

Machine Learning: An Applied Econometric Approach

Jann Spiess

based on work with Sendhil Mullainathan

in collaboration with Susan Athey and Niall Keleher

3. Prediction vs Estimation

Structure of first chapter of webinar

1. Introduction



2. The Secret Sauce of Machine Learning

3. Prediction vs Estimation

ML basics recap

1. Flexible functional forms
 2. Limit expressiveness (regularization)
 3. Learn how much to regularize (tuning)
-
- What do these features imply for the properties of \hat{f} ?
 - And how can we therefore use \hat{f} in applied work?

Prediction problem set-up

Given:

- Training data set $(y_1, x_1), \dots, (y_n, x_n)$ (assume iid)
 - Usually called “regression” when y continuous, “classification” when y discrete
- Loss function $\ell(\hat{y}, y)$

Goal:

- Prediction function \hat{f} with low average loss (“risk”)
$$L(\hat{f}) = E_{(y,x)}[\ell(\hat{f}(x), y)]$$
where (y, x) distributed same as training

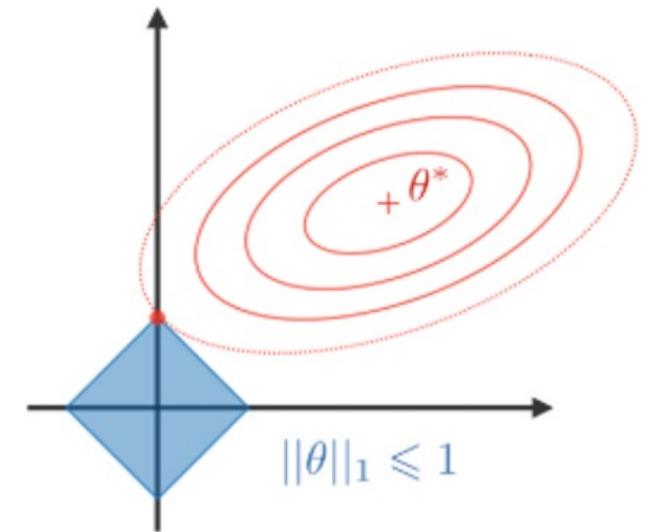
What can we learn about the world from \hat{f} ?

- Often interested in inference on $f(x) = E[y|x]$
→ what can we learn from ML output \hat{f} ?
- Particularly tempting when output has common form
$$\hat{f}(x) = \hat{\beta}'x$$
- Unbiasedness
- Consistency
- Inference: asymptotic Normality, standard errors, tests
- Robustness

LASSO regression

$$\min_{\hat{\beta}} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}' x_i)^2 + \lambda \sum_{j=1}^k |\hat{\beta}_j|$$

- Selects *and* shrinks
- “Capitalist” – in doubt give all to one
- Produces sparse solutions



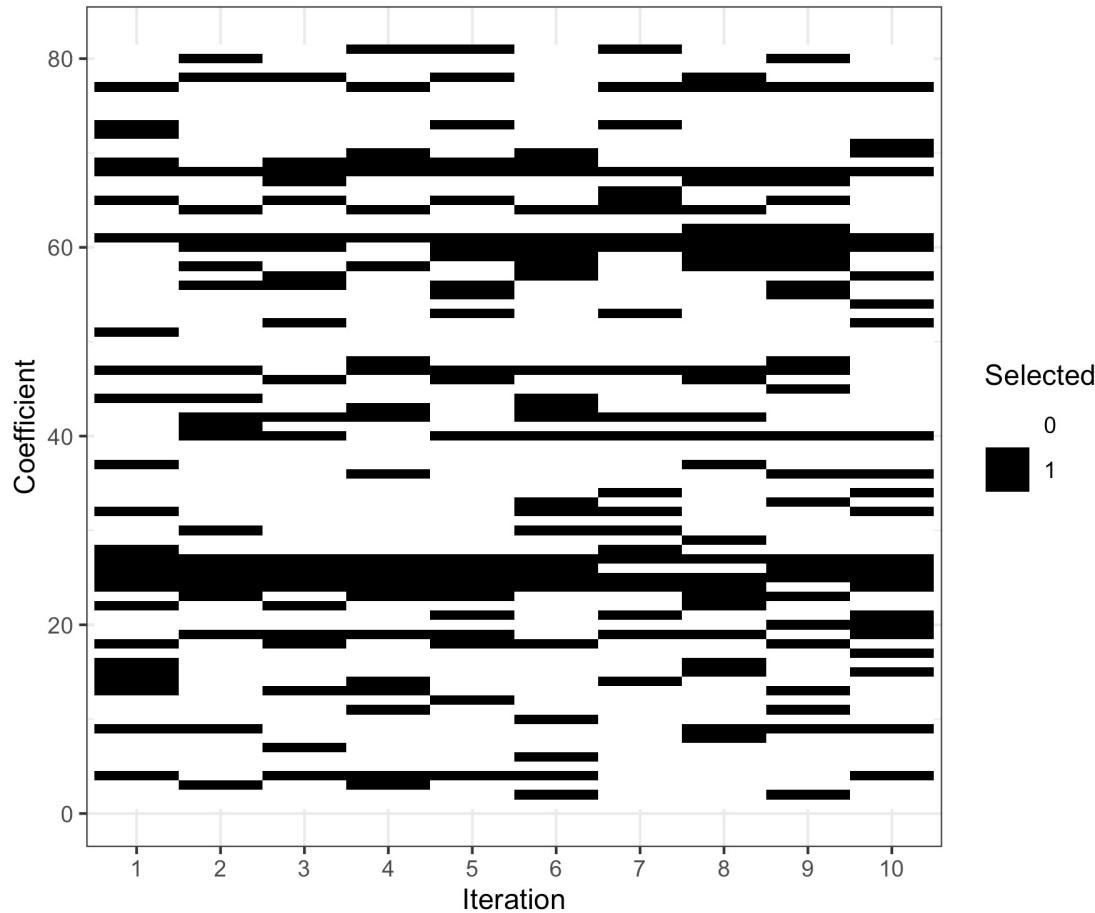
From OLS to LASSO

$$\begin{aligned} \text{default} = & \alpha + \beta_1 \text{ income} + \beta_2 \text{ age} \\ & + \beta_3 \text{ education} + \beta_4 \text{ creditscore} \\ & + \beta_5 x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80} \end{aligned}$$

LASSO selection

$$\begin{aligned} \text{default} = & \alpha + \beta_1 \text{ income} + \beta_2 \text{ age} \\ & + \beta_3 \text{ education} + \beta_4 \text{ creditscore} \\ & + \beta_5 x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80} \end{aligned}$$

LASSO selection



LASSO selection

$$\begin{aligned} \text{default} = & \alpha + \beta_1 \text{ income} + \beta_2 \text{ age} \\ & + \beta_3 \text{ education} + \beta_4 \text{ creditscore} \\ & + \beta_5 x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80} \end{aligned}$$

LASSO selection

$$\begin{aligned} \text{default} = & \alpha + \beta_1 \text{ income} + \beta_2 \text{ age} \\ & + \beta_3 \text{ education} + \beta_4 \text{ creditscore} \\ & + \beta_5 x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80} \end{aligned}$$

LASSO selection

Endogeneous???



Endogeneous???

- potential endogeneity problem



$$\begin{aligned} \text{default} = & \alpha + \beta_1 \text{ income} + \cancel{\beta_2 \text{ age}} \\ & + \beta_3 \text{ education} + \beta_4 \text{ creditscore} \\ & + \beta_5 x_5 + \dots + \beta_{27} x_{27} + \dots + \beta_{80} x_{80} \end{aligned}$$

$$\text{age} \approx f(\text{income}, \text{creditscore}, \dots, x_{27}, \dots)$$

LASSO biases

Fit LASSO with x_1, x_2, x_3 on

$$y = 2x_2 - x_3 + \epsilon$$

x_3 = crucial factors in the equation
If not selected, beta = biased due to omitted var

Selection biases:

- $\rho(x_2, x_3)$ large $\rightarrow \hat{\beta}_3 = 0$ (compactification bias)
- $\rho(x_1, 2x_2 - x_3)$ large $\rightarrow \hat{\beta}_1 \neq 0$ (expansion bias)

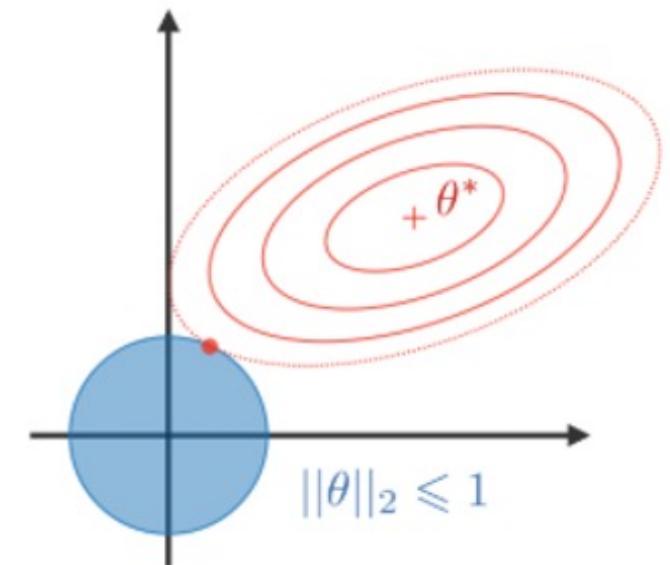
Size biases:

- x_3 not selected $\rightarrow \hat{\beta}_2$ biased (omitted variable bias)
- Even if x_2, x_3 selected, biased toward zero (shrinkage bias)
- In high dimensions, correlations (empirically) ubiquitous

Ridge regression

$$\min_{\hat{\beta}} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}' x_i)^2 + \lambda \sum_{j=1}^k \hat{\beta}_j^2$$

- Shrink towards zero, but never quite
- “Socialist” – in doubt distribute to multiple
- Can be interpreted as Bayesian posterior



Ridge biases

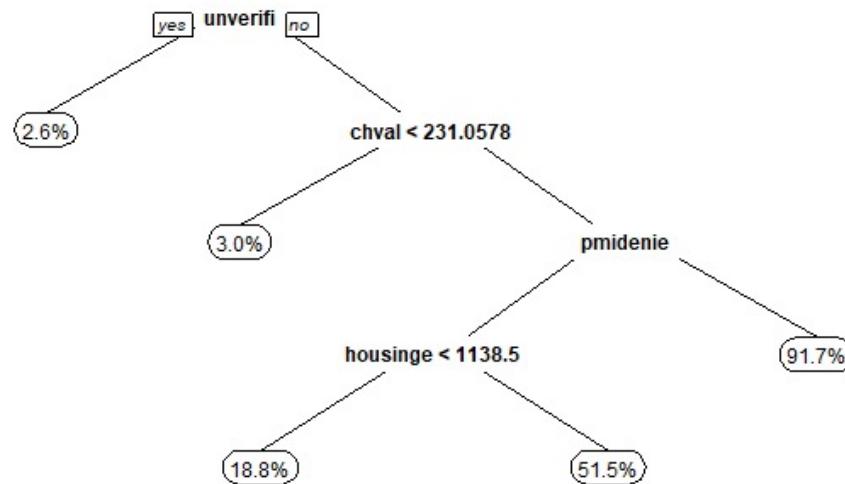
Fit ridge with x_1, x_2 with correlation ρ on

$$y = \beta_2 x_2$$

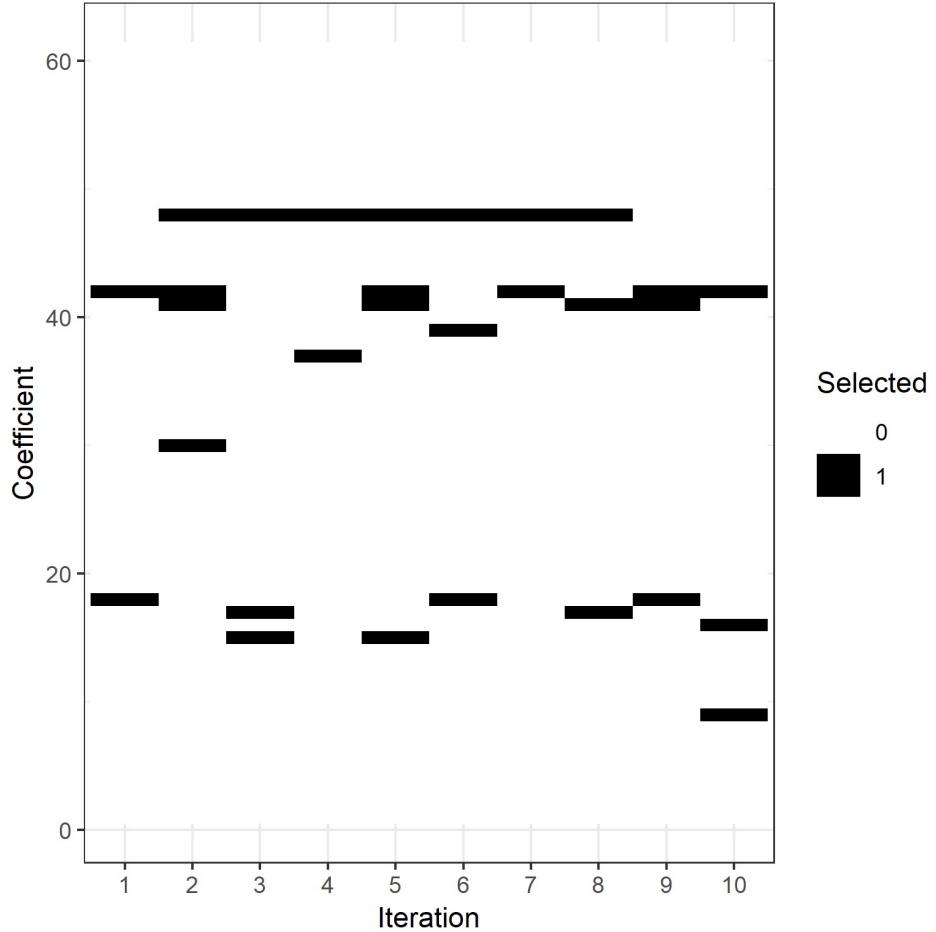
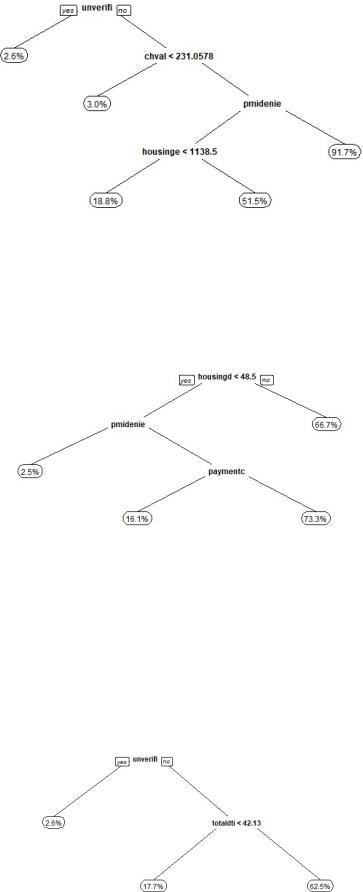
$$\rightarrow \hat{\beta}_1 = \frac{\rho\lambda\beta_2}{(1+\lambda^2)-\rho^2}, \quad \hat{\beta}_2 = \frac{(1+\lambda-\rho^2)\beta_2}{(1+\lambda^2)-\rho^2}$$

(expansion bias) (shrinkage bias)

Tree estimation



Instability, inconsistency



\hat{y} vs $\hat{\beta}$

- Prediction (\hat{y}): care about out-of-sample loss min
- Estimation ($\hat{\beta}$): inference on coefficients
- In high dimensions, can have good predictions even when coefficients unstable, biased, inconsistent
- In big data, many functions that look different can have similar prediction properties, distinguishing hard
- The very features (complexity, regularization, tuning) that make prediction successful make estimation hard

What I mean by prediction (\hat{y})

- A sense in which ML does *not* deliver prediction:
predict what happens under alternative policy
- Counterfactual requires structural/causal knowledge
- I mean: good fit of \hat{y} to y on *same* distribution

What I mean by estimation ($\hat{\beta}$)

- A sense in which ML *does* deliver estimation:
 $\hat{f}(x)$ gets close to $f(x) = E[y|x]$ for minimal loss
- But only in loss norm, say $E_x (\hat{f}(x) - f(x))^2$
 - consistency not invariant to distribution of x
- I mean: estimation consistency $\hat{\beta} \rightarrow \beta$

Take-aways for applied work

- ML provides quality predictions \hat{y}
- The prediction quality of given \hat{f} comes with guarantees from hold-out
- Typically no estimation ($\hat{\beta}$) consistency
- Hence, by itself no structural interpretation or counterfactual extrapolation (causal inference)
- As a side note, inference hard (bootstrap may fail)

Application areas for a \hat{y} tool

1. Data pre-processing
2. \hat{y} tasks (prediction policy problems)
3. \hat{y} in the service of $\hat{\beta}$

Structure of first chapter of webinar

1. Introduction



2. The Secret Sauce of Machine Learning (getting good \hat{y})

3. Prediction vs Estimation (\hat{y} vs $\hat{\beta}$)

Thank you!

jspiess@stanford.edu