# Interpretable Machine Learning Methods

Zachary M. Jones

# Conventional Methods in (Observational) Social Science

- lots of linear models of various flavors fit to tabular, labelled, data (features usually have meaning)
- use of variance estimates/NHST as "importance"/validation/hurdle for publication
- little model evaluation (strange epistemology)

# Problems

- theory/experience suggests social systems are often complex and high dimensional
- little reason to trust observational models without predictive validation
- variance estimates are bogus

# Goals

- be (more) clear about our modeling goals
- assume as little as possible
- predictive validation

# Argument

machine learning methods

- (generally) weakened functional form assumptions (+)
- predictive power (+)
- computational complexity (-)
- uninterpretable black-box outputs (-)

meta-modeling can make any black-box function "interpretable"

# Contribution

- software to interpret black-box models
  - `mlr`, `edarf`, `mmpf`, `fanova` in `R`
  - would like to contribute to scikit-learn as well
- survey/re-analysis of prominent papers
- applications

# Interpretable Machine Learning

- normal ML (preprocess, tune, evaluate, deploy)
- meta-modeling on deployed model (or part of evaluation)

# Why Not Directly Estimate the Interpretable Model?

- do predictions have to be made by an interpretable model?
- what sort of interpretations are required? does all of the model need to be interpretable?
- exploratory data anlysis and/or model evaluation (contrast prediction w/ explanatory model)

# Common Interpretation Tasks

1. is this feature important
   - on its own? in combination with other features?
   - what does important mean?
2. what is shape of the relationship between this (these) feature(s) and the outcome(s)?
3. how reliable is my model's representation of these things?

# Interpretation Methods

to decompose $\hat{f}$ or $\mathcal{L}(\hat{f})$

- fit a constrained model to $\{\hat{f}, \mathbf{x}\}$ (e.g., a parametric or semi-parametric model)
- marginalize out variables iteratively (e.g., $\mathbf{x}_{23}$ to obtain $f_1(x_1)$)

# Meta-Models

partial or full factorization of $f(x_1, x_2, x_3)$

$$f_1(x1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2) + f_{23}(x_2, x_3) + f_{123}(x_1, x_2, x_3)$$

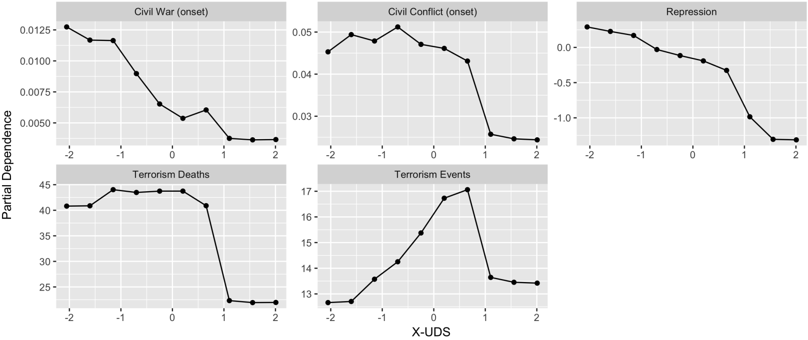
"interpretability" is treated as a function of the dimension of the factors

# Estimation

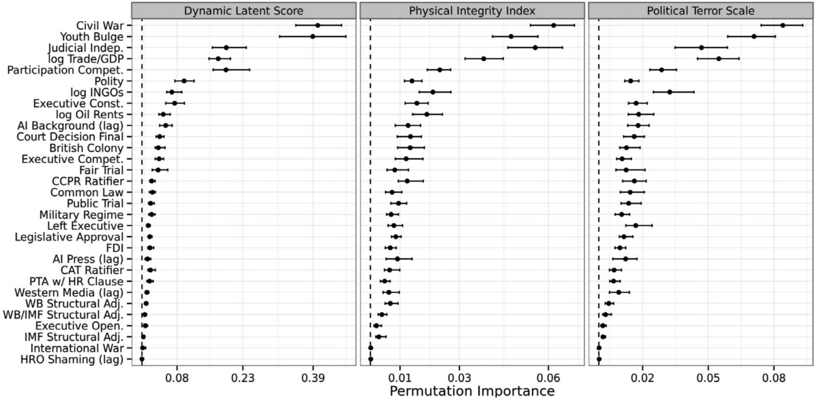
pointwise estimation on a grid (points are effects)

$$\hat{f}(\mathbf{x}) \approx g_u(x_u) + \sum_{v \subset u} g_v(x_v) + g_{-u}(x_{-u}) + \sum_{i \in u \subset v' \subseteq -i} g_{v'}(x_{v'})$$

# Political Violence (Lupu and Jones 2018)



# Repression (Hill and Jones 2014)



# Problems

- goals and how to evaluate them
- variance estimation (additional error from approximation)

# Future Problems/Directions

- different ideas of interpretability (task-specific)
- model evaluation
- variance estimation
- for python
- FATML