Rise of the Machines
Statistical machine learning for observational studies: confounding adjustment and subgroup identification

Armand Chouzy, ETH (summer intern)

Jason Wang, Celgene

PSI conference 2018

Rise of the Machines

Larry Wasserman

ELSEVIER

Statistics & Probability Letters

Volume 136, May 2018, Pages 4-9

Statistics in the big data era: Failures of the machine

David B. Dunson ✉

⊞ Show more

https://doi.org/10.1016/j.spl.2018.02.028

Abstract

There is vast interest in automated methods for complex data analysis. However, there is a lack of consideration of (1) interpretability, (2) uncertainty quantification, (3) applications with limited training data, and (4) selection bias. Statistical methods can achieve (1)-(4) with a change in focus.

https://normaldeviate.wordpress.com/2013/02/16/rise-of-the-machines/

Statistical machine learning (SML) for observational studies

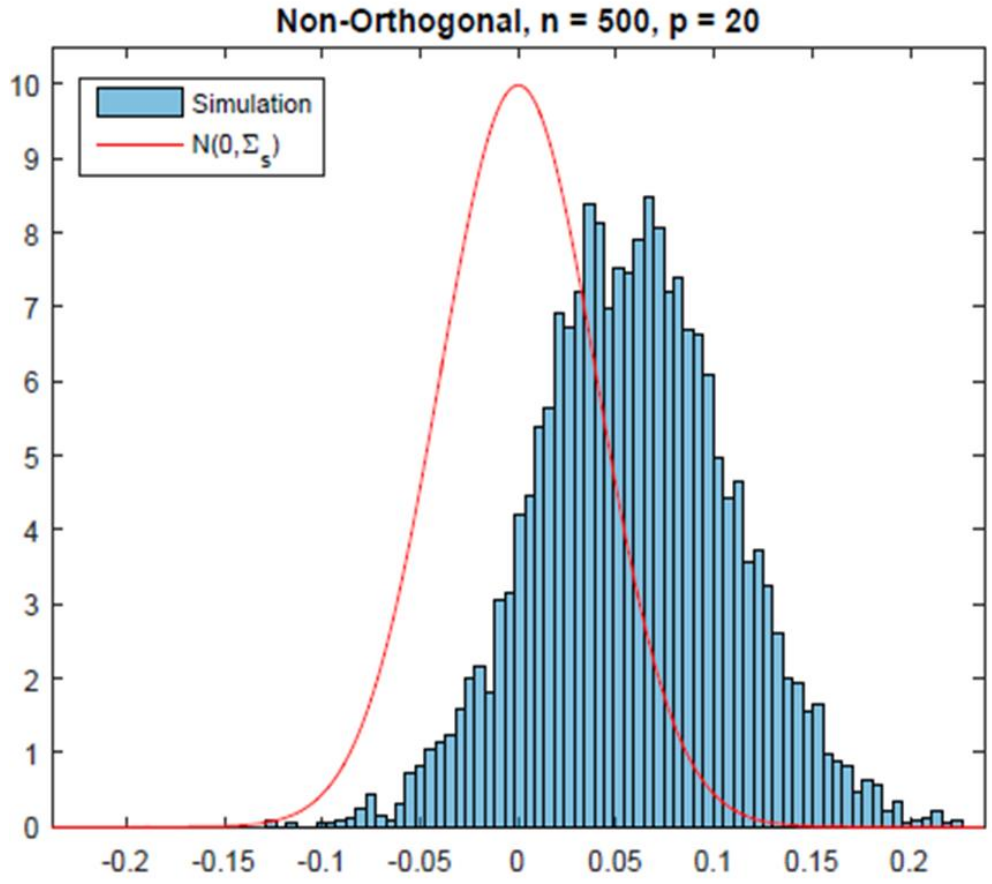
- A conceptual model for outcome
  - **Outcome = Treatment effects + prognostic effects + Error**
- Both effects may contain multiple covariates, some of them may also affect treatment allocation (confounders):
  - **Propensity score = Prob(Treated): a function of covariates**
- SML helps to find covariates in the models, their relationships with treatment allocation as well as with outcomes (prognostic score).
- Learning both relationships is often very beneficial.
- SML approaches are mostly developed for prediction and classification, rather than adjustment and estimation.
- But existing SML approaches can be easily adapted for these purposes.

Treatment effect estimation in observational studies

- When confounders exist, simple estimators for treatment effects, e.g., group mean differences, are not valid.
- To estimate them correctly, we can use 1 of the 4 approaches
  1. Pre-specified direct adjustment with confounders in the outcome model.
  2. Matching and stratification on confounders.
  3. Inverse probability weighting: weights individuals with the inverse probability of getting his treatment (propensity score, PS).
  4. Covariate balancing: weight individuals for covariate balancing between treatment groups.
- With many covariates (even only a few are in the model), none of the approaches performs well.
- In this case, SML can be applied to the above approaches.

Why learning only one relationship may not be sufficient?

- One example shows unsatisfactory performance of direct adjustment with a simple ML application.
- Simulated data with n=500, p=20
- Random forest learns prognostic score for direct adjustment.
- There is an obvious bias in the estimator.



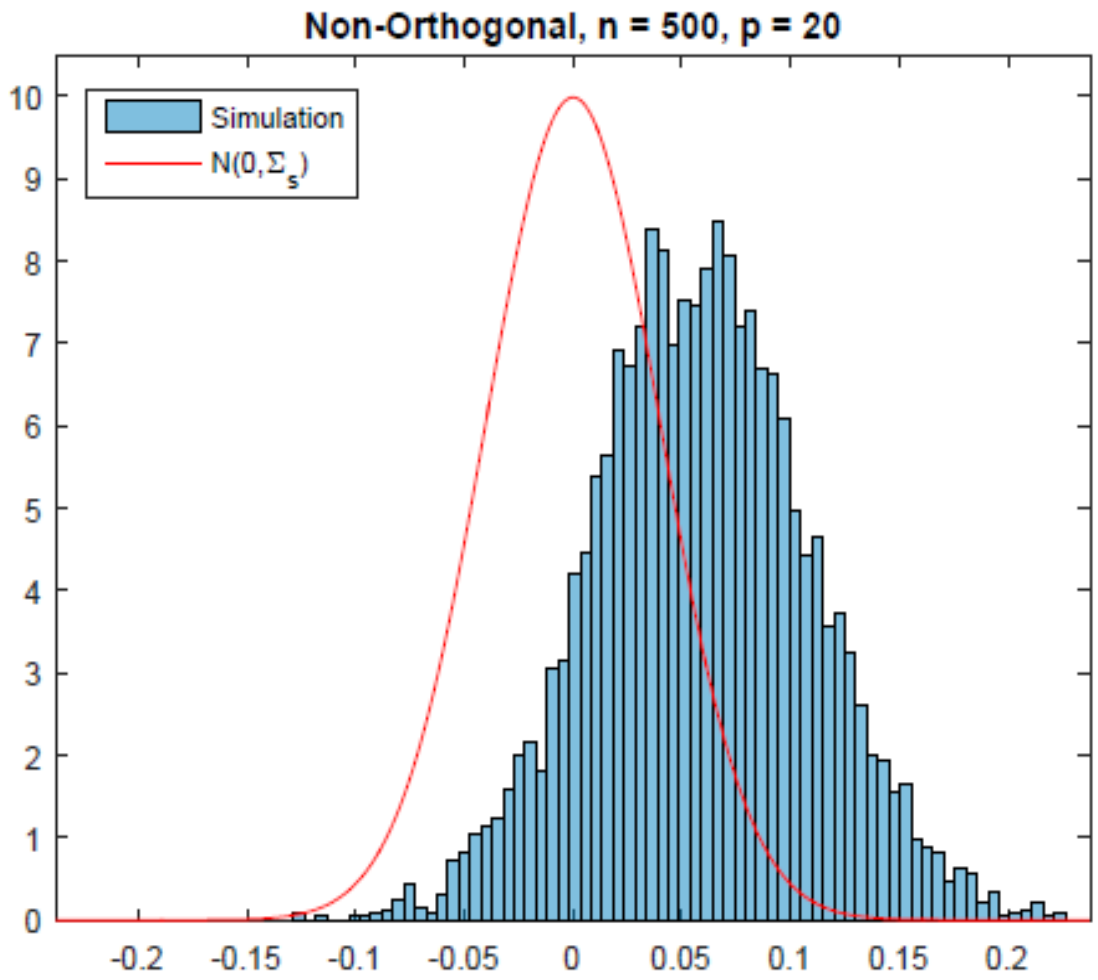
Source: Chernozhukov et al, 2018

Double machine learning (DML) (Chernozhukov et al, 2018)

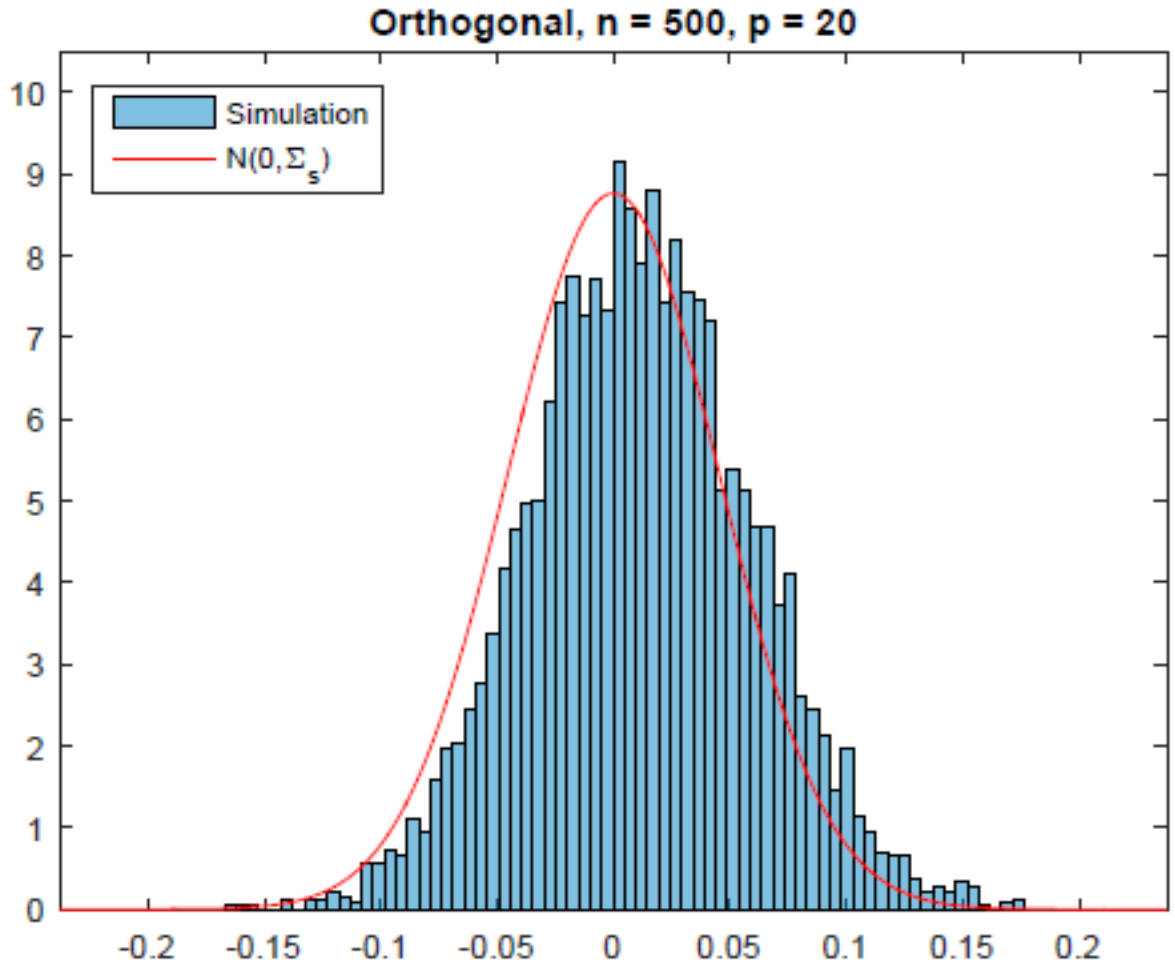
1. Machine-learn prognostic scores (covariate effects on outcome among controls).
2. Machine-learn propensity scores (covariate effects on treatment allocation).
3. Regress the residuals of 1. on the residuals of 2. to estimated treatment effects.
4. The error of the estimator
   $\propto$(error in prognostic score) x (error in propensity score)

- This property is often referred as double robustness.

Random forest learns prognostic score and adjust



Random forest learns prog and propensity scores. Regress residuals on residuals



Source: Chernozhukov et al, 2018

Sample splitting (Rinaldo et al, 2018) and bagging

- The (naïve) DML approach use the same dataset to learn prognostic and propensity scores.
- The following sample splitting steps (1. & 2. in previous algorithm) can improve the performance of DML.

1. Split data into two sets: D1 and D2.
2. Machine-learn prognostic scores with D1.
3. Machine-learn propensity scores with D2.
4. Regress the residuals of 1. on the residuals of 2. to estimated treatment effects based on D1.

- To further gain efficiency, we repeat steps 1-3 K times by resampling D1 and D2, then average the K treatment effect estimates (bagging)

Bagging sample splitting

Randomly split data K times and repeat the analysis K times. Take the mean of K estimates as the estimator.

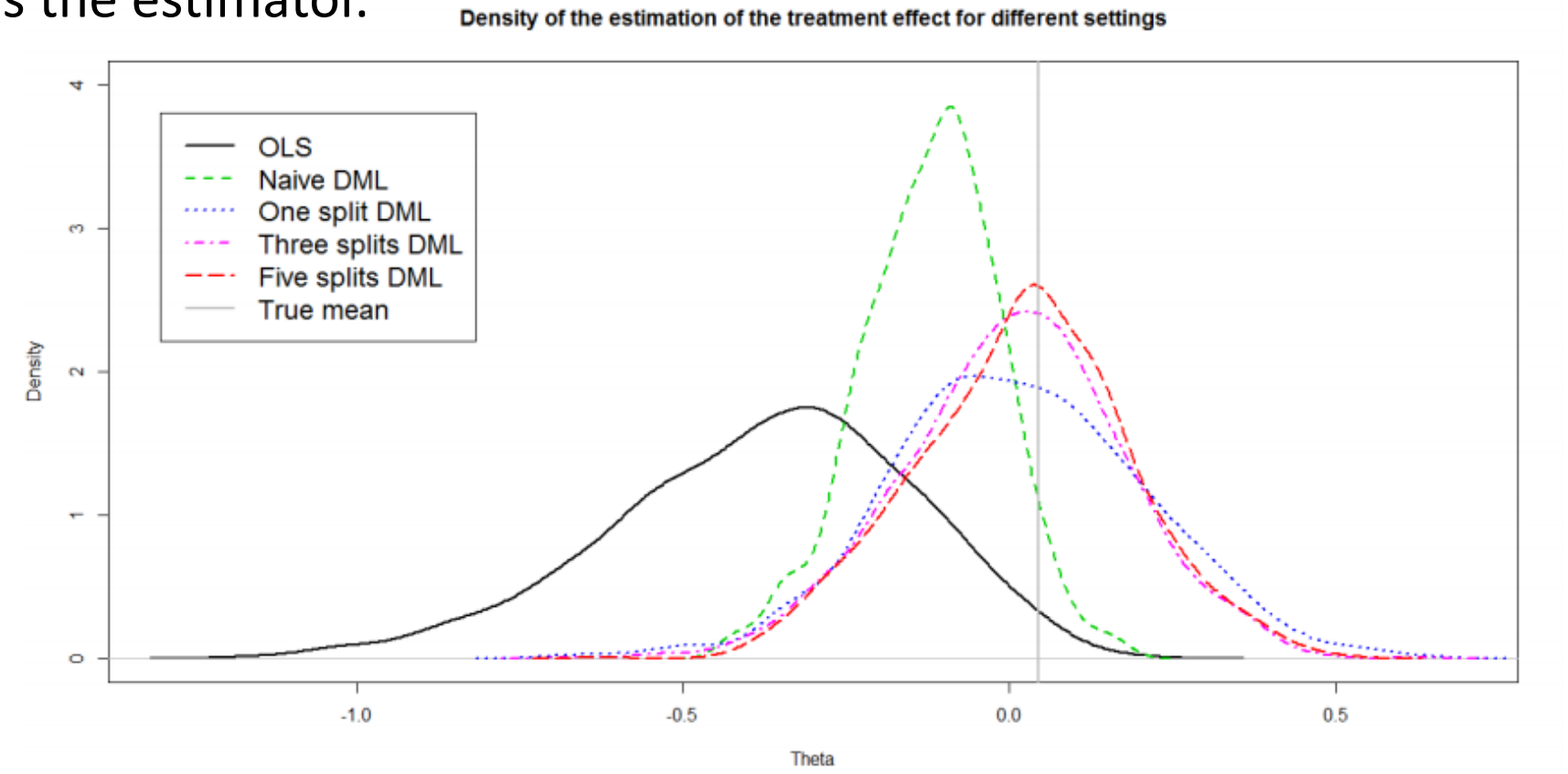
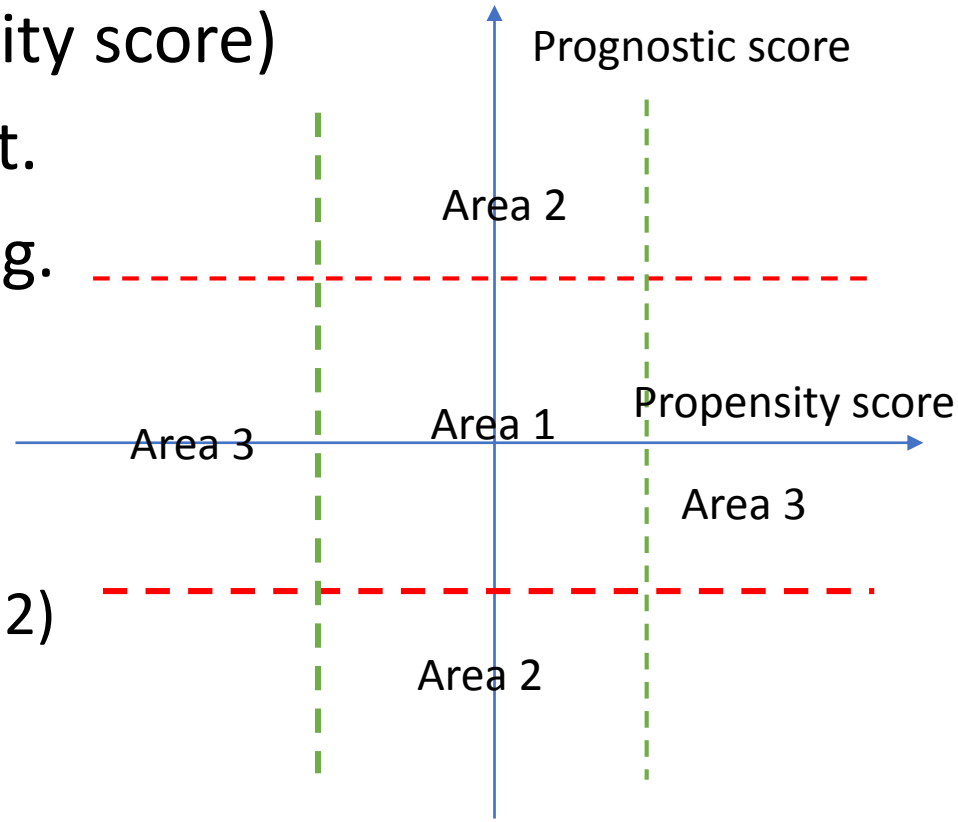


Figure 11: Density plot for the estimated treatment effect of ordinary least square, naive DML and 1,3 and 5- sample splitting DML, $n = 300$ and $p = 50$

Doubly robust matching

- Machine learn both propensity and prognostic score models.
- Then match based on both propensity and prognostic scores.
- The error of the estimator
  $\propto$ (error of prognostic score) x (error of propensity score)
- Hence it is correct if one of the models is correct.
- But there are more unmatched than PS matching.
- Alternatives: sequential Swiss cross matching
  - First match in Area 1 (for double robustness)
  - For those can't be matched, match in Area 2 (or 3)
  - For those still can't be matched, match in Area 3 (or 2)
- Sample splitting for prognostic scores.



Heterogeneity and estimands

- Treatment heterogeneity is more important in observational studies than in randomized trials.
  - Recall the conceptual model:
  - **Outcome = Treatment effects + prognostic effects + Error**
- Which effect to estimate?
  - Average treatment effects: effects in the whole population.
  - Average effects on the treated.
  - Average effects in a subgroup (e.g., is the treatment cost-effective in a special population?)
  - Individual treatment effects (e.g., that of a 60 years old male)

Data driven subgroup identification and honest estimation

- Pre-specified subgroup analyses may be difficult to perform if there are many potential predictive factors and confounders.
- Is data driven subgroup identification useful?
  - Are treatment effect estimates biased (not only the confounding bias)?
- Can we eliminate the bias due to data driving?
  - Sample splitting: training data, cross-validation data, estimation data.
- Closely related:
  - Optimal individual treatment assignment.

Machine learning for subgroup identification

- Outcome = Treatment effects + prognostic effects + Error
- Directly learn treatment-cov interactions with the outcome model is difficult.
- More effective approaches:
  - Predict outcomes with and without the treatment for each subject, then apply an ML approach on the differences (virtual twins, Foster et al, 2011).
  - Outcome transformation: Use $Y_i(2xT_i-1)$ as the differences (Y_i: outcome, T_i=0, 1: treatment indicator).
  - Convert to a classification problem, e.g., by weighting classification error (taking the other treatment than as classified) by the outcome.
  - Many others.
- When treatments are not randomized:
  - Use inverse probability weight together with an approach above.
  - Match by PS, take differences within pairs, then apply an ML approach on them.
  - Doubly robust approaches (e.g., doubly robust matching) .

For illustration: right heart catheterization

- Apply ML approaches to a large study of 5735 patients (Connors *et al.,* 1996), a part of SUPPORT study, 2184 (38%) had right heart catheterization (RHC).
- One main outcome is 30 days survival.
- Unadjusted, those who had RHC had 7.4% (SE=1.3) higher mortality
  - Indication bias? Can we adjust?
  - Subgroups with clear benefit from RHC?
- Adjusted for 65 covariates after univariate screening (drop those with standardized diff <0.1), the mortality of RHC patients is still 5.5% (SE=1.3) higher.
- ML approaches?

Machine learning adjustment for RHC data

ML methods	Risk difference (SE)
Double selection (Belloni et al, 2014): Select covariates for prognostic score and PS models with LASSO, taken all (common) selected covariates and adjust for them in the outcome model.	5.7% (1.3%) (all) 3.7% (1.2%) (common)
Double machine learning (Chernozhukov et al, 2018)	5.7% (1.3%)
Doubly robust matching (Leacy et al, 2014; Antonelli et al, 2018)	4.8% (1.5%)
Covariate balancing propensity score (CBPS, Imai et al, 2013) after LASSO selection	6.0% (1.3%)
Residual balancing (Athey et al, 2017): balance remaining covariate difference after direct adjustment for selected covariates	5.8% (1.4%)
Entropy balancing (Hainmueller, 2011): maximize the weight entropy while balancing covariates between treatment groups (which is also doubly robust (Zhao, 2018)).	5.3% (1.2%)

Machine learning for subgroups in RHC data

- Average treatment effects in the whole population (ATE) or among the treated (ATT)?
  - Doubly robust matching: ATE= 4.8% (1.5%); ATT=6.7% (1.8%).
  - Covariate balancing propensity score: ATE= 6.0% (1.3%); ATT=6.7%(1.3%).
- Subgroup identification:
  - Virtual twins with random forest for prediction: No subgroup identified.
  - Outcome transformation with IPW: No subgroup identified.
  - Doubly robust matching: No subgroup identified.

A randomized trial (n=~2000) showed no difference between RHC and no RHC

ORIGINAL ARTICLE

A Randomized, Controlled Trial of the Use of Pulmonary-Artery Catheters in High-Risk Surgical Patients

James Dean Sandham, M.D., Russell Douglas Hull, M.B., B.S., Rollin Frederick Brant, Ph.D., Linda Knox, R.N., Graham Frederick Pineo, M.D., Christopher J. Doig, M.D., Denny P. Laporta, M.D., Sidney Viner, M.D., Louise Passerini, M.D., Hugh Devitt, M.D., Ann Kirby, M.D., and Michael Jacka, M.D. for the Canadian Critical Care Clinical Trials Group *

January 2, 2003
N Engl J Med 2003; 348:5-14

Summary and discussion

- Statistical machine learning (SML) is a powerful tool for analysis of observational data.
- SML can improve classical confounding adjustment approaches for large and complex studies.
- SML helps subgroup identification, effect estimation, and optimal individual treatment rules.
- Many innovative approaches can be implemented by combining off-the-shelf software/algorithms.
- Want to know more about recent development on machine learning?

http://www.jmlr.org/

Reference

- Belloni, A., V. Chernozhukov, and C. Hansen. “Inference on treatment effects after selection amongst high-dimensional controls.” https://arxiv.org/abs/1201.0224
- Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 76 243–263.
- Leacy, Finbarr P, & Stuart, Elizabeth A. 2014. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in medicine,* 33(20), 3488–3508.
- Foster JC, Taylor JMC, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med.* 2011;30:2867-2880.
- Hainmueller, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20, 25–46
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.
- Susan Athey, Guido W. Imbens, Stefan Wager. Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions. 2017.
- Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei. Bootstrapping and Sample Splitting For High-Dimensional, Assumption-Free Inference. 2018.
- Connors AF et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA. 1996;276:889–897

Backup slides

Residual balancing (Athey et al, 2017)

1. Machine-learn prognostic score from untreated.
2. Predict covariate effects among treated to estimate untreated outcome in treated population, then estimate treatment effect.
3. Machine-learn balancing (matching) covariates between treated and untreated population.
4. Adjusting estimated treatment effect by balanced residual from step 1.
5. The error of this estimator
   =(error in covariate balancing) x (error of para. estimates in step 1.)
6. The performance is close to the oracle (as if we knew the prognostic factors).