AnEconometric History and Perspective to Machine Learning

and

Suggested Next Steps

Gregory M Duncan

University of Washington

Introduction

- Machine Learning is a chair with three legs
  - Statistics
  - Computer Science
  - Domain knowledge(Bio-Science, Econometrics)
  - Each group seems to give at best lip service to the others
    - Biotech has shown the way
    - Social Sciences and Statistics Lag

Introduction

- Computer Science will tell you that it invented "big data" methods
- True to some extent BUT
- Actually, a lot comes from economics

Much of the new stuff is quite old.

Impossible to do in its day

Made Possible by Advances in CS and Statistics

Econometric History of Machine Learning

- Causality and identification
  - Economics 1950
    - Leonid Hurwicz (Economics 2007, Nobel Prize winner)(U Minnesota)
  - Computer Science 1998
    - Judea Pearl (UCLA)

- DAGS and Causality
  - Economics 1921
    - Phillip and Sewell Wright
  - Computer Science 1988
    - Judea Pearl (UCLA)
- Tree methods
  - Economics 1963
    - James N. Morgan (U of Michigan)
  - Statistics 1984
    - Leo Brieman (Berkeley)
- Map Reduce
  - Economics 1980
    - Me (Northwestern)
  - Computer Science 2004
    - Jeffrey Dean and Sanjay Ghemawat (Google)

Econometric History of Machine Learning

- Bagging/Ensemble Methods
  - Economics 1969
    - Bates and Granger (UCSD)
  - ML 1996
    - Brieman (almost of course)
- Principal Components (PCA)
  - Economics 1938
    - Harold Hotelling
  - ML
    - Something they don't take credit for
- Multi-armed Bandits
  - Economics 1974
    - Rothschild
  - ML ~2005
- Neural Nets
  - Economists early contributors to neural net/deep learning literature (late 1980's early 1990's)
    - Hal White, Jeff Wooldridge, Ron Gallant, Max Stinchcombe 1990s
  - Quantile neural nets, learning derivatives, guarantees for convergence, distributions etc.
    - Needed for counterfactuals when out of sample validation make no conceptual sense

Computers made this possible

- Much of machine learning
  - rediscovers old statistical and econometric methods
  - but making them blindingly fast and practicable

This is not damning with faint praise

- Many of the tools used today have been enabled purely by computers
  - Bootstrap (1978)
    - resampling
  - SAS (1966, but really 1976)
    - Huge datasets
  - TSP (1966)
    - UI
      - Albeit command line
  - Principled Model Selection (Lasso, L2-boosting, Double LASSO)

Black Box Models

- Basis function Models
  - Expand non-linear function into linear combination of large number of basis functions
    - Number of basis functions > number of observations
      - Difficult to interpret
- Regression Trees and Variants
  - CART
    - Interpretable
  - Random forests
    - Uninterpretable
- Neural Nets
  - Difficult if not impossible to interpret

- Machine learning allows more complex models than economists are used to
- These models are spectacularly predictive
  - Out of sample
- In part this has led to the

Culture Wars

- Brieman(2000) Statistical Modeling: The Two Cultures
- Shmueli(2010) To Explain or To Predict
  - ML
    - How does the model predict out-of-sample
    - Proofs of properties more-or-less useless
    - Let data reveal structure
    - Doesn't care about guarantees or properties
      - If I beat you out of sample every time
      - Who cares if you are best in class on paper
  - Statistics and Economics
    - What is the underlying probability model
    - How to infer structure
    - How to predict using structure
    - Can we determine causality
    - Though economist and nearly as famous Economist Milton Friedman was clearly amenable to the ML camp
      - I would add Sims as well.

Some ML Contributions

Pointed out Bias Variance Tradeoff

- Bias falls with the complexity of the model
- Variance increases with complexity
  - Constant models are simple
  - Basis function regression models or neural nets complex
- Interpretability decreases with complexity or flexibility
- Yielding the

Interpretability Flexibility Tradeoff

Some ML Contributions

- Value of Out of Sample Prediction
  - Models evaluated on hold out sample
  - When economists miss this we disastrously overfit
    - Excellent in sample predictions R^2 ~.9
    - Corr(Out of Sample Actual, fitted)^2~0
      - Just fit noise
- Cross validation
  - Variant on hold out sample
- Regularization
  - Penalize complex models to reduce variance
  - LASSO, Ridge and Elastic Net are good examples in Linear Regression Framework
    - Work outside linear regression framework

Some Things Economists Can Contribute

- Handling Causality
  - Your perfect AI/ML model predicts 3% revenue decline next month
    - No kudos from management for perfection
  - Just two questions

WHY?

And

WHAT CAN WE DO ABOUT IT?

Causality

- Generally want to identify variables endogenous and exogenous
- Split exogenous into actionable and not
- Economists have contributed lots of work on the identification side starting with Reirosol, Marschak and Hurwicz.
- More recently Chesher, Chernozhukov, Matzkin, Newey, White
  - No idea how to do this at scale
    - Without theory
    - Not enough econometricians in the world to do a day's work in a lifetime.

The Causal Counterfactual Challenge

- These require causal and counterfactual models
- ML models not designed for these
- Problem 1
  - Can’t do out-of-sample cross validation for counterfactuals
- Problem 2
  - Best predictive models uninterpretable

The Causal Counterfactual Challenge

- Height and weight highly correlated
  - Dieting doesn't make you shorter
- City altitude and average temperature highly negatively correlated
  - Putting heaters in all homes
  - Or global warming
  - Won't decrease altitude
- New models required
- Some effort here
  - Athey, Imbens, Pearl, Jantzig, Schölkopf, Wager, White among others
  - Lots of work needed

Causality

Hal White addressed this at the end of his life

- Factor the joint characteristic function of Y and X given Z
  - If factors into CF(Y|Z)CF(X|Z) then Y⫫X given Z.
  - This requires a functional approach
    - Some interesting results on detecting global non-causality
    - No code
    - No implementation
- Research program needs to be finished

Interpretability

- A person doesn't get a loan
- Reasonably asks why
- Reasonably asks what can I do to get approved
  - Thinking credit score, paying quicker, lowering debt
- The answer: "Don't know. The model says you don't get the loan." will not fly.
  - Class action attorneys will have a field day

An Approach to Interpretability

- Black-box predictive models
  - Interpret by old fashioned comparative statics
- Intuition
  - Black box model filters out the error
  - If model is differentiable, take derivatives along coordinate axes
  - If model is not differentiable
    - Smooth model
    - Take derivatives of smoothed model

Framework

- Usual independent and identically distributed framework

$$y = f(x) + \varepsilon$$

- By some method learn

$$f(\cdot)$$

- Prediction denoted by

$$\hat{f}(x)$$

Surrogate Models

- For the prediction matrix X, get the predictions

$$\hat{y}_i = \hat{f}(x_i), x_i \in X$$

- Choose an interpretable model (linear model, decision tree, …).
- Train the interpretable model on the predictions as pseudo-data
- Surrogate represents fitted model.
  - Different from whether the fitted model is representative of reality.
  - Explains model
  - Not necessarily data
- Interpret or visualize surrogate model

Local Surrogate Models

- Choose a reference point or point of interest(POI)
  - For policy often the current state
- Simulate along single coordinate
- Fit a univariate local linear regression or spline
  - Very much like LIME
- Repeat in each coordinate direction
- Interpret coefficients as gradient at POI
  - Numerically
  - Visually

OF COURSE

All this

is old fashioned

first year

micro

COMPARATIVE STATICS

Importance Measures

- Discussed only lest we get misled
  - ML is fixated on the Feature Importance Question
    - How important is a variable relative to others in reducing Loss
    - Cannot answer the comparative statics question
  - Shapley Values for Importance per feature
    - Hot think in ML right now
    - Answers an uninteresting question
      - What is a fair method for determining which variables are the “most” important
    - Cannot answer questions about changes in prediction for changes in the input
      - “If I you to earn $300 more each month would you qualify for a loan.”

Confidence Intervals

- ML relatively uninterested in these
- Confidence intervals available in some cases
  - If predictor is random forest or tree
    - Use Efron-Hastie-Stein infinitesimal jackknife approach
    - Followed by Athey-Wager approximate normality result
  - Likely for boosted regression but haven't shown
    - Zhou and Hooker (2018)
  - Works on neural nets
    - Chen and White (1999) but need extension to deep learned nets

Confidence Intervals

- In the above Spline fit predictions with Generalized Least Squares
  - Covariances of predicted values as Weight Matrix
  - Confidence intervals follow directly
- Gives presumably more efficient derivative estimates
- Gives accurate variances for confidence intervals

Time Series

- Almost all ML have underlying their approaches the independent and identically distributed model
  - Useful in some experimental situations
  - Problematic in most economic and business contexts

Finance

- High frequency financial data
  - Time series based
  - Vast
- A lot of propriety solutions out there
- Known only within the companies
- Known completely only by a few
- Backcasting is a bad way to cross validate

Some Examples: Supply Chain and Inventory Control

- Millions of products and sub products
- Many sell only a few in a year
  - Called the intermittent demand problem
  - Bane of spare parts inventory planning
    - Boeing
    - GM
  - Bane of retail
- Vendor managers must have what's needed on hand
- Vendor managers must not have too much on hand

Supply Chain

- Classic Newsboy Problem
  - Need demand predictive distribution (quantiles)
  - But with only 2 sales in a year what to do?
  - Learn across “similar” products
  - What products are “similar” in this sense?
    - How to “cluster” millions of products into sensible groups
    - In real time
    - As products enter and exit

Clustering: A Lasso Idea

- Regression model with fixed effects

$$y_{it} = x_{it}^T \beta + \alpha_i + u_{it}, t = 1,..,T, i = 1,...,N$$

- Generalized LASSO

$$\min_{\beta,\alpha} \sum_{t=1}^{T} \sum_{i=1}^{n} \left(y_{it} - x_{it}^T \beta + \alpha_i \right)^2 + \lambda \sum_{i=1}^{n} \sum_{j=i}^{n} \left| \alpha_i - \alpha_j \right|$$

- Regularization penalty: sum of absolute differences of all pairs of fixed effects.

Clustering

- Drives the differences of similar fixed effects to zero.
- Natural Clusters
  - Thus if a group of observations have the same fixed effect they are in the same category, segment or cluster.
    - Gertheiss and Tutz (2011)

Issues Economists Need Address

- Vast data available today
  - hundreds of billions of observations and millions of features
  - A 100,000,000,000 x 1,500,000 dimensional tables/matrices common
- Allows analyses unthinkable 15 years ago
- Presents challenges to econometricians, computer scientists and statisticians
  - How to get that matrix into a computer
    - Can you get that matrix into a computer
  - If so how do you interpret 1.5 million independent variables
    - Can you interpret 1.5 million independent variables

Operating At Scale

- Econometricians my generation
  - Programmed their own stuff
    - FORTRAN, C, APL, Assembly, Cobol
      - The weak used BASIC
- This changed somewhere in the mid 1980's
  - Programming became using SAS, TSP, SPSS, GAUSS
- Now it means STATA
  - For a few MatLab and R

This Is Insufficient

- Data won't fit in memory.
  - Sampling fails with low signal to noise ratios
- Standard methods fail with on-line real time data
- New PhD seem not to know the rudiments of CS
  - Github for version control
  - Basic Database management
  - Choosing the right language for the problem
    - The language should be a detail
  - Parallel Processing

Pet Peeve

- We typically spend a year or more teaching new PhD Economists rudimentary CS
- Their training usually allows them to grasp the statistical part very well.
- Except many do not know how to handle truly observational data
- In business unacceptable to answer:
  - Nature did the wrong experiment. Too bad. So sad.
- Their response
  - either you can help us or not.
- CS, Finance and Marketing filled gap
  - And economics got left behind

Need for CS and Stat

- Need basic course in CS
  - To say otherwise like saying theorists don't need calculus
- Need serious training in Stat and Probability
  - Unlike parametric models where one can specify optimal or best models
  - No Free Lunch Theorem
    - There is no universal best learner
    - Different datasets require different approaches

Some Newer Stuff: Double LASSO

- To avoid data mining problems
  - Overfitting with very high numbers of uninteresting variables
  - Incorrect inference
- Divide independent variables into two parts
  - Focal Variables
  - Uninteresting variables
    - Like to omit the uninteresting variables
    - But omitted variables problem if important and correlated with the focal variables
  - The lasso finds the features with non-zero coefficients
- LASSO dependent and each focal variable on all the uninteresting variables
  - Keep ones that are kept in any of LASSOs
  - Using these and the focal features in subsequent OLS of the dependent variable requires no adjustment for the lasso
    - All tests are correct
    - In large enough samples

Even Newer: SAFER LASSO

- In double LASSO
  - You don't need to do any LASSOs!
- Work with the dual of the LASSO
- Then turns out can work with the correlations of
  - y with nuisance
  - Focal with nuisance
  - Keep any variable where the correlation is too Large!
    - Ghaoui, Viallon, and Rabbani 2010
    - Xiang, James and Ramadge (2012)
    - Fercoq, Gramfort, and Salmon 2015
- Final regression of y on focal and kept nuisance variables
- Usual asymptotics hold!

The End

Appendix

- The Lasso

$$\min_{\alpha} \sum_{i=1}^{n} \left(y_i - x_i^T \alpha \right)^2 + \lambda \left\| \alpha \right\|_1$$

  - Drives coefficients that should be zero to zero with high probability

Example of Black Box Models: Trees and Forests

- Partition the feature space into a set of mutually exclusive regions, R_m
- Fit a simple model in each
  - Average for regressions
  - 0/1 prediction for classification
- In essence, a piecewise constant approximation

$$E(y|x) \approx \sum_{m=1}^{M} c_m I(x \in R_m)$$

  - $I(x \in R_m)$ indicator function for set R_m

- Choose M, c_m and R_m to minimize

$$\sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} c_m I(x_n \in R_m) \right)^2$$

- If we knew the regions R_m, this would be a straightforward dummy variable regression

Trees

- But we don't know the regions.
  - So find them by searching
- As stated, this problem is too hard to solve
  - Typically need at least 5 observations in a region
  - How many possible 5 observation regions are there with N observations with p (say 1000) independent variables?
    - I would not know how to even approach the counting problem
- SO we solve a simpler one using simple regions defined recursively
- Brings us to the second ML method we will discuss

CART

- For each independent variable
  - split range into two regions,
- Calculate mean and sum of squares of Y in each region.
- Split point minimizes the SSR of Y in the two regions
  - (best fit)
- Choose the variable to split on as the one with the lowest SSR
- Both of these regions are split into two more regions,
- this process is continued, until some stopping rule is applied.
  - The split points are called nodes
  - The end points are called leaves
- Stop splitting a leaf if
  - all the values in the leaf are the same
  - the number of branches exceeds some tuning parameter
  - the tree gets too complex by some criteria
- Quit when no leaves can be split.

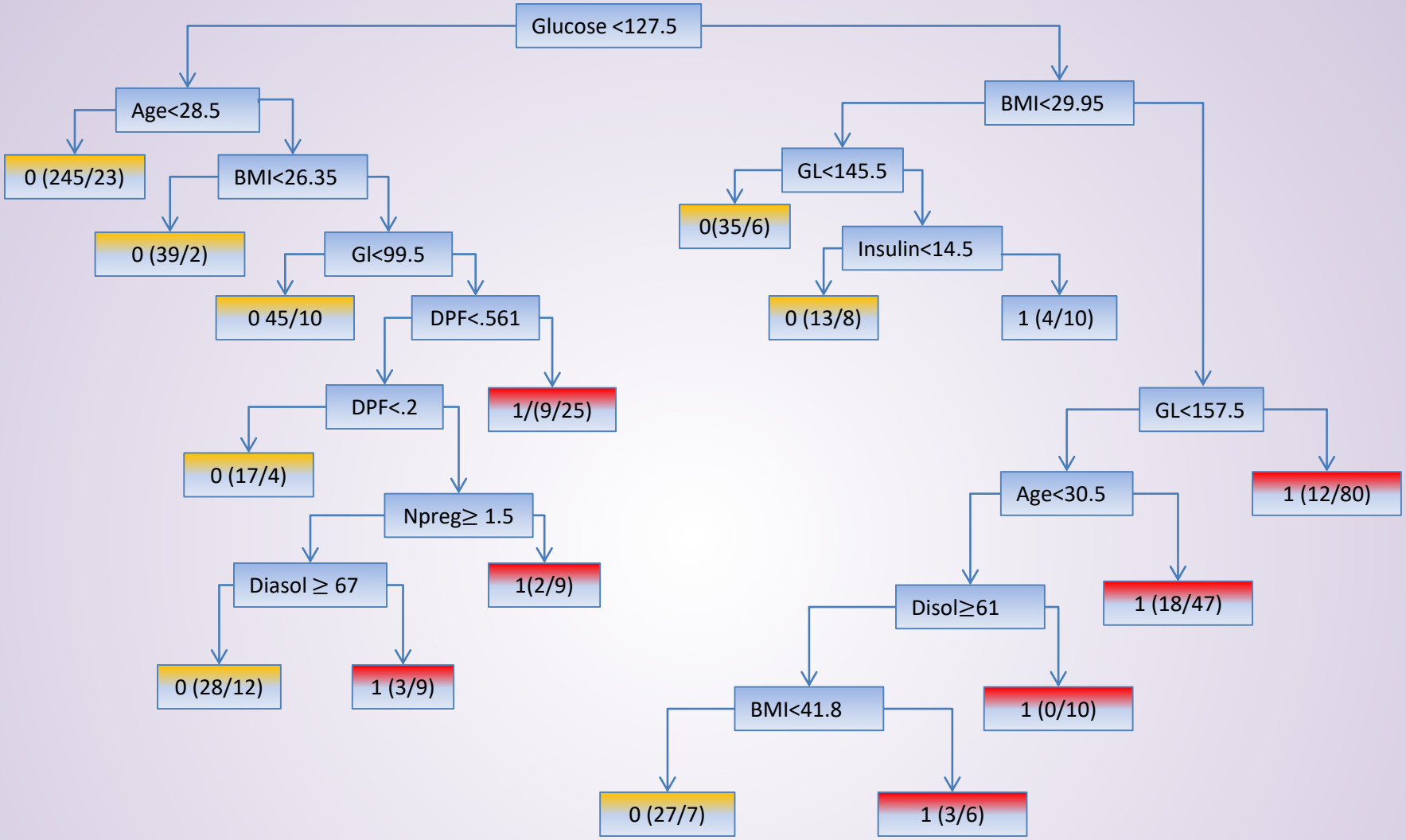
Example: Pima Indian Diabetes Study

- The binary-valued variable tested positive for diabetes
- The population lives near Phoenix, Arizona, USA.
- Number of Instances: 768
- Number of Attributes: 8 plus class

1. Npreg- Number of times pregnant
2. Glucose- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastol- Diastolic blood pressure (mm Hg)
4. TSF- Triceps skin fold thickness (mm)
5. Insulin- 2-Hour serum insulin (mu U/ml)
6. BMI- Body mass index (weight in kg/(height in m)^2)
7. DPF- Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

(class value 1 is interpreted as "tested positive for diabetes")

Class Value	Number of instances
0	500
1	268



PIMA Indian Diabetes Study

Problems

- Overfitting
  - Horrid out of sample accuracy
- What tuning parameters to use
  - How complex a tree, how deep
- Huge literature on what to do
  - Not very persuasive
- Legitimate researchers of good will using acceptable criteria would come up with VERY different answers.

Rescue 1:Boosting

- Train a tree $y = h_o(x) + r_0$
- Train a second tree $r_0 = h_1(x) + r_1$
  - Probably, you'll see error decrease.
- Continue m=2,..,M

$$r_{m-1} = h_m(x) + r_m$$

$$\hat{y}_m = \sum_{i=1}^{m} \eta_i h_i(x)$$

- Stop when sufficient accuracy is achieved.
- Questions: what to use for a loss function, what to use for the *h*, what to use for the η, stopping.

Boosting Rarely Overfits

Repeat!!

Boosting Rarely Overfits

Rescue 2: Random Forests

- Brieman (2001)
- Simple trees:
  - if independent and unbiased
  - average would be unbiased and have a small variance.
- Called ensemble learning
  - averaging over many small models tends to give better out-of-sample prediction than choosing a single complicated model.
- New insight for ML/Statistics
- Economists have done this for years
  - We call it model averaging
    - Primarily in macro modeling
      - Bates and Granger (1969)
      - Granger and Ramanathan (1984)

Random Forest Details

- For *b* = 1 to *B*:
  - Draw *N* bootstrap observations from the training data.
  - Grow a single tree T(b) from the bootstrapped data
    - repeat the following steps for each node, until minimum node size n_{min}
      - Select m variables at random from the p variables.
      - Pick the best variable/split-point among the m.
      - Split the node into two daughter nodes.
    - The output of tree is average of the y in the regression case or a 0/1 for classification.

Random Forest Details

- To predict the outcome of a new x
  - Run x through each tree to get the prediction for that tree.
  - Average the predictions

TA!DA!

You are DONE

References of Particular Interest

Breiman, L.,(2001), Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author), *Statistical Science*, Volume 16, Issue 3 (2001), 199-231. https://projecteuclid.org/euclid.ss/1009213726

Efron, B., and Trevor Hastie (2016*). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* (Institute of Mathematical Statistics Monographs). Cambridge: Cambridge University Press. https://web.stanford.edu/~hastie/CASI/

Donoho, David (2015) 50 years of Data Science. *Tukey Centennial Workshop*, Princeton NJ Sept 18 2015.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) *An Introduction to Statistical Learning, with Applications in R,* Springer.

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf

Gertheiss, Jan; Tutz, Gerhard. Sparse modeling of categorial explanatory variables. Ann. Appl. Stat. 4 (2010), no. 4, 2150--2180. doi:10.1214/10-AOAS355. https://projecteuclid.org/euclid.aoas/1294167814

http://pages.cs.wisc.edu/~anhai/courses/784-fall15/50YearsDataScience.pdf

El Ghaoui, L., Viallon, V., and Rabbani, T. Safe feature elimination in sparse supervised learning. Pacific Journal of Optimization, 8(4):667–698, 2012.

Harry J. Paarsch and Konstantin Golyaev (2016) *A Gentle Introduction to Effective Computing in Quantitative Research: What Every Research Assistant Should Know*, MIT Press.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf(2017) *Elements of Causal Inference:Foundations and Learning Algorithms*, MIT Press

John W Tukey (1962) The future of data analysis. *The Annals of Mathematical Statistics*, 1–67. https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711

References

Angrist ,J. and Alan B. Krueger, (2001) Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments, *Journal of Economic Perspectives* –Vol. 15, 69-85

Angrist, J., Imbens, G., and Rubin, D.,(1996) Identification of Causal effects Using Instrumental Variables, *Journal of the American Statistical Association*.

Angrist, Joshua D. and Jörn-Steffen Pischke, (2008) *Mostly Harmless Econometrics*, Princeton University Press

Bates, J., and C. Granger (1969), The Combination of Forecasts, *Operations Research Quarterly*, 20, 451-468.

Berk, Richard A., (2009) *Statistical Learning from a Regression Perspective*, Springer

Belloni, Alexandre and Victor Chernozhukov (2013) Least squares after model selection in high-dimensional sparse models, *Bernoulli* 19(2), 521–547 DOI: 10.3150/11-BEJ410

Belloni, A., Chernozhukov, V., & Hansen, C. (2014) Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81(2), 608-650.

Breiman, L., (1996) Bagging Predictors*, Machine Learning*, 24(2), pp.123-140.

Breiman, L.,(2001), Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author), *Statistical Science*, Volume 16, Issue 3 (2001), 199-231.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984) *Classication and Regression Trees*. Wadsworth and Brooks/Cole, Monterey.

Blundell, Richard; Matzkin, Rosa L. (2013) Conditions for the existence of control functions in nonseparable simultaneous equations models, http://www.ucl.ac.uk/~uctp39a/Blundell_Matzkin_June_23_2013.pdf

Blundell R.W. & J.L. Powell (2003) Endogeneity in Nonparametric and Semiparametric Regression Models. In Dewatripont, M., L.P. Hansen, and S.J. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications*, Eighth World Congress, Vol. II. Cambridge: Cambridge University Press.

References

Blundell, R.W. & J.L. Powell (2004) Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies* 71, 655-679.

Chesher, A.D. (2005) Nonparametric Identification under Discrete Variation *Econometrica* 73, 1525-1550.

Chesher, A.D. (2007) Identification of Nonadditive Structural Functions. In R. Blundell, T. Persson and W. Newey, eds., *Advances in Economics and Econometrics, Theory and Applications*, 9th World Congress, Vol III. Cambridge: Cambridge University Press.

Chesher, A.D. (2010) Instrumental Variable Models for Discrete Outcomes. *Econometrica* 78, 575-601.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, and C. Hansen (2016): "Double Machine Learning for Treatment and Causal Parameters". Preprint, arXiv:1608.00060. [237,258]

Donoho, David (2015) 50 years of Data Science. *Tukey Centennial Workshop*, Princeton NJ Sept 18 2015, http://pages.cs.wisc.edu/~anhai/courses/784-fall15/50YearsDataScience.pdf

Duncan, Gregory M. (1980) Approximate Maximum Likelihood with Datasets That Exceed Computer Limits, *Journal of Econometrics* 14 257-264.

Einav, Liran and Jonathan Levin. The data revolution and economic analysis (2013) NBER Innovation Policy and the Economy Conference,2013.

O. Fercoq, A. Gramfort, and J. Salmon(2015) Mind the duality gap: safer rules for the lasso. *Proceedings of the 32 nd International Conference on Machine Learning,* Lille, France, 2015. *JMLR:* W&CP volume 37.

Friedman, Jerome and Bogdan E. Popescu (2005) Predictive learning via rule ensembles. Technical report, Stanford University http://www-stat.stanford.edu/~jhf/R-RuleFit.html

Friedman, Jerome and Peter Hall (2005) On bagging and nonlinear estimation. Technical report, Stanford University, http://www-stat.stanford.edu/~jhf/ftp/bag.pdf

References

Friedman, Jerome (1999) Stochastic gradient boosting. Technical report, Stanford University. http://www-stat.stanford.edu/~jhf/ftp/stobst.pdf

El Ghaoui, L., Viallon, V., and Rabbani, T.(2012) Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8(4):667–698.

Haavelmo, T. (1943). "The Statistical Implications of a System of Simultaneous Equations". *Econometrica*, Vol. 11, 1–12.

Haavelmo, T. (1944). "The Probability Approach in Econometrics" *Econometrica*, Vol. 12, Supplement, iii-115

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, 2 ed http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html

Heckman James J. (2010) Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy, *Journal of Economic Literature*, Vol. 48, No. 2

Heckman, J.J. (2008) Econometric Causality. *International Statistical Review*, Vol. 76, 1--27.

Heckman, J.J., (2005) The scientific model of causality, *Sociological Methodology*, Vol. 35, 1-97.

Heckman, James J. and Rodrigo Pinto (2012) Causal Analysis After Haavelmo: Definitions and a Unified Analysis of Identification of Recursive Causal Models, *Causal Inference in the Social Sciences*, University of Michigan

References

Hendry, David F. and Hans-Martin Krolzig(2004) We ran one regression. *Oxford Bulletin of Economics and Statistics*, 66(5):799-810

Holland, Paul W., (1986) Statistics and Causal Inference, *Journal of the American Statistical Association*, Vol. 81, No.396., pp. 945–960

Huang, Meng & Sun, Yixiao & White, Halbert, 2016. "A Flexible Nonparametric Test For Conditional Independence," Econometric Theory, Cambridge University Press, vol. 32(06), pages 1434-1482, December.

Hurwicz, L. (1950) Generalization of the concept of identification. *In Statistical Inference in Dynamic Economic Models* (T. Koopmans, ed.). Cowles Commission, Monograph 10, Wiley, New York, 245–257.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013) *An Introduction to Statistical Learning with Applications in R*. Springer, NewYork.

Xun Lu & Liangjun Su & Halbert White, 2016. "Granger Causality and Structural Causality in Cross-Section and Panel Data," Working Papers 04-2016, Singapore Management University, School of Economics.

Mallows, C. L. (1964). Choosing Variables In A Linear Regression: A Graphical Aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7-9.

References

Marschak, J. (1950) Statistical inference in economics: An introduction. In T. C. Koopmans, editor, *Statistical Inference in Dynamic Economic Models,* pages 1–50. Cowles Commission for Research in Economics, Monograph No. 10.

Morgan, James N. and John A. Sonquist (1963) Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415-434. URL http://www.jstor.org/stable/2283276.

Morgan, Stephen L. and Christopher Winship, (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Cambridge University Press

Neyman, Jerzy. (1923) Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. Master's Thesis (1923). Excerpts reprinted in English, *Statistical Science*, Vol. 5, pp. 463-472.

Pearl, J., (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press .

Pearl, J., (2009) Causal inference in statistics: An overview, *Statistics Surveys*, Vol. 3, 96-146.

Reiersol, Olav. (1941) Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis, *Econometrica* , Vol. 9, 1-24.

Rubin, Donald (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology*, 66 (5), pp. 688–701.

References

Rubin, Donald (1978) Bayesian Inference for Causal Effects: The Role of Randomization, *The Annals of Statistics*, 6, pp. 34–58.

Rubin, Donald (1977) Assignment to Treatment Group on the Basis of a Covariate, *Journal of Educational Statistics*, 2, pp. 1–26.

Shpitser, I. and Pearl, J.,(2006) Identification of Conditional Interventional Distributions. In R. Dechter and T.S. Richardson (Eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, Corvallis, OR: AUAI Press, 437-444.

Wright, P. G. (1928). The tariff on animal and vegetable oils. The Macmillan Company

Wright, S. (1921). Correlation and causation. J. Agric. Res. 20, 557-85.

Wu, Xindong and Vipin Kumar, editors. The Top Ten Algorithms in Data Mining. CRC Press, 2009. URL http://www.cs.uvm.edu/~icdm/algorithms/index.shtml

Xiang, Zhen James and Peter J. Ramadge (2012) Fast Lasso Screening Tests Based On Correlations. *IEEE International Conference on Acoustics, Speech, and Signal Processing*

Shpitser, I. and Pearl, J.,(2006) Identification of Conditional Interventional Distributions. In R. Dechter and T.S. Richardson (Eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, Corvallis, OR: AUAI Press, 437-444.

References

Wright, P. G. (1928). The tariff on animal and vegetable oils. The Macmillan Company

Wright, S. (1921). Correlation and causation. J. Agric. Res. 20, 557-85.

Wu, Xindong and Vipin Kumar, editors. The Top Ten Algorithms in Data Mining. CRC Press, 2009. URL http://www.cs.uvm.edu/~icdm/algorithms/index.shtml

Xiang, Zhen James and Peter J. Ramadge (2012) Fast Lasso Screening Tests Based On Correlations. *IEEE International Conference on Acoustics, Speech, and Signal Processing*