensional Confounding Adjustment

ml4econ, HUJI 2025

Itamar Caspi

May 25, 2024 (updated: 2025-05-25)

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,
)
```

```
## package 'checkmate' successfully unpacked and MD5 sums checked
## package 'hdm' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##     C:\Users\internet\AppData\Local\Temp\RtmpOWU69M\downloaded_packages
```

Outline

- Lasso and Variable Selection
- High Dimensional Confoundedness
- Empirical Illustration using `hdm`

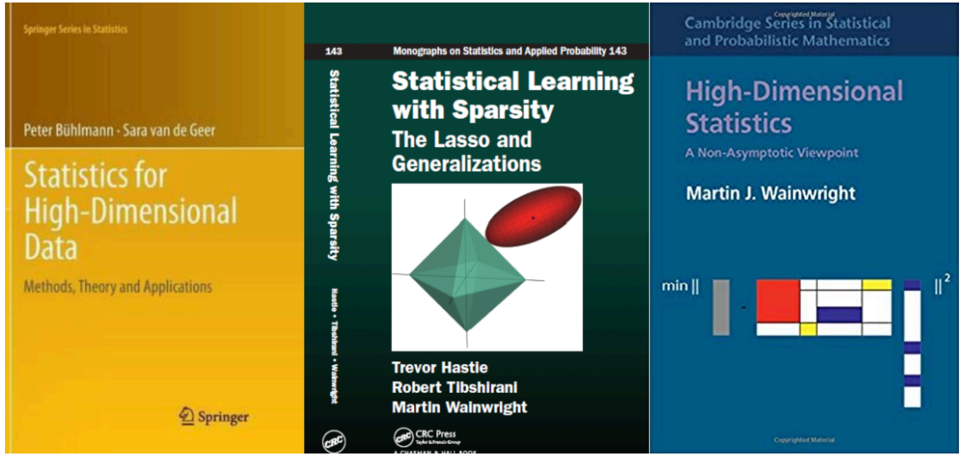
Lasso and Variable Selection

Key Lasso Theory Resources

- *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.

- *High Dimensional Statistics - A Non-Asymptotic Viewpoint* (Wainwright), Chapter 7: Sparse Linear Models in High Dimensions



Guidance vs. Guarantees: Fundamental Differences

- We've primarily relied on *guidance* for our work:
  - Selection of folds in CV
  - Size determination of the holdout set
  - Tuning parameter(s) adjustment
  - Loss function selection
  - Function class selection
- But in causal inference, *guarantees* become vital:
  - Selecting variables
  - Deriving confidence intervals and p-values
- To attain these guarantees, we generally need:
  - Assumptions regarding a "true" model
  - Asymptotic principles, such as $n \to \infty$, $k \to ?$

Key Notations in Lasso Literature

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

- ℓ_0-norm is $||\boldsymbol{\beta}||_0 = \sum_{j=1}^{k} \mathbf{1}_{\{\beta_j \neq 0\}}$, indicating the count of non-zero elements in $\boldsymbol{\beta}$.
- ℓ_1-norm is $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{k} |\beta_j|$.
- ℓ_2-norm or Euclidean norm is $||\boldsymbol{\beta}||_2 = \left(\sum_{j=1}^{k} |\beta_j|^2\right)^{\frac{1}{2}}$.
- ℓ_∞-norm is $||\boldsymbol{\beta}||_\infty = \mathbf{sup}_j \, |\beta_j|$, signifying the maximum magnitude among $\boldsymbol{\beta}$'s entries.
- Support of $\boldsymbol{\beta}$ is $S \equiv \mathbf{supp}(\boldsymbol{\beta}) = \{\beta_j \neq 0, j = 1, \ldots, j\}$, the subset of non-zero coefficients.
- Size of the support $\boldsymbol{s} = |S|$ is the count of non-zero elements in $\boldsymbol{\beta}$, namely $\boldsymbol{s} = ||\boldsymbol{\beta}||_0$

Understanding the Basic Setup of Lasso

The linear regression model is given as:

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

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

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

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

Here, S is the subset of active predictors, $\mathbf{X}_S \in \mathbb{R}^{n \times s}$ corresponds to the subset of covariates in the sparse set, and $\mathbf{X}_{S^C} \in \mathbb{R}^{n \times k-s}$ refers to the subset of "irrelevant" non-sparse variables.

Lasso: The Optimization

The Lasso (Least Absolute Shrinkage and Selection Operator), introduced by Tibshirani (1996), poses the following optimization problem:

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

In this setup, Lasso places a "budget constraint" on the sum of *absolute* values of β's.

Differing from ridge, the Lasso penalty is linear (shifting from 1 to 2 bears the same weight as moving from 101 to 102).

A major strength of Lasso lies in its ability to perform model selection - it zeroes out most of the β's in the model, making the solution *sparse*.

Any penalty involving the ℓ_1 norm will achieve this.

Evaluating the Lasso

Suppose β^0 is the true vector of coefficients and $\widehat{\beta}$ represents the Lasso estimator. We can evaluate Lasso's effectiveness in several ways:

I. Prediction Quality

$$\text{Loss}_{\text{ pred}}\left(\widehat{\boldsymbol{\beta}};\boldsymbol{\beta}^0\right)=\frac{1}{N}\left\|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\mathbf{X}\right\|_2^2=\frac{1}{N}\sum_{j=1}^{k}\left[(\hat{\beta}_j-\beta_j^0)\mathbf{X}_{(j)}\right]^2$$

II. Parameter Consistency

$$\text{Loss}_{\text{param}}\left(\widehat{\boldsymbol{\beta}};\boldsymbol{\beta}^0\right)=\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_2^2=\sum_{j=1}^{k}(\hat{\beta}-\beta^0)^2$$

III. Support Recovery (Sparsistency)

For example, score +1 if $\mathbf{sign}(\beta^0)=\mathbf{sign}(\beta_j)$ for all $j=1,\ldots,k$, and `0` otherwise.

Leveraging Lasso for Variable Selection

- Variable selection consistency is crucial for causal inference, considering omitted variable bias.

- Lasso frequently serves as a tool for variable selection.

- The successful selection of the "true" support by Lasso depends heavily on strong assumptions about:

  - Distinguishing between relevant and irrelevant variables.
  - Identifying $\boldsymbol{\beta}$.

Lasso Sparsistency: Necessary & Sufficient

Exact support recovery ⇔ Satisfying key conditions:

1.Active Gram invertible: $\lambda_{\min}\left(\frac{X_S^\top X_S}{n}\right) > 0$

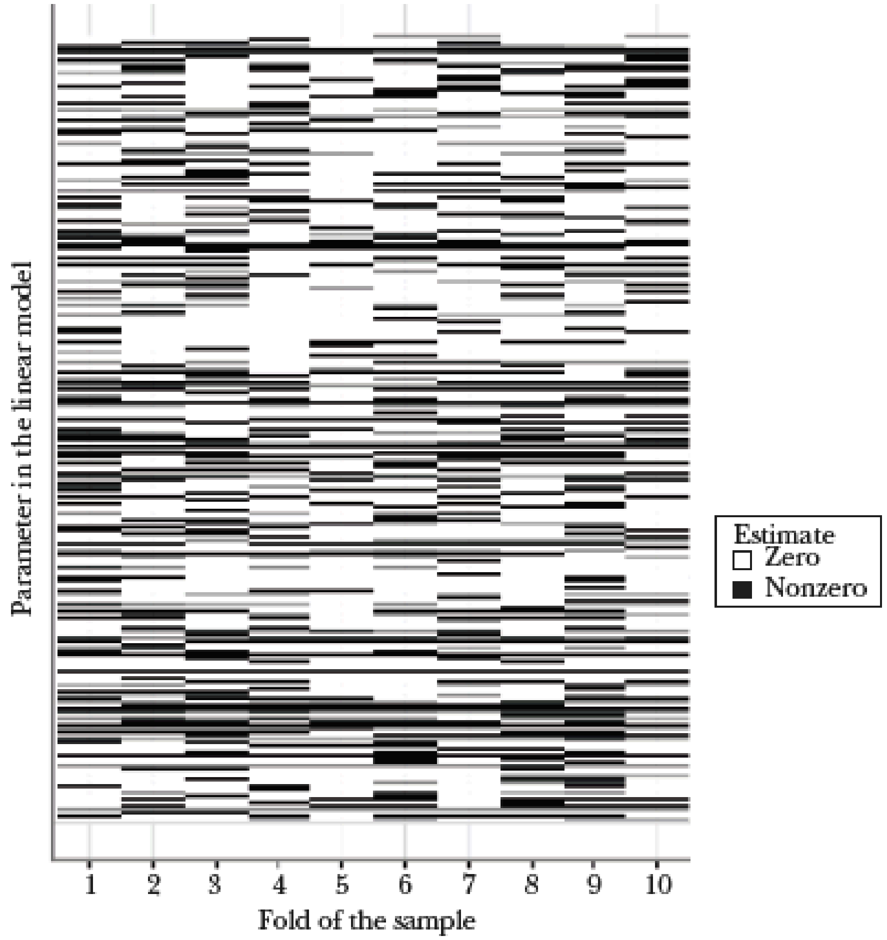
1. Irrepresentable condition: $\left\|\left(X_S^\top X_S\right)^{-1} X_S^\top x_j\right\|_1 \leq 1-\gamma$
2. β-min condition: $\min_{j\in S}|\beta_j| > c\cdot\lambda_n$
3. Penalty level: $\lambda_n \geq c\cdot\sigma\cdot\sqrt{\frac{\log p}{n}}$
4. Sample size requirement: $n \gtrsim k\cdot\log p$

Plain English:

1. The true variables aren't collinear
2. Noise variables don't mimic the true ones
3. True coefficients are large enough to detect
4. The noise is tame
5. You have enough samples to work with

Figure 2

Selected Coefficients (Nonzero Estimates) across Ten LASSO Regressions



Source: Mullainathan and Spiess (JEP 2017).

Setting the Optimal Tuning Parameter

- Throughout this course, we have frequently chosen λ empirically, often by cross-validation, based on its predictive performance.
- In causal analysis, however, the end goal is inference, not prediction. These two objectives often conflict (bias vs. variance).
- Ideally, the choice of λ should provide assurances about the model's performance.
- Generally, for satisfying sparsistency, we set λ such that it selects non-zero β's with a high probability.

High Dimensional Confoundedness

"Naive" Implementation of the Lasso

Run `glmnet`

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

where `DX` is the feature matrix which includes both the treatment D and the features vector X. The estimated coefficients are:

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

ISSUES:

1. Both $\hat{\tau}$ and $\widehat{\boldsymbol{\beta}}$ experience shrinkage, thus biased towards zero.
2. Lasso might eliminate D_i, i.e., shrink $\hat{\tau}$ to zero. The same can happen to relevant confounding factors.
3. The choice of λ is a challenge.

Moving Towards a Solution

To avoid eliminating D_i, we can adjust the Lasso regression:

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

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

ISSUES: The Lasso might eliminate features that are correlated with D_i because they are not good predictors of Y_i, leading to *omitted variable bias*.

Problem Solved?

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

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

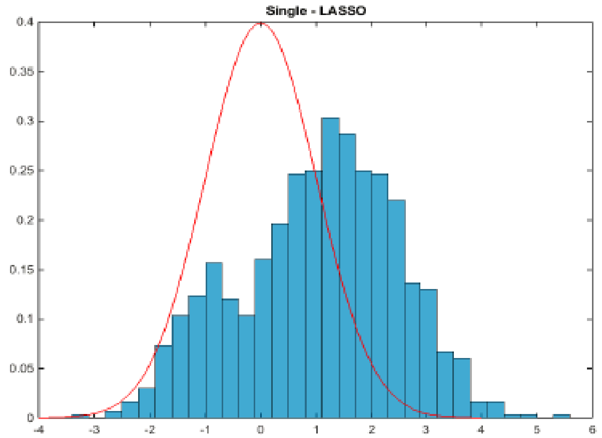
$$\alpha = 0, \quad \beta = .2, \quad \gamma = .8,$$

$$n = 100$$

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

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

- selection done by **Lasso**



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

Source: https://stuff.mit.edu/~vchern/papers/Chernozhukov-Saloniki.pdf

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

Step 1: Perform two Lasso regressions: Y_i on X_i and D_i on X_i:

$$\widehat{\gamma} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{\gamma}' X_i\right)^2 + \lambda_\gamma \left(\sum_{j=2}^{p} |\gamma_j| \right)$$

$$\hat{\delta} = \underset{\boldsymbol{\delta} \in \mathbb{R}^{q+1}}{\arg\min} \sum_{i=1}^{n} \left(D_i - \boldsymbol{\delta}' X_i\right)^2 + \lambda_\delta \left(\sum_{j=2}^{q} |\delta_j| \right)$$

Step 2: Refit the model using OLS, but only include the $\mathbf{X}$'s that were significant predictors of both Y_i and D_i.

Step 3: Proceed with the inference using standard confidence intervals.

> The tuning parameter λ is set in a way that ensures non-sparse coefficients are correctly selected with high probability.

Does it Work?

Double Selection Works

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

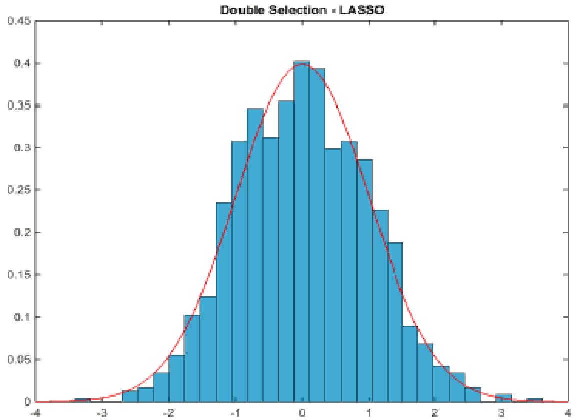
$$\alpha = \mathbf{0}, \quad \beta = .\mathbf{2}, \quad \gamma = .8,$$

$$n = 100$$

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

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

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



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

Source: https://stuff.mit.edu/~vchern/papers/Chernozhukov-Saloniki.pdf

Statistical Inference

Uniform Validity of the Double Selection

Theorem (Belloni, Chernozhukov, Hansen: WC 2010, ReStud 2013)

Uniformly within a class of approximately sparse models with restricted isometry conditions

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \to_d N(0, 1),$$

where σ_n^2 is conventional variance formula for least squares. Under homoscedasticity, semi-parametrically efficient.

- Model selection mistakes are asymptotically negligible due to double selection.
- Analogous result also holds for *endogenous* models, see Chernozhukov, Hansen, Spindler, *Annual Review of Economics*, 2015.

Source: https://stuff.mit.edu/~vchern/papers/Chernozhukov-Saloniki.pdf

Intuition: Partialling-out regression

Consider two methods for estimating the effect of X_{1i} (a scalar) on Y_i, while adjusting for X_{2i}:

Alternative 1: Run

$$Y_i = \alpha + \beta X_{1i} + \gamma X_{2i} + \varepsilon_i$$

Alternative 2: First, run Y_i on X_{2i} and X_{1i} on X_{2i} and keep the residuals, i.e., run

$$Y_i = \gamma_0 + \gamma_1 X_{2i} + u_i^Y, \quad \textbf{and} \quad X_{1i} = \delta_0 + \delta_1 X_{2i} + u_i^{X_1},$$

and keep $\widehat{u}_i^Y$ and $\widehat{u}_i^{X_1}$. Next, run

$$\widehat{u}_i^Y = \beta^* \widehat{u}_i^{X_1} + v_i.$$

According to the Frisch-Waugh-Lovell (FWV) Theorem, $\widehat{\beta} = \widehat{\beta}^*$.

Guarantees of Double-selection Lasso (VERY Wonkish)

Approximate Sparsity Consider the following regression model:

$$Y_i = f(W_i) + \varepsilon_i = X_i' \boldsymbol{\beta}^0 + r_i + \varepsilon_i, \quad 1, \ldots, n$$

where r_i is the approximation error.

Under *approximate sparsity*, it is assumed that $f(W_i)$ can be approximated sufficiently well (up to r_i) by $X_i' \boldsymbol{\beta}^0$, 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.

Regularization Event The tuning parameter λ is to a value that it selects to correct model with probability of at least p, where p is set by the user. Further assumptions regarding the quantile function of the maximal value of the gradient of the objective function at $\boldsymbol{\beta}^0$, and the error term (homoskedasticity vs. heteroskedasticity). See Belloni et al. (2012) for further details.

Additional Extensions of Double-selection

1. **Other Function Classes (Double-ML):** Chernozhukov et al. (AER 2017) proposed using other function classes, such as applying random forest for $Y \sim X$ and regularized logit for $D \sim X$.

2. **Instrumental Variables:** Techniques involving instrumental variables have been developed by Belloni et al. (Ecta 2012) and Chernozhukov et al. (AER 2015). For further understanding, please refer to the problem set.

3. **Heterogeneous Treatment Effects:** Heterogeneous treatment effects have been studied by Belloni et al. (Ecta 2017). We'll explore this topic more thoroughly next week.

4. **Panel Data:** Consideration for panel data was made by Belloni et al. (JBES 2016).

Evidence on the Applicability of Double-Lasso

"Machine Labor" (Angrist and Frandsen, 2022 JLE):

- ML can be useful for regression-based causal inference using Lasso.
- Post-double-selection (PDS) Lasso offers data-driven sensitivity analysis.
- ML-based instrument selection can improve on 2SLS, but split-sample IV and limited information maximum likelihood (LIML) perform better.
- ML might not be optimal for Instrumental Variables (IV) applications in labor economics. This is due to the creation of artificial exclusion restrictions potentially resulting in inaccurate findings.
- Empirical takeaway: ML is useful in regression settings but ill-suited for instrument selection or IV control inferences unless handled with extreme care.

More from "Labor Machine"

quality of PDS bias mitigation depends on design features like regressor variance and the extent of OVB. Moreover, we have examined a scenario in which OLS with full-dictionary control is feasible and effectively removes OVB. Even so, PDS seems a useful tool for sensitivity analysis in a regression context, where analysts may choose from an abundance of possible control variables. Findings where the target causal estimate remains reasonably stable while the list of selected controls varies from one routine to another reinforce claims of robustness.

It is worth emphasizing that a causal interpretation of the ML estimates in table 2 turns on a maintained conditional independence assumption. ML methods do not *create* quasi-experimental variation. Rather, ML uses data to pick from among a large set of modeling options founded on a common identifying assumption. This facilitates estimation in high-dimensional control scenarios and may increase precision (although that is not the finding here). We have also noted considerable sensitivity to implementation details, specifically to software choice and lasso penalty determination.

Source: Angrist and Frandsen (2022).

Table 2
Post-lasso Estimates of Elite College Effects

	Double Selection (PDS)			Outcome Selection			All Controls
	Plug-In (1)	CV λ (2)	cvlasso (3)	Plug-In (7)	CV λ (8)	cvlasso (9)	OLS (7)
	A. Private School Effects						
Estimate	.038	.020	.040	.046	.043	.042	.017
	(.040)	(.039)	(.041)	(.041)	(.043)	(.043)	.039
Number of controls	18	100	112	10	35	50	303
	B. Effects of School-Average SAT Score/100						
Estimate	−.009	−.013	−.009	−.008	−.009	−.008	−.012
	(.020)	(.018)	(.019)	(.020)	(.019)	(.019)	(.018)
Number of controls	24	151	58	10	34	43	303
	C. Effects of Attending Schools Rated Highly Competitive or Better						
Estimate	.068	.051	.073	.076	.080	.082	.053
	(.033)	(.033)	(.033)	(.031)	(.032)	(.032)	.033
Number of controls	17	185	106	10	34	43	303

Bottom line: ML-selected control strategies do not overturn the main conclusion from Dale and Krueger (2002).

DML: A Cautionary Tale

Hünermund, Louw, and Caspi (2023 JCI):

- DML is highly sensitive to a few "bad controls" in the covariate space, leading to potential bias.

- This bias varies depending on the theoretical causal model, raising questions about the practicality of data-driven control variable selection.

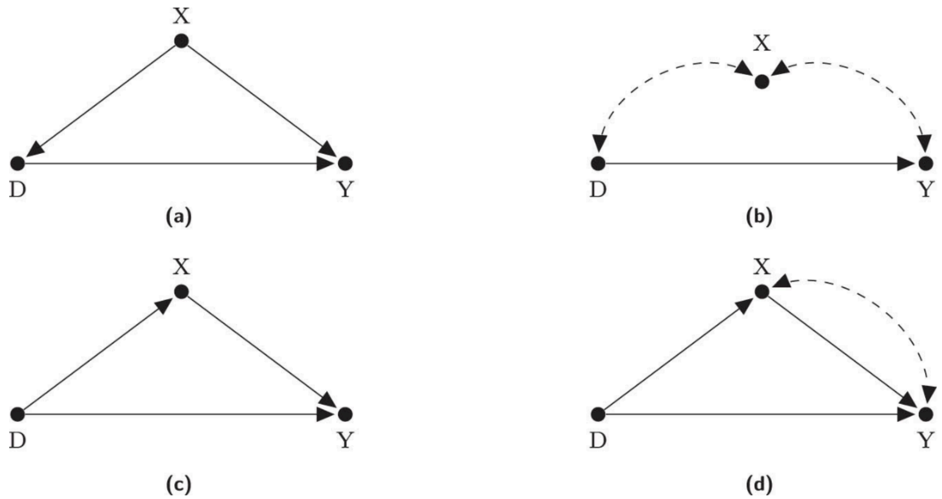
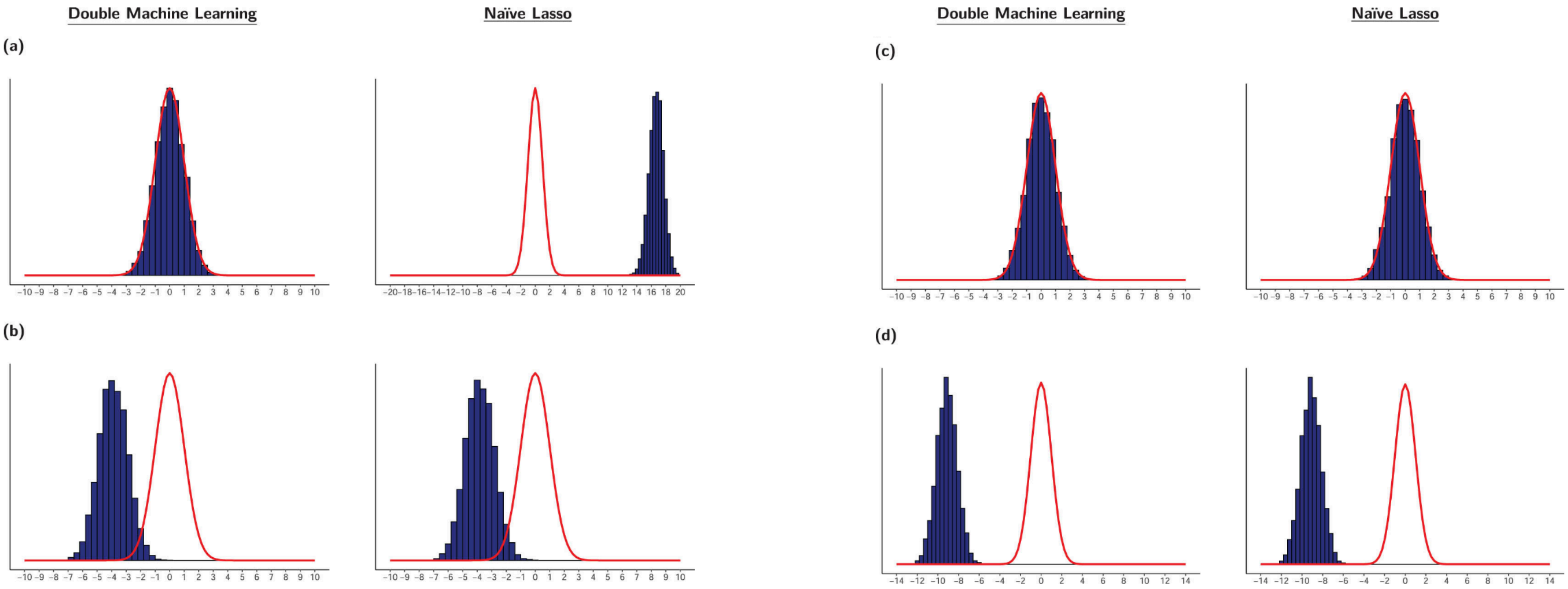


Figure 1

Directed acyclic graphs representing different structural causal models. (a) Good control, (b) M-graph, (c) mediator, and (d) confounded mediator.

Simulation Results



Empirical Relevance

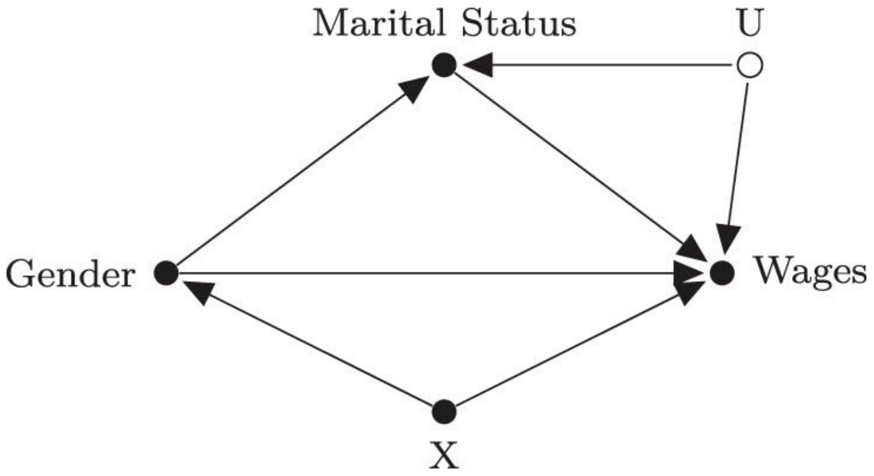


Table 2

Effect of gender on log wages using PSID data from [25] (standard errors in parentheses)

Wave	1981	1990	1999	2007	2009	2011
OLS	−0.249	−0.137	−0.158	−0.168	−0.157	−0.145
	(0.016)	(0.014)	(0.016)	(0.015)	(0.015)	(0.016)
DML	−0.268	−0.139	−0.158	−0.164	−0.157	−0.136
	(0.017)	(0.015)	(0.016)	(0.016)	(0.016)	(0.017)
DML incl.	−0.270	−0.154	−0.173	−0.190	−0.179	−0.163
Marital status	(0.022)	(0.019)	(0.020)	(0.019)	(0.020)	(0.021)

Bottom line

Marital status is double-edged:

- Mediator edge: cutting it removes a genuine slice of the gender wage gap.
- Collider edge: cutting it invites invisible traits to contaminate the estimate.

Good causal inference tools (instrumental variables for wage-relevant marriage shocks, panel fixed effects that soak up stable U, double-machine learning that orthogonalises the gender regressor) are ways to keep both edges under control.

Empirical Illustration using `hdm`

Introducing the `hdm` R 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.

[*] A Stata module named `Lassopack` offers a comprehensive set of programs for regularized regression in high-dimensional contexts..]

Illustration: Testing for Growth Convergence

The standard empirical model for growth convergence is represented by the equation:

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

where

- $Y_{i,T}$ national growth rates in GDP per capita for the periods 1965-1975 and 1975-1985.
- $Y_{i,0}$ is the log of the initial level of GDP at the beginning of the specified decade.
- X_{ij} covariates which might influence growth.

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

Growth Data

To test the growth convergence hypothesis, we will employ the Barro and Lee (1994) dataset.

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

The data features macroeconomic information for a substantial group of countries over various decades. Specifically,

- n equals 90 countries
- k equals 60 country features

While these numbers may not seem large, the quantity of covariates is substantial compared to the sample size. Hence, variable selection is crucial!

Let's Have a Look

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

```
## # A tibble: 2 x 63
##   Outcome intercept gdpsh465 bmp1l freeop freetar   h65  hm65  hf65   p65  pm65  pf65   s65
##     <dbl>     <int>    <dbl> <dbl>  <dbl>   <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  0.29  0.37  0.21  0.04
## 2  0.100          1     6.83 0.614  0.314  0.0618 0.019 0.032 0.007  0.91  1     0.65  0.16
## # i 50 more variables: 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>, ..
```

Data Processing

Rename the response and "treatment" variables:

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

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

```
YT <- df %>% dplyr::select(YT) %>% pull()

Y0 <- df %>% dplyr::select(Y0) %>% pull()

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

Y0_X <- df %>%
  dplyr::select(-YT) %>%
  as.matrix()
```

Estimation of the Convergence Parameter α_1

Method 1: OLS

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

Method 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

Method 3: Partialling-out Lasso

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

Method 4: Double-selection Lasso

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

Tidying the Results

```
# OLS
ols_tbl <- tidy(ols) %>%
  dplyr::filter(term == "Y0") %>%
  dplyr::mutate(method = "OLS") %>%
  dplyr::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-selection 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])
```

Summary of the Convergence Test

```
bind_rows(ols_tbl, naive_Lasso_tbl, part_Lasso_tbl, double_Lasso_tbl) %>%
  kable(digits = 3, format = "html")
```

method	estimate	std.error
OLS	-0.009	0.030
Naive Lasso	NA	NA
Partialling-out Lasso	-0.050	0.014
Double-selection Lasso	-0.050	0.016

The use of double-selection and partialling-out methods lead to significantly more precise estimates and lend support to the conditional convergence hypothesis.

An Advanced R Package: DoubleML

- The Python and R packages {DoubleML} offer a modern implementation of the double / debiased machine learning framework.

- For more details, visit the Getting Started and Examples sections.

- The package is constructed on the {mlr3} ecosystem.

```
slides |> end()
```

Source code

Selected References

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2019). lassopack: Model selection and prediction with regularized regression in Stata.

Angrist, Joshua D, and Alan B Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106(4): 979–1014.

Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, 40(S1), S97-S140.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica* 80(6): 2369–2429.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547.

Belloni, A., Chernozhukov, V., & Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608–650.

Selected References

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. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5), 486–490.

Chernozhukov, V., Hansen, C., & Spindler, M. (2016). hdm: High-Dimensional Metrics. *The R Journal*, 8(2), 185–199.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5), 261–265.

Selected References

Dale, Stacy Berg, and Alan B Krueger. 2002. "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables." *The Quarterly Journal of Economics*, 117(4): 1491–1527.

Hünermund, P., Louw, B., & Caspi, I. (2023). Double machi*ne learning and automated confounder selection: A cautionary tale. _Journal of Causal Inference*, 11(1), 20220078.

Mullainathan, S. & Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87–106.

Van de Geer, S. A., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3, 1360–1392.

Zhao, P., & Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541–2563.