Lecture 23:
Machine Learning and Causal Inference

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

April 26, 2018

Machine Learning vs. Traditional Statistics

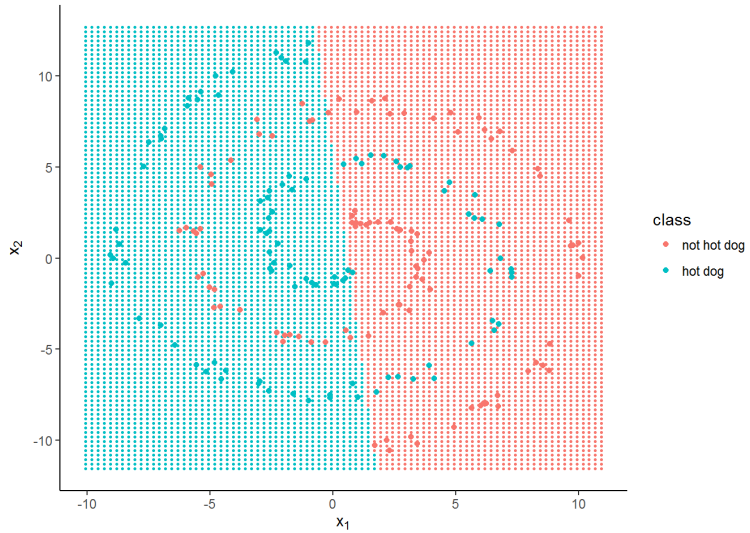
A priori specification vs. "letting the data tell us" about

- specification of conditioning sets (which Xs to include and how to do so),
- sources of effect heterogeneity, or
- causal structure.

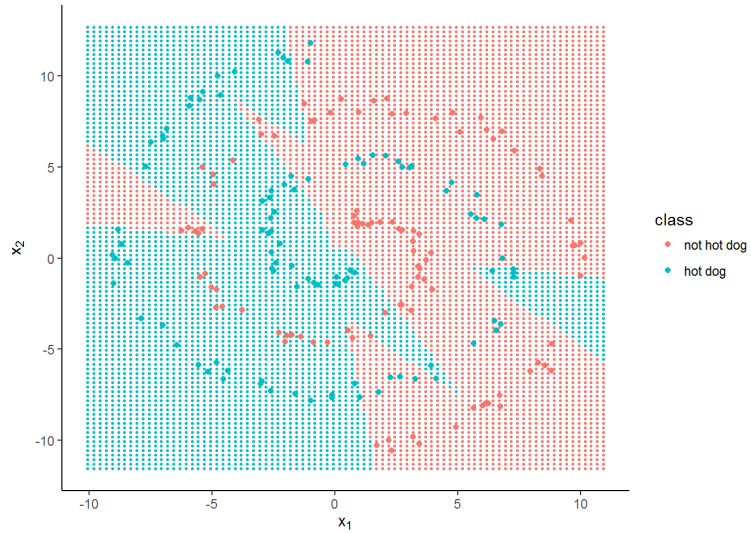
Machine learning emphasizes regularization and predictive validity so that:

- Models *grow in complexity* but regularization creates friction in doing so,
- Models are assessed in their predictive validity, typically using *hold out samples* and cross-validation.

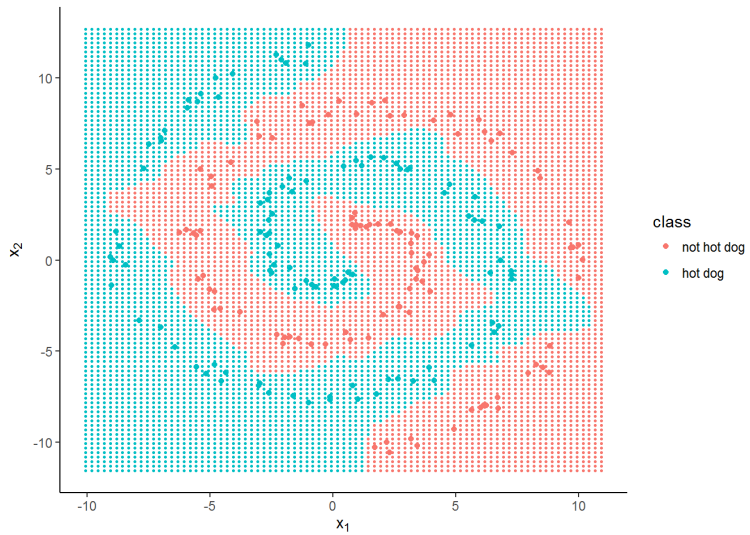
Machine Learning vs. Traditional Statistics



(i)



(ii)



(iii)

From D. Selby *Tea & Stats* blog (http://selbydavid.com/2018/01/09/neural-network/): Neural network classification regions with (i) 1 node, (ii) 5 nodes, (iii) 30 nodes.

Ensemble Learning for Retrospective Causal Inference

(Samii, Paler, and Daly, 2016)

Machine Learning and Causal Inference

Illustrations:

- CIA with high-dimensional X.
- Effect heterogeneity.

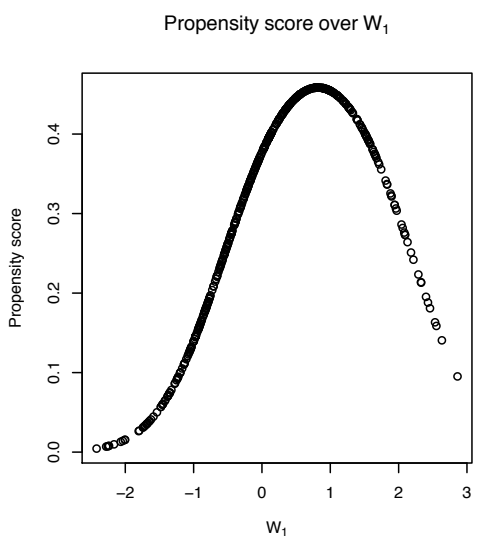
Ensemble Learning for Retrospective Causal Inference

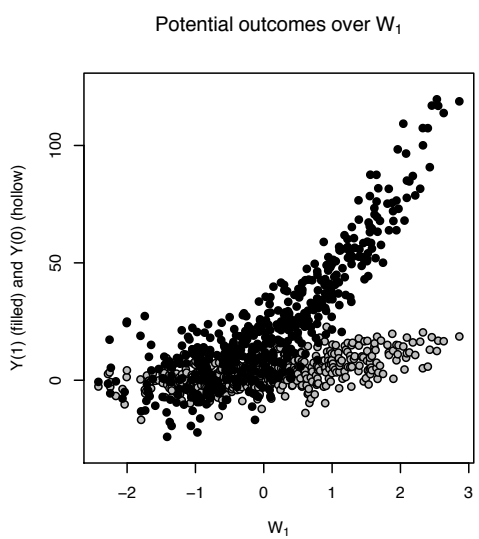
- Retrospective causal inference necessary, but thorny.
- Necessary because of slow to emerge outcomes, limits on experiments, non-availability of natural experiments for some populations.
- Thorny because of endogeneity.

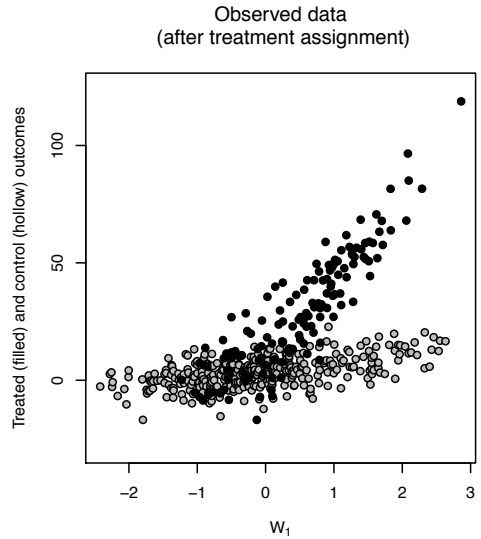
Retrospective Causal Inference

- Idea: what if we have *tons* of covariate data?
- Makes CIA more plausible.
- But implementation is challenging.
- What Xs to include? How to do it?
- Maybe machine learning can help?
  - By targeting the propensity score, we can let the machine learn an identifying covariate set and specification!
  - (In principle, could also target the potential outcome distributions, but one would have to do that separately for all outcomes of interest.)

Perils of Standard Practice, Promise of Machines







- Suppose we want to estimate the ATT.
- Suppose we have a set of covariates.
- But unbeknownst to us, treatment and outcome confounded by only *one* of them; the rest are just noise.
- And, it is confounded in an irregular way.
- How well do conventional methods do in these circumstances? How sensitive are they to increasing noise?

Perils of Standard Practice, Promise of Machines

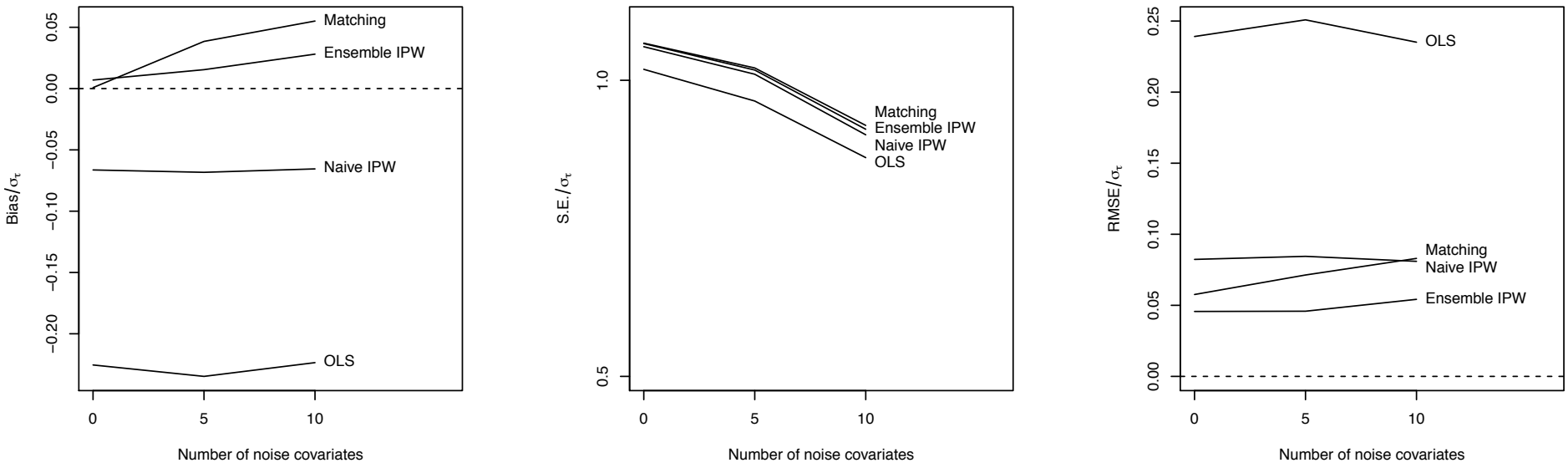
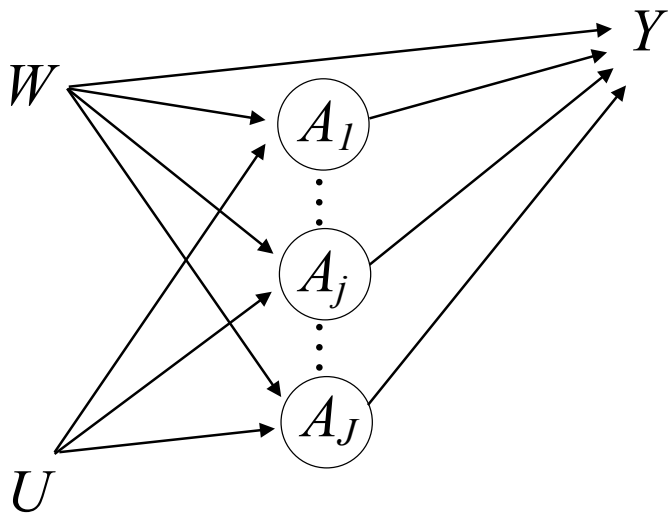


Figure 3: Simulation results. From left to right, the graphs show bias, standard error (S.E.), and root mean square error (RMSE) for the different estimators from 250 simulation runs as the number of noise covariates increases from 0 to 10. All results are standardized relative to the standard deviation of the true sample ATT across the simulation runs.

Application: Reducing Recidivism in Colombia

- What interventions might be useful to reduce ex-combatant recidivism?
- Consider a set of "hypothetical interventions":
  - Employment,
  - Ensure physical security,
  - Build trust in government,
  - Ensure good psychological health,
  - Reduce social ties to group.
- Estimate what such interventions could have accomplished, using actual variation in these risk factors, measured against what actually happened.

Assumptions



- CIA.
- Also need positivity for intervention factors of interest.

Target and Estimator

For outcome Y_k, define the retrospective intervention effect (RIE) for A_j as,

$$\psi_j = \underbrace{\mathrm{E}[Y(\underline{a}_j, A_{-j})]}_{\text{counterfactual mean}} - \underbrace{\mathrm{E}[Y]}_{\text{observed mean}} ,$$

where A_{-j} refers to elements of A other than A_j. The RIE differs slightly from the average

We use this identification result to construct an inverse-propensity score weighted (IPW) estimator of the RIE:

$$\hat{\psi}_j^{IPW} = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{I(A_{ji} = \underline{a}_j)}{\hat{g}_j(\underline{a}_j|W_i, A_{-ji})}Y_i\right) - \bar{Y} \tag{2}$$

where N is the sample size and $\hat{g}_j(\underline{a}_j|W_i, A_{-ji})$ is a consistent estimator for $\Pr[A_j = \underline{a}_j|W_i, A_{-ji}]$. In

- Need to estimate the propensity score ($\hat{g}_j(.)$).
- Have 114 covariates reduced to 23 indices on war, political, economic, and social characteristics at demobilization; 9 demographic traits; and 47 municipality fixed effects.
- Conventional approaches would buckle under this.
- Try a machine learning ensemble instead.

Propensity Score Ensemble

Ensemble: why pick one approach when you can try them all?

- Vanilla and t-regularized logistic regression (benchmarks).
- Kernel regularized least squares.
- Bayesian additive regression trees (a version of random forest).
- ν-regularized support vector machine.

Ensemble

- Kernel regularized least squares:
  - "Duality" of regression as basis expansion and regression as kernel weighted average ("kernel trick").
  - For each unit, solve for c based on

$$f(x^\star) = c_1 k(x^\star, x_1) + c_2 k(x^\star, x_2) + \ldots + c_N k(x^\star, x_N)$$

$$= c_1(\text{similarity of } x^\star \text{ to } x_1) + c_2(\text{sim. of } x^\star \text{ to } x_2) + \ldots + c_N(\text{sim. of } x^\star \text{ to } x_N).$$

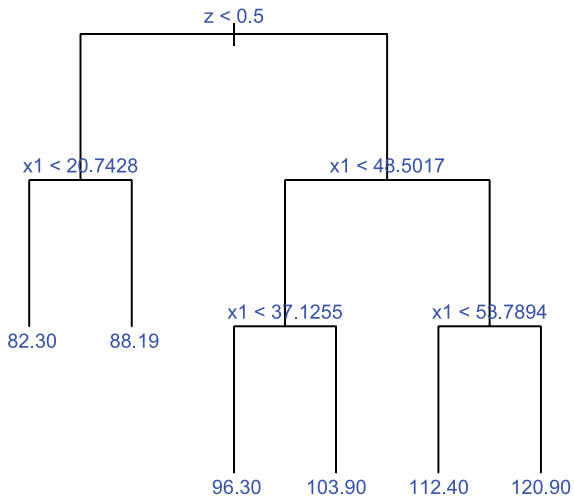
(Hainmueller & Hazlett, 2013).

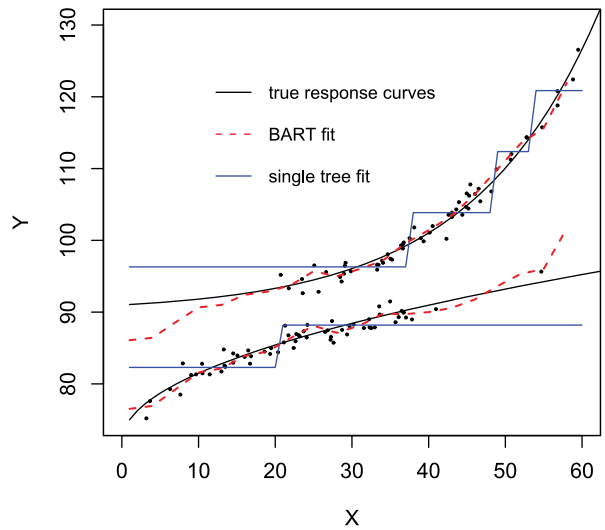
  - Regularized to penalize complexity in the c vectors.
  - Generate pscores from KRLS fits.
  - Effective for characterizing local nonlinearities and interactions.

Ensemble

- Bayesian additive regression trees:
  - Predict pscores with:

$$Y = g(z, x; T_1, M_1) + g(z, x; T_2, M_2) + \cdots + g(z, x; T_m, M_m) + \epsilon,$$





(Hill, 2011).

- Bayesian regularization to penalize tree complexity.
- Effective for characterizing nonlinearities, interactions, and non-smooth relationships.

Ensemble

- ν-support vector classification and regression:
  - Also works off the "kernel trick."
  - Fit a classifier for pscores:

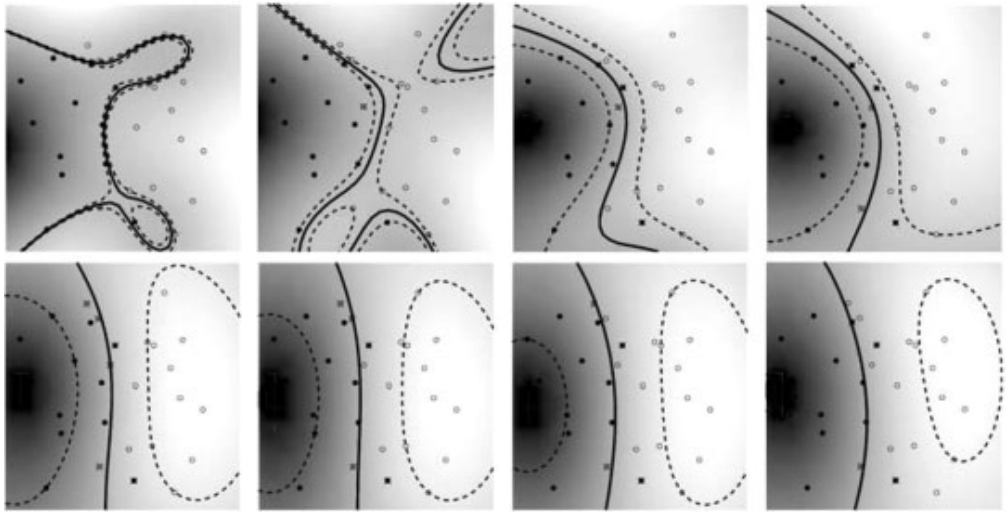


Figure 4. Toy problem (task: to separate circles from disks) solved using ν-SV classification, with parameter values ranging from $\nu = 0.1$ (top left) to 0.8 (bottom right). The larger we make ν, the more points are allowed to lie inside the margin (depicted by dotted lines). Results are shown for a Gaussian kernel, $k(x, x') = \exp(-\|x - x'\|^2)$ (from [1]).

(Chen et al., 2005).

- For binary outcomes, minimize classification error.
- *nu*-regularization to penalize complexity in the support vectors.
- Effective for characterizing nonlinearities and interactions.

Ensemble

- SuperLearner prediction takes mse-minimizing combination:

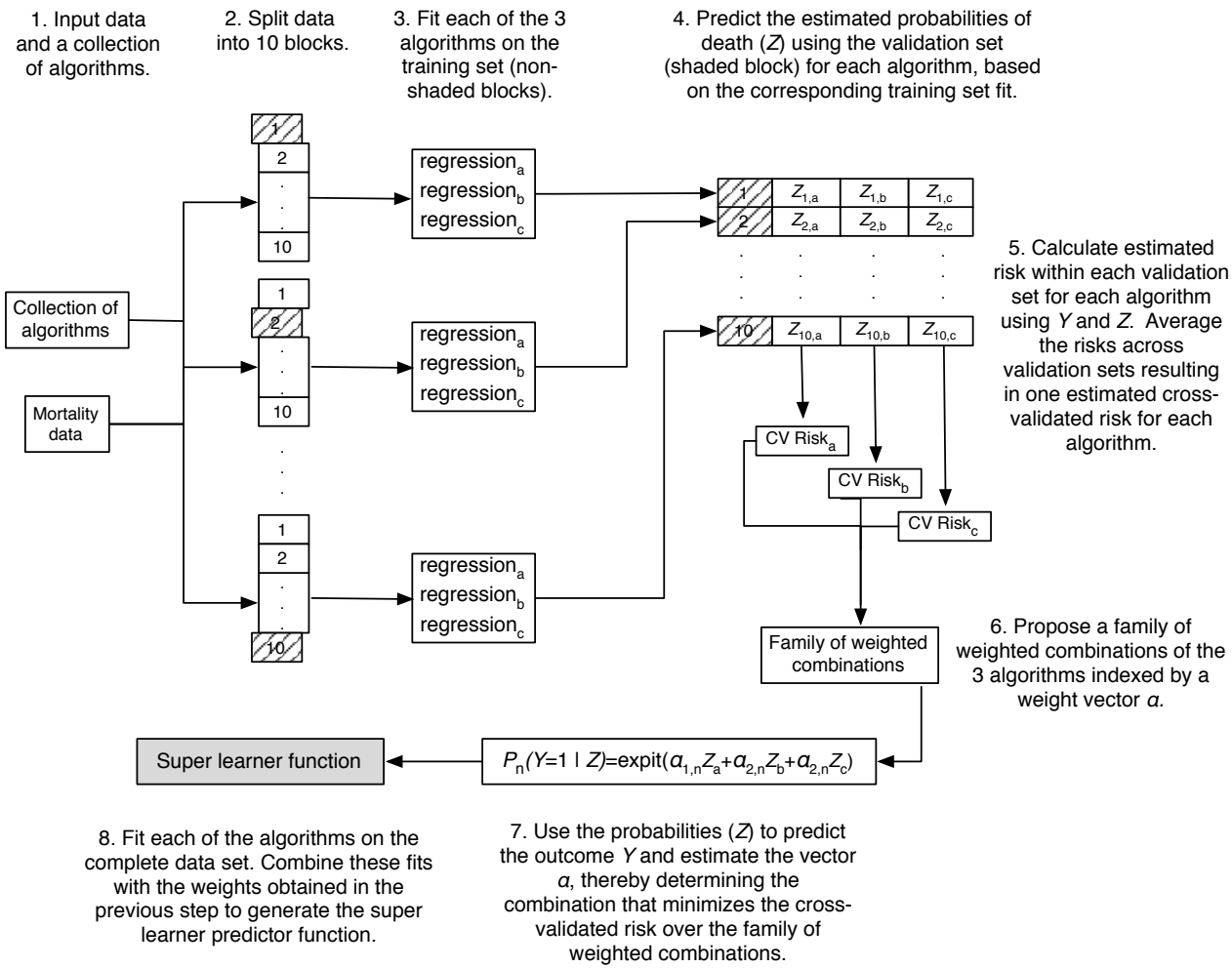


Fig. 3.2 Super learner algorithm for the mortality study example

(Van Der Laan & Rose, 2011).

Results

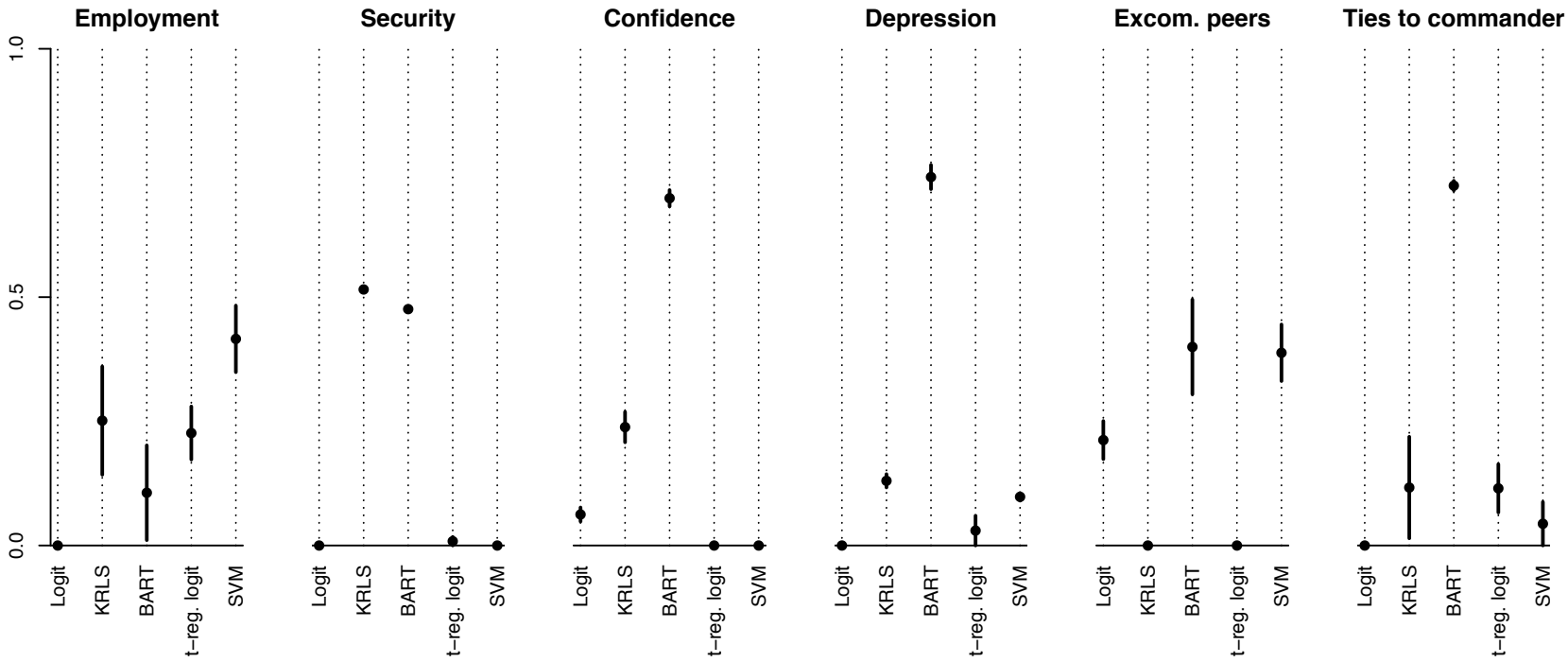
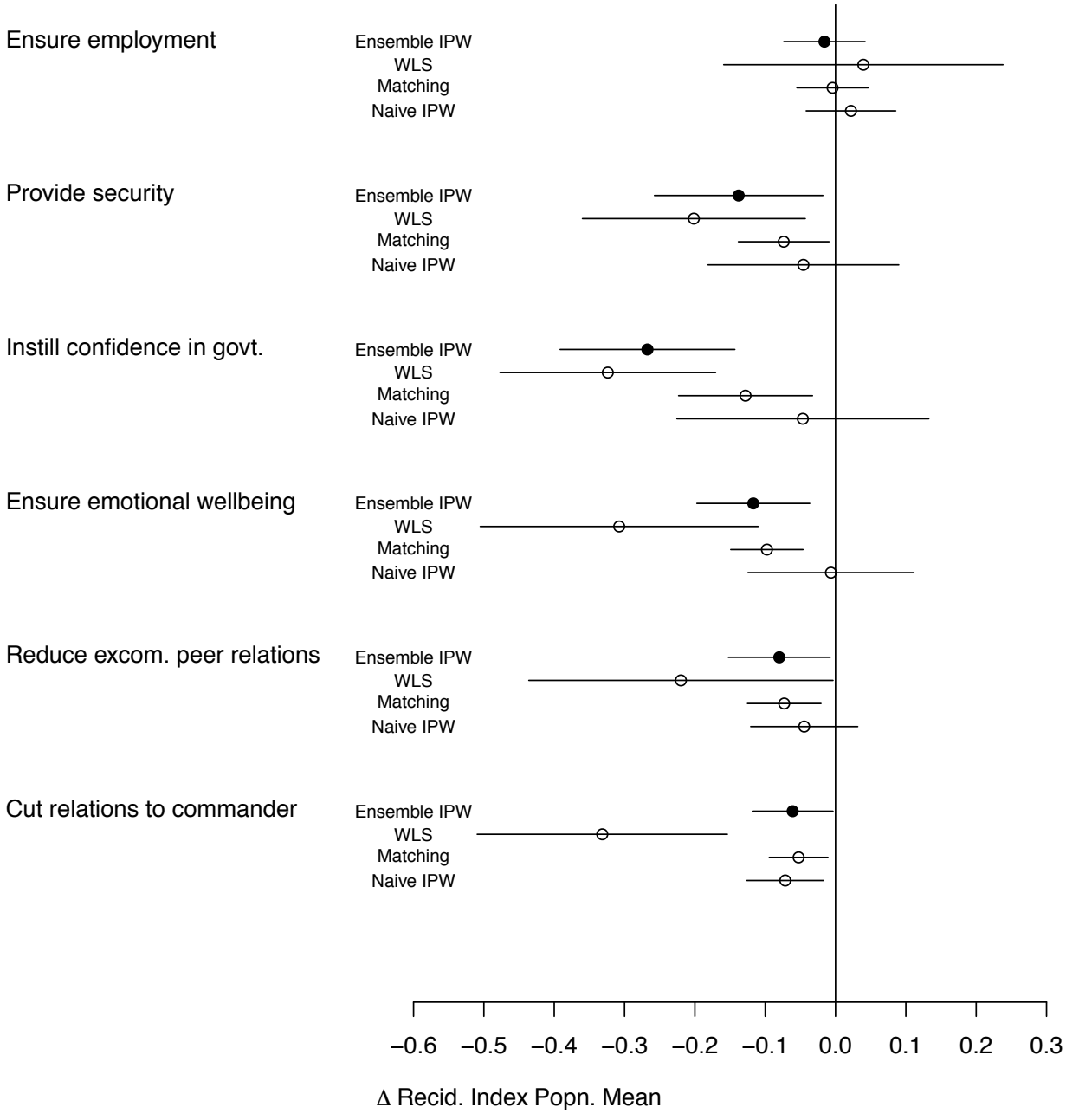


Figure 8: Weights applied to propensity score predictions from each prediction method. The values of weights run along the *y*-axis, and prediction methods run along the *x*-axis. Results are grouped by intervention. The weights are constrained to be no less than zero and to sum to one for each intervention. The black bars show the range of the weights over the 10 imputation runs, and the dots show the means.

Results


Ensure employment
Ensemble IPW
WLS
Matching
Naive IPW
Provide security
Ensemble IPW
WLS
Matching
Naive IPW
Instill confidence in govt.
Ensemble IPW
WLS
Matching
Naive IPW
Ensure emotional wellbeing
Ensemble IPW
WLS
Matching
Naive IPW
Reduce excom. peer relations
Ensemble IPW
WLS
Matching
Naive IPW
Cut relations to commander
Ensemble IPW
WLS
Matching
Naive IPW
−0.6 −0.5 −0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3
Δ Recid. Index Popn. Mean

Principles for Regularization

(Belloni, Chernozhukov, and Hansen, 2014; Chernozhukov et al. 2017)

Principles for Regularization

- Previous example illustrated potential benefits of regularization.
- Estimator is was an IPW estimator and so consistency depended on *consistent pscore estimation*. Ensemble and regularization were targeted toward this.[1]
- IPW has some attractive features, but may be less efficient than methods that incorporate outcome modeling.
- Belloni et al. (2014) develop some principles for regularization and covariate selection with methods rooted in outcome modeling.

[1] Nice discussion is here: `http://www.unofficialgoogledatascience.com/2016/06/to-balance-or-not-to-balance.html`

Principles for Regularization

- Following Belloni et al., suppose CIA and linearity.
- Outcome and treatment equation:

$$Y_i = \alpha D_i + X_i'\theta_Y + \zeta_i \quad \text{and} \quad D_i = X_i'\theta_D + \nu_i.$$

  with $\mathrm{E}[\zeta_i|D,X] = 0$, $\mathrm{E}[\nu_i|X] = 0$, $\mathrm{E}[\zeta_i\nu_i|X] = 0$.
- Implies a "reduced form" equation in terms of X:

$$\begin{aligned} Y_i &= \alpha(X_i'\theta_D + \nu_i) + X_i'\theta_Y + \zeta_i \\ &= X_i'(\alpha\theta_D + \theta_Y) + (\alpha\nu_i + \zeta_i) \\ &= X_i'\pi + \varepsilon_i \end{aligned}$$

- Parameter of interest is α. By partial regression we know:

$$\mathrm{Cov}[\nu,\varepsilon] = \mathrm{Cov}[\nu,\alpha\nu+\zeta] = \alpha\mathrm{Var}[\nu] \Leftrightarrow \alpha = \frac{\mathrm{Cov}[\nu,\varepsilon]}{\mathrm{Var}[\nu]}.$$

- Suppose we have lots of potential Xs. This motivates a machine learning approach to fit the treatment and reduced form equations.

Principles for Regularization

$$
\begin{aligned}
Y_i &= \alpha D_i + X_i'\theta_Y + \zeta_i \\
D_i &= X_i'\theta_D + \nu_i \\
Y_i &= X_i'(\alpha\theta_D + \theta_Y) + (\alpha\nu_i + \zeta_i) = X_i'\pi + \varepsilon_i.
\end{aligned}
$$

- Regularizing wrt D misses Xs for which θ_Y large, θ_D small.
- Regularizing wrt Y misses Xs for which θ_D large if α small, or for which θ_Y small.
- The "missed" Xs will lead to effects misattributed to the treatment that are actually due to selection.
- That is why we want to choose X wrt to both the D and Y equations.
- Works *if* outcome and treatment equations *are* indeed sparse.

Skepticism

Chuck: Regarding counterfactual forecasting, of course, that's out-of-population. If there is going to be a policy change, a within-population prediction says nothing per se.

Something that makes no sense to me is the work that connects machine learning and causal inference. The two phrases have nothing to do with each other as far as I can see.

Elie: This is through the use of conditional exogeneity.

Chuck: If you want to do conditional exogeneity and if you take this view that if you condition it on enough stuff...

Elie: That's what it is. That's how they connect causal inference to...

Chuck: I don't take that seriously. If that's all it amounts to, then it's just...

Elie: How would they solve the identification problem?

Chuck: They can't, obviously. It's just a marketing job that they've...

Elie: The idea is that as we are able to include more and more covariates, then conditional exogeneity becomes more palatable.

Chuck: I associate that claim with your colleague Don Rubin. Rubin has pushed this his whole career. I have had so many discussions with serious people on this over the years. We ask: "For what class of real-world problems is there a credible basis for thinking that as you add more covariates you get closer to random treatment selection?"

Elie: Yes, maybe, biology, disease. I don't know.

Skepticism

Causal inference is not just about adding Xs:

- Tamer and Manksi: adding more covariates could lead to a condition of perfect prediction of treatment, and so no overlap. Thus, adding more covariate per se is not a recipe for CIA.
- D'Amour et al. (2017): recall that CIA has an overlap condition. This limits how different covariate distributions can really be across treatment and control, either in terms of number of covariates or extent of difference for any given covariate.

Characterizing Effect Heterogeneity

Characterizing Effect Heterogeneity

- Another area of active develop is machine learning methods to characterize effect heterogeneity.
- Various uses:
  - Optimal treatment regimes (e.g., Imai & Strauss 2011).
  - Extrapolation (Hotz et al. 2005; Dehejia et al. 2017; more in future lecture).
  - Exploratory analyses (e.g., Angrist et al. 2013; Athey & Imbens 2016).

Characterizing Effect Heterogeneity

Recent example: Athey & Imbens (2016):

- Want to identify effect heterogeneity in an exploratory (not confirmatory) way, but avoid statistical recklessness.
- Using a tree approach: "deriving a partition of the population according to treatment effect heterogeneity"—a "Causal Tree."
  - Similar to what Imai and Strauss sought to do for optimal treatment regime.
- The tree is appealing for exploratory work, because you can look at full profiles of covariate distributions for the strata created by the tree partitions.
- Split sample approach to overcome challenges to estimation for treatment effects and inference based on an adaptive estimator.

Setting

- Treatment $W_i = 0, 1$, with $(Y_i(1), Y_i(0))$, $\tau_i = Y_i(1) - Y_i(0)$, and $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$.
- X_i is K-length covariate vector.
- Assume complete randomization $W_i \perp\!\!\!\perp (Y_i(1), Y_i(0), X_i)$.
- CATE : $\tau(x) = E[Y_i(1) - Y_i(0) | X_i = x]$

Partitioning

- The support of X_i is $\mathbb{X}$.
- Partition $\mathbb{X}$ using a tree $\mathscr{T}$ with $\#(\mathscr{T})$ leaves.
- Π is the partition yielded by a tree algorithm (e.g., maximize variance explained, subject to penalty for number of splits).

Partitioning

- In a prediction problem for outcomes, Y_i, on N data points, we could target MSE as,

$$MSE_Y = \frac{1}{N}\sum_i (Y_i - \hat{\mu}(\Pi(X_i)))^2,$$

  where $\hat{\mu}(\Pi(X_i))$ is the outcome mean in leaf $\Pi(X_i)$.
- Doing this for treatment effects, τ_i, is infeasible because we do not actually observe the τ_is.
- Instead, Athey and Imbens target conditional treatment effects across training and test samples, in essence

$$MSE_\tau = \frac{1}{N}\sum_{i\in\mathscr{S}^{test}} (\tau_i(\Pi(X_i);\mathscr{S}^{test}) - \tau_i(\Pi(X_i);\mathscr{S}^{train}))^2.$$

- Then, they modify the criterion to include a term that rewards variability in the conditional (leaf-specific) treatment effects.
- Finally, estimation is done on an estimation split of the sample for which the tree partition can be taken as exogenous.

Remarks

- Basic idea here is machine learning to work with large covariate space for causal inference problems (operationalizing CIA, characterizing heterogeneous effects).
- Can be applied to other problems:
  - Exploring what features of complex treatments matter (e.g., "texts" as treatments Egami et al 2018).
  - Dynamic learning of optimal treatments ("causal bandits"—Lattimore et al. 2016).
- That said, these types of tools can only *complement* or *supplement*, not replace, the randomization, conditional randomization, or discontinuities that allow for causal identification.