Machine Learning Methods That Economists Should Know About

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Big Data Economics

Outline I

Introduction to Statistical Paradigms

Breiman's Insight on Statistical Cultures

- **Model-Based Approach**: Assumes data from a known stochastic model.
- **Algorithmic Approach**: Treats the data mechanism as unknown, suitable for complex/large datasets.

Evolution in Statistics

- **Past**: Strong preference for data models, leading to "irrelevant theory" (Breiman, 2001, *Statistical Modeling: The Two Cultures*).
- **Now**: Embrace of Machine Learning (ML) as part of standard statistical tools.

Economics & Econometrics

- **Adoption**: Slower integration of ML methods in economics.
- **Advocacy**: Urging economists to diversify their toolkit beyond traditional data models.

The Case for ML in Economics

Why ML in Economics?

- **Data Complexity**: ML excels in "big data" contexts with vast units and variables.
- **Cultural Shift**: Economics journals now recognize the need for ML's empirical efficiency.

Challenges and Adaptations

- **Formal Properties**: Traditional emphasis on estimators' large-sample properties vs. ML's algorithmic performance focus.
- **Customization**: Requirement for tuning ML techniques to economic problems, exploiting structural insights from econometrics.

Future Directions

- **Method Inclusion**: Nonparametric regression, classification, clustering, treatment effects, experimental design, matrix completion, and text analysis.
- **Pedagogical Integration**: Updating graduate curricula to include ML tools alongside traditional methods.

References for Further Reading

- **Economics Focus**: Varian (2014), Mullainathan & Spiess (2017), Athey (2017)
- **Comprehensive Texts**: Hastie et al. (2009), Efron & Hastie (2016), Burkov (2019), Alpaydin (2009), Knox (2018)

Econometric Approaches

Traditional Econometrics:

- Focus on specifying a target or estimand.
- Estimand: a function of the joint distribution of data.
- Often a parameter of a statistical model.
- Parameters estimated via objective functions like sum of squared errors or likelihood function.
- Emphasis on estimator quality and large sample efficiency.
- Construction of confidence intervals, reporting of point estimates and standard errors.

Machine Learning Approaches

Focus in Machine Learning:

- Development of algorithms for prediction and classification.
- Example: Wu et al. (2008) "Top 10 Algorithms in Data Mining".
- Algorithms predict variables or classify units based on features.
- Different performance metrics, not necessarily tied to estimand properties.

Econometrics vs. Machine Learning

Least Squares in Econometrics:

- Model: $Y_i|X_i \sim N(\alpha + \beta^\top X_i, \sigma^2)$.
- Least squares used to estimate $\theta = (\alpha, \beta)$.
- Properties if the model is correct:
  - Unbiased
  - Best linear unbiased estimator (BLUE)
  - Maximum likelihood estimator (MLE)
  - Large sample efficiency

Prediction in Machine Learning

Prediction Focus in ML:

- Predict outcome Y_{N+1} for a new unit based on X_{N+1}.
- Restrict to linear predictors: $\hat{Y}_{N+1} = \hat{\alpha} + \hat{\beta}^\top X_{N+1}$.
- Loss function: $(Y_{N+1} - \hat{Y}_{N+1})^2$.
- Goal: Estimators with good properties related to the loss function.
- Least squares not necessarily the best choice when feature dimension exceeds two.
- Decision theory suggests other estimators can dominate least squares.

Machine Learning Terminology in Econometrics

Training Sample: Sample used to estimate parameters.
Model Training: Estimation of the model's parameters.

Features: Known as regressors, covariates, or predictors in econometrics.
Weights: Regression parameters or coefficients.

Supervised Learning: Observing both predictors (features) and outcomes.
Unsupervised Learning: Only predictors are observed; tasks include clustering or distribution estimation.
Classification: Refers to unordered discrete response problems.

Validation and Cross-Validation

Traditional Econometric Focus:

- Linear regression models given by economic theory.
- Emphasis on parameter estimation efficiency.
- Large sample efficiency as a key criterion.
- Model selection often via hypothesis testing.

Predictive Perspective:

- Interest in prediction for new, similar population units.
- Model with intercept vs. model with intercept and scalar X_i.
- Consideration of β close to zero leading to simplification.

Out-of-Sample Cross-Validation:

- Guides model selection decisions.
- Focus on predictive power, not solely on parameter estimation.
- Utilizes out-of-sample data to ensure unbiased fit comparison.

Overfitting in Machine Learning

- Overfitting is a primary concern in ML, unlike traditional econometrics.
- Emphasis on model selection that balances fit and out-of-sample prediction.
- Regularization is a fundamental concept to prevent overfitting.
- Vapnik: Regularization theory was one of the first signs of intelligent inference.

Regularization and Model Complexity

- Regularization penalizes model complexity to improve prediction.
- Complexity measured by number of parameters or VC dimension.
- Traditional econometrics used similar concepts like AIC or BIC.
- Modern ML uses data-driven approaches for regularization.

Regularization in Practice

- Regularization in regression: Adding penalty terms to objective function.
- Bayesian approaches use priors to regularize parameter estimates.
- ML methods tune regularization via cross-validation for predictive performance.

Example: Regularization in Linear Regression

- Consider the linear regression model with K regressors:

$$Y_i|X_i \sim \mathcal{N}(\beta X_i, \sigma^2)$$

- Prior for slope coefficients β_k:

$$\beta_k \sim \mathcal{N}(0, \tau^2)$$

- Posterior mean for β with prior variance τ^2:

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{N} (Y_i - \beta X_i)^2 + \frac{\sigma^2}{\tau^2} \|\beta\|_2^2 \right\}$$

- ML approach with penalty parameter λ:

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{N} (Y_i - \beta X_i)^2 + \lambda \|\beta\|_2^2 \right\}$$

  λ chosen by cross-validation to optimize predictive performance.

The Challenge of High-Dimensional Data

- High feature-to-sample size ratios are common in ML.
- Many features are suspected to be marginally important.
- **Key Challenge:** Identifying impactful features without prior knowledge.

The Sparsity Principle:

- Assumes a sparse true underlying model.
- L_1 penalty (Lasso) is leveraged for feature selection.

Implications and Practices of Sparsity

- **Effective Recovery:** If sparsity holds, L_1 penalty can accurately identify relevant signals.
- **In the Absence of Sparsity:** No selection method may substantially outperform others.
- Approximate sparsity is often adequatefew features have strong explanatory power.
- Traditional social science research often limited explanatory variables *manually*.
- Data-driven selection is seen as an improvement but comes with strong assumptions.
- **Inference Challenge:** Making reliable inferences with data-dependent selection.

Computational Issues and Scalability

- Machine Learning (ML) emphasizes computationally efficient methods.
- Large datasets necessitate scalable solutions.
- Preference for methods that balance statistical efficiency with computational practicality.

LASSO vs Subset Selection in Linear Regression

- Subset selection identifies a subset of regressors, using least squares for estimation.
- LASSO (Least Absolute Shrinkage and Selection Operator) adds a penalty term to the regression, aiding in feature selection and regularization.
- LASSO scales well with large datasets, capable of handling millions of features.
- Best subset selection is NP-hard, traditionally limited to around 30 regressors, though recent research extends this to the 1000s.

Current Debates and LASSO Performance

- Ongoing debate on LASSO vs best subset selection where both are computationally feasible.
- LASSO may perform better in low signal-to-noise ratio settings, common in social science.
- In many social science problems, computational concerns are secondary to the substantive questions.

Stochastic Gradient Descent (SGD)

- SGD is a key optimization tool in ML for parameter estimation.
- It minimizes an empirical loss function for parameter θ.
- Classic gradient descent updates parameters iteratively:

$$\theta^{(k)} = \theta^{(k-1)} - \eta_k \frac{1}{N} \sum_i \nabla Q_i(\theta^{(k)})$$

- This can be expensive for large N.

Implementing SGD in Practice

- SGD takes many small, noisy steps towards the minimum:

$$\theta^{(k)} = \theta^{(k-1)} - \eta_k \frac{1}{\frac{N}{B}} \sum_{i:B_i=k} \nabla Q_i(\theta^{(k)})$$

- B_i denotes the batch that observation i belongs to.
- With a decreasing learning rate η_k, SGD converges to the global minimum (convex cases) or a local minimum (non-convex cases).

Advanced SGD Applications

- In cases where $\nabla Q_i(\theta)$ is an expectation, Monte Carlo integration can be applied:

$$\nabla Q_i(\theta) \approx \frac{1}{M} \sum_{m=1}^{M} \nabla q_i(\theta, \omega_m)$$

- M is the number of Monte Carlo draws, ω_m represents the m-th draw.
- This is efficient for evaluating gradients with fewer or single Monte Carlo draws.
- Used in economic applications (Ruiz et al. 2017, Hartford et al. 2016).

Ensemble Methods: An Introduction

- Ensemble methods enhance prediction accuracy by combining multiple models.
- Model averaging, a common technique, involves weighting and combining different models' predictions.
- Modern ensemble methods may integrate diverse models, unlike traditional methods that average similar ones.
- The primary objective is to optimize out-of-sample performance, rather than just in-sample fit.

Example: Combining Predictive Models

- Consider three models: a random forest ($\hat{Y}_{\text{RF}}$), a neural net ($\hat{Y}_{\text{NN}}$), and a LASSO-estimated linear model ($\hat{Y}_{\text{LASSO}}$).
- The ensemble prediction is a weighted sum of these models.
- Weights are chosen to minimize prediction error on a validation or test dataset.

Optimizing Weights for Model Averaging

- The weights $(p_{\text{rf}}, p_{\text{nn}}, p_{\text{lasso}})$ are obtained by solving the following optimization problem:

$$\begin{aligned} \underset{p_{\text{rf}}, p_{\text{nn}}, p_{\text{lasso}}}{\arg\min} \quad & \sum_{i=1}^{N_{\text{test}}} \left(Y_i - p_{\text{rf}} \hat{Y}_{\text{RF},i} - p_{\text{nn}} \hat{Y}_{\text{NN},i} - p_{\text{lasso}} \hat{Y}_{\text{LASSO},i} \right)^2 \\ \text{subject to} \quad & p_{\text{rf}} + p_{\text{nn}} + p_{\text{lasso}} = 1, \\ & p_{\text{rf}}, p_{\text{nn}}, p_{\text{lasso}} \geq 0. \end{aligned} \tag{1}$$

Advantages of Model Averaging

- Model averaging improves out-of-sample predictions by leveraging diverse model strengths.
- Ensemble methods reduce overfitting risk compared to complex single models.
- Effective against irrelevant features, nonlinearities, and interactions.

Ensemble Methods in Panel Data

- In panel data, ensemble methods can combine synthetic control and matrix completion methods.
- Athey et al. (2019) show that combining these methods can outperform individual approaches.
- The key is to exploit the unique advantages of each method in context.

The Challenge of Inference in Machine Learning

- ML emphasizes out-of-sample performance, often neglecting inferential capabilities.
  - Efron & Hastie (2016) noted that algorithmic progress has outpaced inferential justifications.
  - Advances in inference are noted in areas like **random forests** and **neural networks**.
  - The relevance of inference varies; often, decision-making favors prediction.
  - Inferential efforts may compromise predictive accuracy, evident in kernel regression's bias-variance tradeoff.
- Optimal predictors might exhibit asymptotic bias, impacting confidence intervals.
- Lowering bandwidth reduces inference bias but raises variance and diminishes prediction accuracy.

Linear Regression and Least Squares

- Linear approximation to conditional expectation: $g(x) = \sum_{k=1}^{K} \beta_k x_k$
- **Least Squares Estimation**:

$$\hat{\beta}_{ls} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - \beta X_i)^2$$

- Issues arise when K is large:
  - Poor repeated sampling properties for $\hat{\beta}_{ls,k}$.
  - Least squares estimator can be inadmissible with $K \geq 3$.
  - Particularly poor properties when $K > N$.
- Regularization can improve predictive performance.

Regularization Techniques

- Regularization adds a penalty term to shrink coefficients toward zero:

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{N} (Y_i - \beta^\top X_i)^2 + \lambda \|\beta\|_q^{1/q} \right\}$$

- Different values of q represent different methods:
  - $q = 1 \rightarrow$ LASSO (Least Absolute Shrinkage and Selection Operator)
  - $q = 2 \rightarrow$ Ridge Regression
  - $q \rightarrow 0 \rightarrow$ Best Subset Regression
- Hybrid approaches: Elastic Nets, Relaxed LASSO, etc.

Conceptual Differences and Computational Considerations

- **Sparsity**: LASSO and Best Subset promote sparse solutions; Ridge does not.
- **Computational Feasibility**:
  - Best Subset Regression is NP-hard, challenging for large N and K.
  - LASSO and Ridge are computationally more feasible.
- **Bayesian Interpretation**:
  - Ridge corresponds to normal priors on coefficients.
  - LASSO corresponds to Laplace priors on coefficients.
  - Regularization parameter λ is chosen via cross-validation, not prior beliefs.
- Variable selection in sparse models can be overinterpreted.

Regression Trees I

Overview of Regression Trees

- Objective: Estimate regression functions optimized for out-of-sample predictive power.
- Nature: Non-parametric, suitable for flexible function estimation without subtle tuning.
- Given a sample $(X_{i1}, ..., X_{iK}, Y_i)$ for $i = 1, ..., N$.
- The idea is to recursively split the sample based on a covariate X_{ik} exceeding a threshold c.

Regression Trees II

Mathematical Formulation

- Pre-split in-sample squared error:

$$Q = \sum_{i=1}^{N} (Y_i - \bar{Y})^2,$$

where $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$.

- Post-split in-sample squared error for a split at covariate k and threshold c:

$$Q(k, c) = \sum_{i:X_{ik} \leq c} (Y_i - Y_{k,c,l})^2 + \sum_{i:X_{ik} > c} (Y_i - Y_{k,c,r})^2,$$

where $Y_{k,c,l}$ and $Y_{k,c,r}$ are the average outcomes in the left and right subsamples, respectively.

Optimization and Regularization

Selection of Splits

- Choose covariate k and threshold c that minimize $Q(k, c)$.
- Consider all covariates $k = 1, ..., K$ and all possible thresholds.
- Repeat optimization over the resulting subsamples (leaves).

Avoiding Overfitting

- Regularization: Penalize sum of squared residuals with a term linear in the number of leaves.
- Coefficient of penalty term chosen through cross-validation.
- Grow a very deep tree, then prune to optimal depth using cross-validation.
- This process ensures capturing potential subtle interactions.

Interpretation and Bias

- Easy interpretation: Prediction in each leaf is a sample average.
- Bias: Sample average within a leaf might not be an unbiased estimate for a new test set.
- Athey & Imbens (2016) suggest sample splitting to address bias and provide unbiased estimates for confidence intervals.

Comparison: Kernel Regression vs. Tree-Based Methods I

Understanding Tree-Based Prediction

- A regression tree's prediction for a leaf is the sample average outcome within the leaf.
- Each leaf defines a "neighborhood" of nearest neighbors for a target observation.
- This approach is akin to a matching estimator, with a unique method for selecting neighbors.

Neighborhood Prioritization

- Covariates are prioritized differently, affecting which observations are considered nearby.
- Tree-based neighborhoods are rectangular, not necessarily centered around the target.

Comparison: Kernel Regression vs. Tree-Based Methods II

Differences from Kernel Regression

- Kernel regression uses Euclidean distance for neighborhood creation.
- Tree-based methods create neighborhoods that are less standardized in shape.

Implications for Prediction

- Trees may not always predict optimally for a specific test point.
- For tailored predictions, generalized tree-based methods are preferable.

Comparison: Kernel Regression vs. Tree-Based Methods III

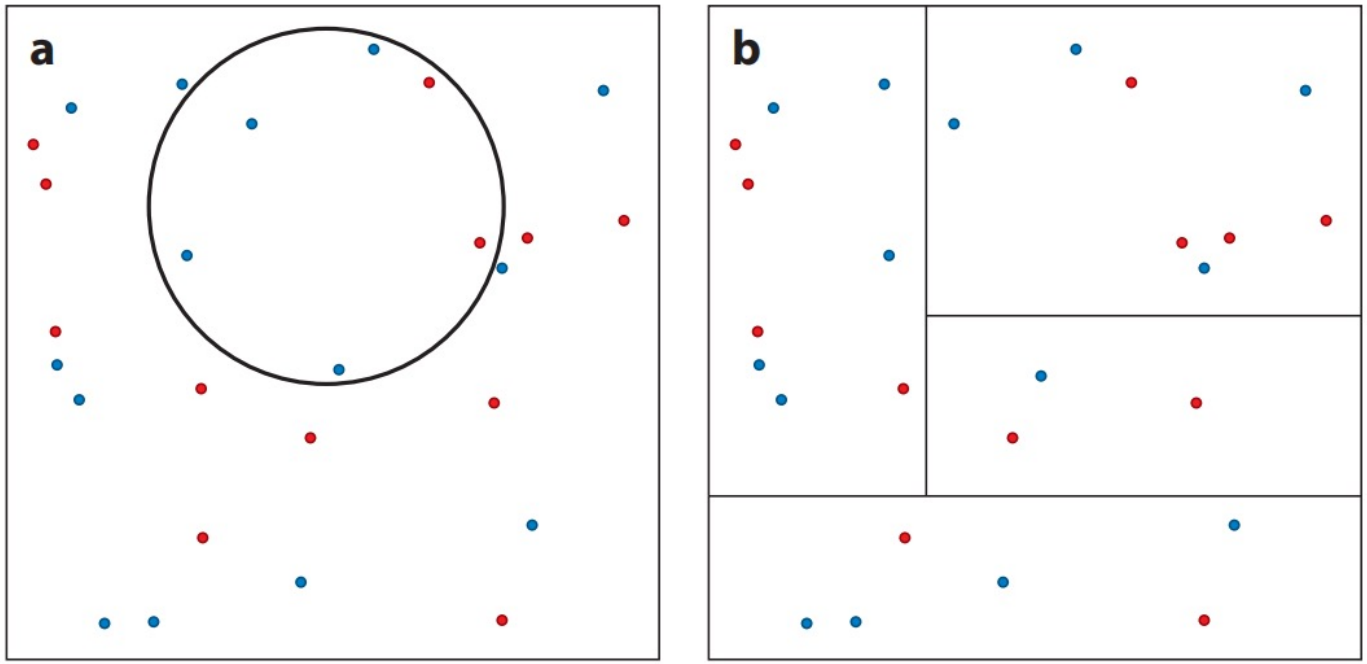

Figure 1

(*a*) Euclidean neighborhood for *k*-nearest neighbor (KNN) matching. (*b*) Tree-based neighborhood.

Introduction to Random Forests

Evolution from Regression Trees

- Random Forests improve on regression trees by addressing discontinuities.
- They introduce smoothness by averaging over numerous trees.

Mechanism of Action

- Each tree uses a bootstrap sample or a subsample of the data (bagging).
- Splits are based on a random subset of covariates, introducing diversity.

Benefits of Random Forests

- Smooth, yet still discontinuous, average with enhanced predictive power.
- Outperform single trees and are less sensitive to irrelevant features.

Random Forests in Practice

Popularity and Performance

- Require minimal tuning and deliver robust performance.
- Particularly effective with sparse datasets containing irrelevant features.

Advantage over Kernel Regression

- Adding non-predictive covariates degrades kernel regression but not random forests.
- Random forests remain efficient by largely ignoring irrelevant covariates.

Statistical Analysis and Advancements

- While statistical analysis of forests was challenging, recent advances have provided insights.
- Particular random forest variants can yield asymptotically normal estimates of $\hat{\mu}(x)$.

Comparison with Kernel Regression

- Random forests can be seen as an average of matching estimators.
- They prioritize important covariates and generate adaptive weighting functions.

Weighting Functions

- Similar to kernel weighting, but with adaptability to covariate importance.
- Prediction at a point x considers nearby points more due to frequent inclusion.

Formulation of Predictions

$$\hat{\mu}_{rf}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x) Y_i}{\sum_{i=1}^{n} \alpha_i(x)},$$

- Where $\alpha_i(x)$ represents the weight given to the ith training example.

Random Forests for Econometric Analysis I

Extension to Causal and Economic Models

- Random forests extended to causal effects and parameters in economic models (Wager & Athey 2017; Athey et al. 2016b).
- Generalized random forests: a two-step algorithm integrating forests with GMM, yielding asymptotically normal parameter estimates.

Local Linear Forests

- Addressing inefficiencies in capturing linear/quadratic effects and boundary bias.
- Local linear regression weighted by forest-derived functions improves predictions.
- Corrects for conditional mean bias near boundaries, enhances performance in multiple dimensions.

Random Forests for Econometric Analysis II

Operationalization and Advantages

- Forests as weighting functions for local maximum likelihood estimation.
- Outperforms traditional forests by capturing broader patterns, ensuring asymptotic normality.
- Efficient in multidimensional datasets, uncovering complex interactions among covariates.

Challenges and Solutions

- Traditional forests struggle with smooth data-generating processes and boundary bias.
- Local linear forests offer a solution through regression corrections and forest-estimated kernel weighting.

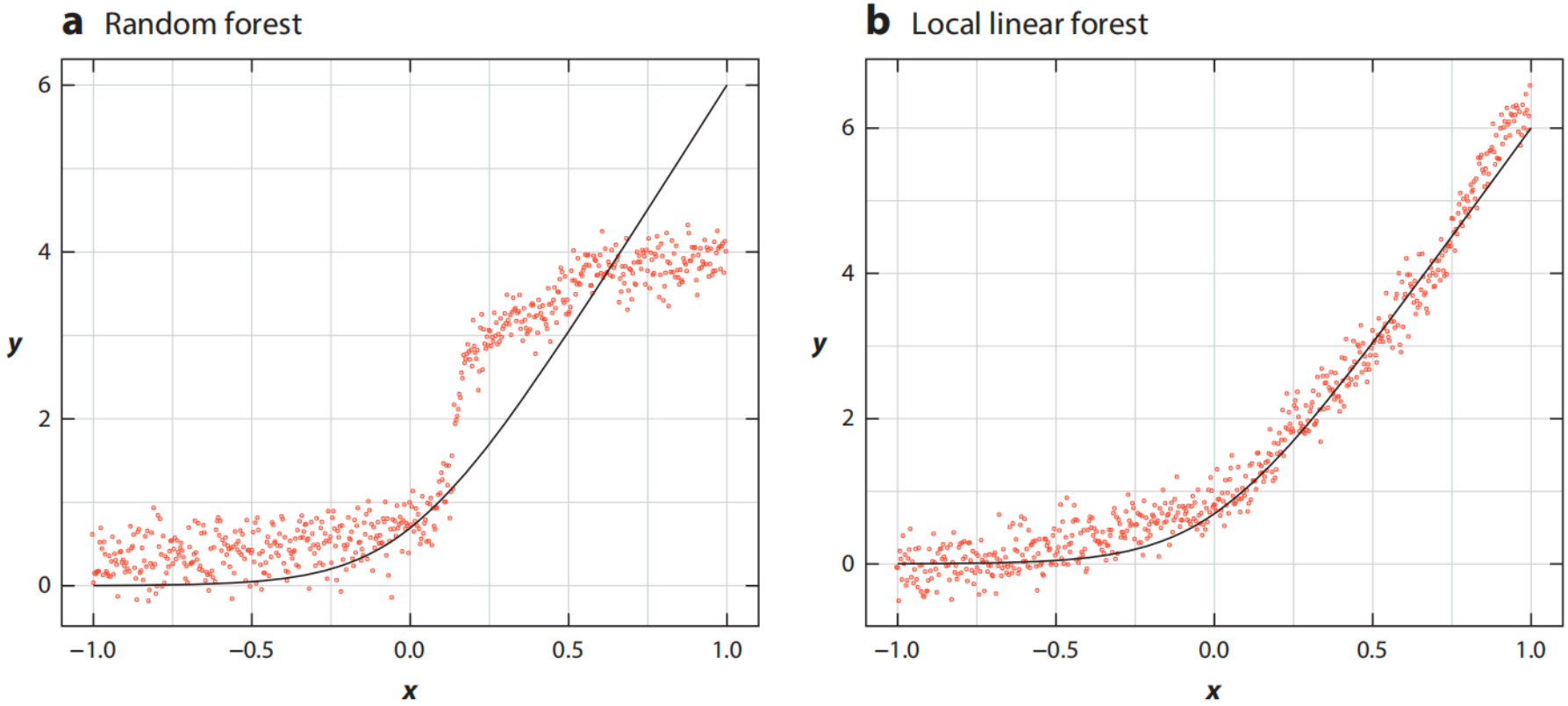

Figure 3

Predictions from random forests and local linear forests on 600 test points. Training and test data were simulated from $Y_i = \log\left(1 + e^{6X_{i1}}\right) + \epsilon_i \sim \mathcal{N}(0,\ 20)$, with X having dimension $d = 20$ (19 covariates are irrelevant) and errors $\epsilon \sim N(0, 20)$. Forests were trained on $n = 600$ training points using the R package GRF and tuned via cross-validation. The true conditional mean signal $\mu(x)$ is shown in black, and predictions are shown in red. Figure adapted with permission from Friedberg et al. (2018).

Introduction to Average Treatment Effects (ATE)

Context and Importance

- Understanding the ATE is crucial for evaluating interventions in econometrics.
- ATE provides insights into the causal impact of a binary treatment on a given outcome.

Definition and Assumptions

- ATE measures the expected difference in outcomes due to treatment: $\tau = E[Yi(1) - Yi(0)]$.
- Requires unconfoundedness assumption for accurate estimation (Rosenbaum & Rubin 1983; Imbens & Rubin 2015).

Mathematical Formalization

$$Wi \perp\!\!\!\perp [Yi(0), Yi(1)]|Xi, \tag{2}$$

- This assumption allows for the identification of ATE when potential outcomes are independent of treatment given covariates.

Estimation Strategies for Average Treatment Effects (ATE)

Characterization of ATE

- ATE can be represented in various ways, including:
  1. Covariate-adjusted difference between treatment groups.
  2. Weighted average of outcomes.
  3. Via influence or efficient score function.

$$\tau = E[\mu(1, Xi) - \mu(0, Xi)] \tag{3}$$

$$= E\left[\frac{YiWi}{e(Xi)} - \frac{Yi(1 - Wi)}{1 - e(Xi)}\right] \tag{4}$$

$$= E\left[\frac{[Yi - \mu(1, Xi)]Wi}{e(Xi)} - \frac{[Yi - \mu(0, Xi)](1 - Wi)}{1 - e(Xi)}\right] + E[\mu(1, Xi) - \mu(0, Xi)], \tag{5}$$

- $\mu(w, x) = E[Yi|Wi = w, Xi = x]$ and $e(x) = E[Wi|Xi = x]$.
- Choose representation based on desired estimation: conditional outcomes, propensity score, or both.

Implications for Econometric Practice

- Model selection is crucial; predictive covariates for treatment and outcome should be included (Belloni et al. 2014).
- Doubly robust methods and covariate balancing approaches offer newer, more stable ATE estimation.

Orthogonalization in ATE Estimation I

The Concept of Orthogonalization:

- Orthogonalization is a process that enhances estimation by making errors in nuisance parameters orthogonal to the estimation errors in the parameter of interest.
- In ATE estimation, this is achieved via the influence function:

$$\psi(y, w, x) = \mu(1, x) - \mu(0, x) + \frac{w}{e(x)}[y - \mu(1, x)] + \frac{1 - w}{1 - e(x)}[y - \mu(0, x)],$$

with $\hat{\tau}_i = \psi(Y_i, W_i, X_i)$.

Orthogonalization in ATE Estimation II

Efficiency of Estimators:

- An estimator based on the influence function is efficient if the nuisance estimators $\hat{\mu}(w, x)$ and $\hat{e}(x)$ are sufficiently accurate, satisfying:

$$\sqrt{E[\hat{\mu}(w, X_i) - \mu(w, X_i)]^2}\sqrt{E[\hat{e}(X_i) - e(X_i)]^2} = o(N^{-1/2}).$$

- This efficiency holds even if each nuisance component converges at a slower rate (e.g., $N^{-1/4}$).

Cross-Fitting in ATE Estimation

Cross-Fitting Technique:

- Cross-fitting enhances performance by estimating nuisance parameters independently of the observation's outcome.
- It employs methods such as sample splitting, out-of-bag prediction, and leave-one-out estimation.

Advantages of Cross-Fitting:

- Reduces overfitting, especially in models with high flexibility and many covariates.
- Enhances robustness in estimation by mitigating the influence of a single observation.
- Allows ML models to better handle many covariates relative to the number of observations.

Applications in Econometrics:

- Cross-fitting and orthogonalization are crucial for estimating nuisance parameters in econometric ML applications.
- They enhance the reliability of estimating heterogeneous effects and models involving unconfoundedness or instrumental variables.

Treatment Effect Heterogeneity

Exploring Heterogeneity:

- Machine learning is adept at uncovering treatment effect heterogeneity with respect to observable covariates.
- Key questions include identifying which individuals benefit most, determining positivity of treatment effects, and analyzing how effects vary with covariates.
- Understanding heterogeneity is crucial for both scientific knowledge and optimal policy assignments (*Athey & Imbens 2017b*).

Conditional ATE (CATE):

- Defined as $\tau(x) = E[\tau_i | X_i = x]$, where $\tau_i = Y_i(1) - Y_i(0)$.
- Identified under the unconfoundedness assumption, addressing the "fundamental problem of causal inference" (*Holland 1986*).

Problems Addressed:

a) Learning low-dimensional representations and hypothesis testing of heterogeneity.

b) Estimating a flexible (nonparametric) $\tau(x)$.

c) Estimating optimal policy for treatment allocation based on covariates x.

Adapting ML for Causal Parameters

Criterion Function Challenges:

- Predictive models use mean squared error (MSE) for model selection.
- Estimating the CATE's MSE is infeasible due to unobservable unit-level causal effects.
- Regularization and estimator comparison are more complex when focusing on structural or causal parameters.

Adapting Cross-Validation:

- Difficulties in cross-validation can be addressed through careful adaptation of regularization methods.
- *Athey & Imbens (2016)* propose criteria for optimizing covariate splits and cross-validation.

Model Selection Insight:

- Comparing models avoids the need to estimate the difficult τ_i^2 term.
- The transformed outcome $Y_i^* = \frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}$ allows $E[Y_i^*|X_i] = E[\tau_i|X_i]$ estimation.
- Estimation of the propensity score $e(X_i)$ introduces dependency on modeling choices.

Introduction and Data Representation

Text Analysis in Machine Learning

- A rich area of study, with a comprehensive review by *Gentzkow et al. (2017)*.
- Focus on a dataset of **N documents** and methods to analyze them.

Data Representation

- Documents represented as rows and words as columns in matrix **C**.
- Information loss concerns lead to richer representations like **bigrams**.

Unsupervised Learning and Topic Modeling

Unsupervised Learning

- Seeks a lower-rank representation of matrix $\mathbf{C}$ to uncover latent document characteristics.
- Includes matrix completion methods to predict elements in a test set.

Topic Modeling: LDA

- Estimates latent topics and their word distributions to understand document composition.

Word Embeddings and Supervised Learning

Word Embedding Methods

- Capture the latent semantic structure of language with vector representations for words.

Supervised Learning

- Targets learning of specific characteristics from text using labeled data.
- Regularization is key due to high dimensionality ($T > N$).

Combining Approaches: Supervised Topic Models

Supervised Topic Models

- Incorporate supervised and unsupervised learning techniques.
- Use observed characteristics to enhance predictions from generative models.
- Useful in labeling documents based on learned topics and characteristics.

Unsupervised Learning I

- Unsupervised learning deals with unlabeled data, focusing on finding patterns or intrinsic structures.
- **Objective**: Partitioning observations into clusters or estimating joint distributions of variables.
- **K-Means Clustering**:
  1. Given observations $\{X_i\}$, partition the feature space into K clusters.
  2. Clusters help in creating new features or organizing sample into types for differential treatments.
  3. No natural benchmark for the "best" solution, making the evaluation subjective.

Unsupervised Learning II

- **Algorithm** (Hartigan & Wong, 1979; Alpaydin, 2009):
  1. Initiate centroids $\{b_1, b_2, ..., b_K\}$.
  2. Assign each point to the nearest centroid, minimizing $||X_i - b_c||^2$.
  3. Update centroids to the mean of assigned points.
  4. Iterate until convergence.
- **Challenges**:
  - Choosing K is non-trivial without a clear cross-validation approach.
  - Often based on substantive reasoning rather than data-driven methods.
- **Unsupervised Learning Landscape**:
  - K-Means is one among many techniques like topic models and unsupervised neural networks.

For Further Reading I

- Susan Athey and Guido W Imbens.
  Machine learning methods that economists should know about.
  Annual Review of Economics, 11:685-725, 2019.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York:
- Alexandre Belloni, Victor Chernozhukov, and Christian Hansen.
  High-dimensional methods and inference on structural and treatment effects.
  Journal of Economic Perspectives, 28(2):29-50, 2014.