De l'économétrie au machine learning, quelles conséquences pour l'évaluation des politiques publiques ?

Emmanuel Flachaire

Aix-Marseille Université, AMSE

Big data and Machine Learning

Ex: Google index, spam, netflix, amazon, bank, cv, kamikazes, cancers ...

1. Introduction
   - General principle
   - Ridge and Lasso
   - Random Forest, Boosting, Deep learning
2. Misspecification
   - Detection of misspecification
   - Interpretable machine learning
3. Causal inference
   - Average treatment effects
   - Detection and analysis of heterogeneity

General Principle: optimization problem

Find the solution $\widehat{m}$ to the optimization problem:

$$\underset{m}{\text{Minimize}} \sum_{i=1}^{n} \mathcal{L}(y_i, m(X_i)) \quad \text{subject to} \quad \|m\|_{\ell_q} \leq t \qquad (1)$$

which can be rewritten in Lagrangian form, for some $\lambda \geq 0$:

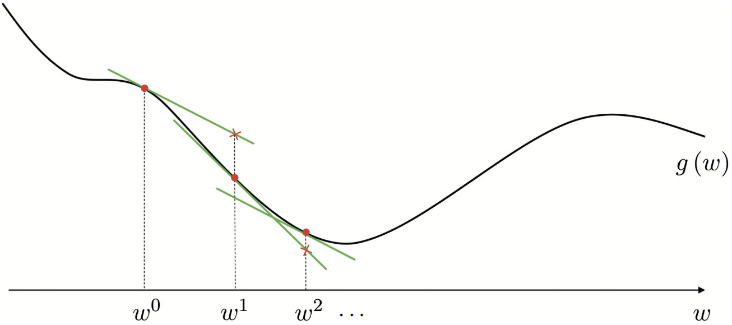
$$\underset{m}{\text{Minimize}} \sum_{i=1}^{n} \underbrace{\mathcal{L}(y_i, m(X_i))}_{\text{loss function}} + \underbrace{\lambda \|m\|_{\ell_q}}_{\text{penalization}} \qquad (2)$$

- The goal is to minimize a loss function under constraint
- It is usually done by numerical optimization

General Principle: resolution by numerical optimization

Gradient Descent

Use *linear* approximations at each steps, from Taylor expansion



(Source: Watt et al., 2016)

Algorithm: Gradient descent

Input: differentiable function g, fixed step length α, initial point x^0

Repeat until stopping condition is met: $w^k = w^{k-1} - \alpha\, g'(w^{k-1})$

Linear regression

$$\underset{m}{\text{Minimize}} \sum_{i=1}^{n} \underbrace{\mathcal{L}(y_i, m(X_i))}_{\text{loss function}} + \underbrace{\lambda \, \|m\|_{\ell_q}}_{\text{penalization}}$$

Let us consider:

- Euclidian distance: $\mathcal{L}(y_i, m(X_i)) = (y_i - m(X_i))^2$
- m is a linear function of parameters: $y_i \approx X_i\boldsymbol{\beta}$ with $\boldsymbol{\beta} \in R^p$
- no penalization: $\lambda = 0$

Thus, we have:

$$\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

It is the minimization of the SSR in a linear regression $\rightarrow \widehat{\boldsymbol{\beta}}_{OLS}$

General Principle

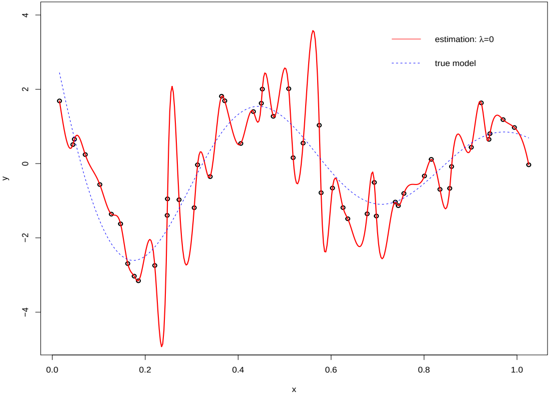
Machine Learning: solve the optimization problem

$$\underset{m}{\text{Minimize}} \sum_{i=1}^{n} \underbrace{\mathcal{L}(y_i, m(X_i))}_{\text{loss function}} + \underbrace{\lambda \, \|m\|_{\ell_q}}_{\text{penalization}}$$

- Choice of the loss function:
  - $\mathcal{L} \to$ conditional mean, quantiles, classification
  - $m \to$ linear, splines, tree-based models, neural networks
- Choice of the penalization:
  - $\ell_q \to$ lasso, ridge
  - $\lambda \to$ over-fitting, under-fitting, cross validation

Over-fitting

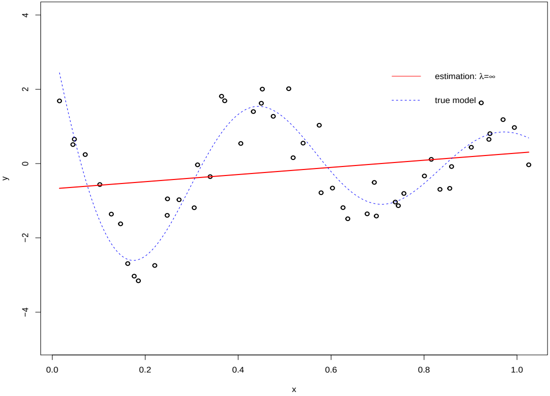
A model with high flexibility may fit perfectly observations used for estimation, but very poorly new observations



→ penalization: put a price to pay for having a more flexible model

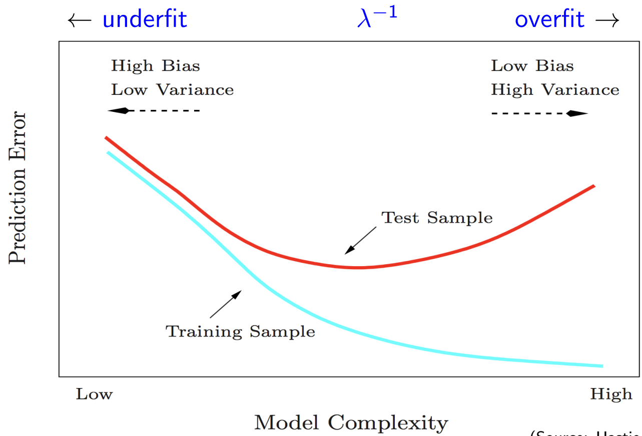
Under-fitting

If we put a huge cost for a more complex model, $\lambda = \infty$, we obtain a linear regression model



→ if the cost is too large: low variance, but high bias

Do not train and evaluate the model with the same sample



(Source: Hastie et al., 2009)

Underfitting: the model performs poorly on training and test samples

Overfitting: performs well on training sample, but generalizes poorly on test sample

→ Control overfitting with MSE computed out-sample by CV

Ridge and Lasso

$$\underset{\beta}{\text{Minimize}} \quad \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=2}^{p} |\beta_j|^q$$

It is equivalent to minimize SSR subject to $\sum_{j=2}^{p} |\beta_j|^q \leq c$

- The constraint restricts the magnitude of the coefficients
- It shrinks the coefficients towards zero as $c \searrow$ (or $\lambda \nearrow$)
- Add some bias if it leads to a substantial decrease in variance
- $q = 2$: Ridge, $\hat{\beta} = (X^\top X + \lambda \mathbb{I}_n)^{-1} X^\top y$ is defined with $p \gg n$
- $q = 1$: Lasso sets some coef exactly to 0, variable selection

$\rightarrow$ High-dimensional problems ($p \gg n$)

Random Forest, Boosting, Deep learning

$$\underset{m}{\text{Minimize}} \quad \sum_{i=1}^{n} (y_i - m(X_i))^2 + \lambda \int m''(x)^2 dx$$

It is equivalent to minimize SSR subject to $\int m''(x)^2 dx \leq c$

- A fully nonparametric model: $y \approx m(X_1, \ldots, X_p)$
- The constraint restricts the flexibility of m
- Choice of m: Random forest, boosting or deep learning
- Similar to nonparametric econometrics (splines)
- Appropriate with many covariates (no curse of dimensionality)

$\rightarrow$ Complex functional form

Why and how to use ML methods in Econometrics?

Pros:

- High-dimensional problems
- Complex functional forms

However,

- Black-box models
- Prediction is not causation[1]

[1] Kleinberg et al. (2015) Prediction policy problems, Athey (2017) Beyond prediction: Using big data for policy problems

Misspecification

ML models outperform parametric econometric models

- Many results report that ML outperform parametric models in terms of predictive performance
- Boston housing dataset:[2]

$\widehat{\mathcal{R}}^{10-\text{CV}}$	OLS	$\text{OLS}_{x^2x^3int}$	R.Forest	Boosting
MSE	23.938	24.079	10.008	9.729

- ML models show impressive improvement in prediction error
- ML models are known to capture complex functional forms
- It suggests that the parametric models miss important nonlinear and/or interaction effects

[2] 14 variables (2 dummies), 78 pairwise interactions, 506 observations

An econometric model for interpretable Machine Learning

Partially linear model:[3]

$$y = g_1(X_1) + \ldots + g_p(X_p) + Z\gamma + \varepsilon$$

with Z a matrix of pairwise interactions $Z = (X_1 X_2, \ldots, X_{q-1} X_q)$. The marginal effect is:

$$\frac{\partial y}{\partial X_j} = g_j'(X_j) + c$$

where c is a constant term which depends on the other covariates.

- Combine non-linearity in X_j and linear pairwise interactions
- The linearity assumption on interaction effects represents the price to pay to keep the model interpretable.
- Estimation: GAM+variable selection (Lasso, Autometrics)

[3] Flachaire, Hacheme, Hué, Laurent (2021)

Parametric models can perform as well as ML models

- Boston housing dataset:

$\widehat{\mathcal{R}}^{10-\text{CV}}$	OLS	R.Forest	Boosting	GAMLA
MSE	23.938	10.008	9.729	9.594

- ML models outperform standard parametric model ... which are not well-specified!
- ML methods can help to detect and correct misspecification in parametric regression

Causal inference

Treatment effects: high-dimensions

Partially linear model

$$y = D\tau + g(X) + \varepsilon$$

- $g(X)$ approx linearly with many controls (2-ways interactions)
- τ variable of interest, $g(X) = Z\gamma$, with $Z = [X, X{:}X]$
- Post-Lasso: inference is valid if perfect selection achieved only
- Concern: wrong exclusion of variables (omitted variable bias)
- Double Lasso: least squares after double selection[4]
  1. Lasso of y on Z: select variables important to predict y
  2. Lasso of D on Z: select variables correlated with the treatment

  OLS of y on D and the union of the selected variables

  $\rightarrow$ valid post-selection inference in high-dimensions

[4] Belloni, Chernozhukov and Hansen (2014): uniformly valid confidence set for τ despite imperfect model selection, and full efficiency for estimating τ

Heterogeneous treatment effects: high-dimensions

Heterogeneity

$$y = D\tau(X) + g(X) + \varepsilon$$

- $\tau(X)$ is a parametric function of X: e.g. $\tau(X) = X\beta$
- $g(X) = Z\gamma$, approximated linearly with 2-ways interactions
- Double Lasso: least squares after double selection
  1. Lasso of y on Z: select variables important to predict y
  2. Lasso of each component of DX on the other regressors

  OLS of y on D and the union of the selected variables
- Bach, Chernozhukov and Spindler (2021) Closing the U.S. gender wage gap requires understanding its heterogeneity

$\rightarrow$ assess heterogeneity with many determinants

Heterogeneous treatment effects: fully nonparametric

Interactive model

$$y = m(D, X) + \varepsilon$$
$$d = h(X) + \eta$$

- ATE: parameter of interest, m(.) and h(.): nuisance functions
- Double Machine Learning:[5]
  1. Neyman orthogonal condition (double residuals, FWL)
  2. Cross-fitting: ATE and m, h estimated from $\neq$ samples
  3. Doubly robust: AIPW robust to misspecification of m or h

  AIPW estimator based on ML estimation of m and h

  $\rightarrow$ ATE estimation and inference with good properties[6]
- No detection and analysis of heterogeneity

[5] Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018)

[6] $\sqrt{n}$-consistent and asymp Normal even if nuisance functions $n^{1/4}$-consistent

Detection and analysis of heterogeneity

Generic Machine Learning:[7]

- Do not attempt to get valid estimation and inference on the CATE itself, but on features of the CATE
- Obtain ML proxy predictor of CATE (auxiliary set) and target features of CATE based on this proxy predictor (main set)

Main interests:

- Test if there is evidence of heterogeneity (BLP)
- ATE for the 20% most (least) affected individuals? (GATES)
- Which covariates are associated to TE heterogeneity? (CLAN)

→ valid estimation and inference on *features* of CATE

[7] Chernozhukov, Demirer, Duflo and Fernàndez-Val (2020)

Detection and analysis of heterogeneity

Causal Random Forest:[8]

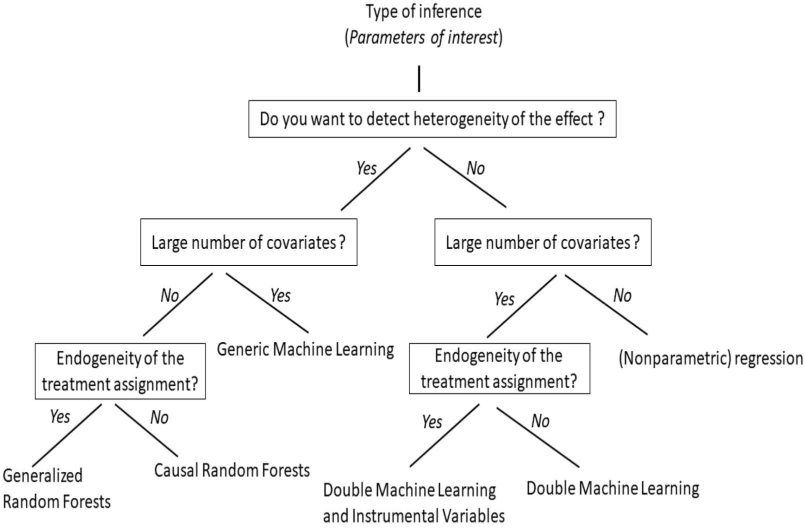
- Random Forest is modified to estimate the CATE directly
- Grow a tree and evaluate its performance based on TE heterogeneity rather than predictive accuracy
- The idea is to find leaves where the treatment effect is constant but different from other leaves
- Split criterion: maximize heterogeneity in TE between leaves
- Honest tree: build tree and estimate CATE from $\neq$ samples

  $\rightarrow$ valid estimation and confidence intervals for CATE[9]

---

[8] Wager and Athey (2018), Athey, Tibshirani and Wager (2019)
[9] RF predictions are asymp unbiased and Gaussian, but cv rates below $\sqrt{n}$

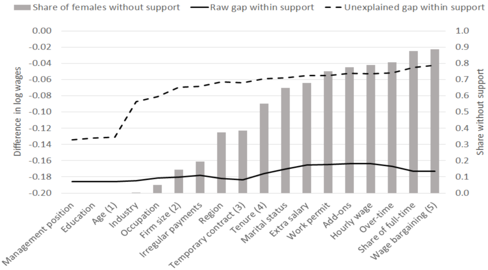
Causal Machine Learning: A brief roadmap



Source: Gaillac and L'Hour (2021)

Underlying assumptions

- Standard hypotheses: SUTVA, CIA and CSC
- Common support condition (CSC): $0 < P(d_i = 1|X_i = x) < 1$
  - ML estimation often provides better predictions
  - Adding covariates makes matching more difficult



Strittmatter and Wunsch (2021) The gender pay gap revisited with big data: Do methodological choices matter?

- Trimming in experiments vs. decomposition methods

$\rightarrow$ Beware of CSC when moving away from RCT framework

Conclusion

The impact of ML for public policy evaluation:

- Dealing with many covariates ($p \gg n$)
- Relying less on a priori specification
- Take care of heterogeneity
- However, do not forget underlying assumptions! (CSC)

Technical literature, where implementation becomes easier

- Double Lasso: R package `hdm`
- Double Machine Learning: R package `DoubleML`
- Generalized Random Forest: R package `grf`
- Generic Machine Learning: R package `GenericML`

An effervescent empirical and theoretical literature

Selected references

- Athey (2017) Beyond prediction: Using big data for policy problems, Science
- Athey (2018) The impact of machine learning on economics
- Athey, Tibshirani and Wager (2019) Generalized random forest, Ann. Statis.
- Bach, Chernozhukov and Spindler (2021) Closing the U.S. gender wage gap requires understanding its heterogeneity, arXiv:1812.04345
- Belloni, Chernozhukov and Hansen (2014) Inference on treatment effects after selection amongst high-dimensional controls, REStud
- Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018) Double/debiased ML for treatment and structural parameters. Econometrics J.
- Chernozhukov, Demirer, Duflo and Fernàndez-Val (2020) Generic ML inference on heterogenous treatment effects in randomized experiments, arXiv:1712.04802
- Gaillac and L'Hour (2020) Machine Learning for Econometrics, Lecture notes
- Kleinberg, Ludwig, Mullainathan and Obermeyer (2015) Prediction Policy Problems, AER P&P
- L'Hour (2020), L'économétrie en grande dimension. INSEE M2020-01
- Strittmatter (2020) What is the value added by using causal machine learning methods in a welfare experiment evaluation.
- Strittmatter and Wunsch (2021) The gender pay gap revisited with big data: Do methodological choices matter? arXiv:2102.09207
- Wager and Athey (2018) Estimation and inference of heterogeneous treatment effects using random forests. JASA