Bayesian Rule Set : A Quantitative Alternative to Qualitative Comparative Analysis

Albert Chiu and Yiqing Xu

Department of Political Science, Stanford University

September 30, 2021

Democratic consolidation

Which countries remain democratic?

Modernization theory

- Wealth, industrialization, education, urbanization
- Which variables matter? For whom?
- **Heterogeneous** treatment effects

Regression

- OLS, LASSO, MLE, Bayesian, etc.
- Common trait: effects are **marginal** and **constant**

Regression

- OLS, LASSO, MLE, Bayesian, etc.
- Common trait: effects are **marginal** and **constant**
- Can relax this assumption at a cost
- E.g., interactions:
  $\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{123} X_1 X_2 X_3$
  - Uninterpretability
  - Dimensionality & model selection: # terms is exponential

Rule Sets as a Classifier

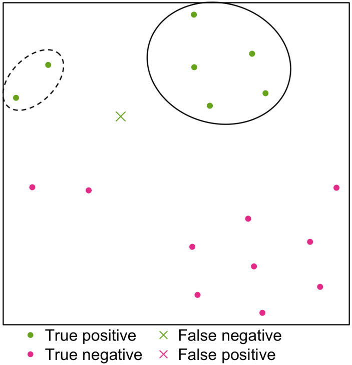
- If-Then statements to classify data
- Qualitative Comparative Analysis (QCA):

  IF (High Wealth) OR (Medium Wealth AND Low Industrialization)
  THEN Stable Democracy

Rule Sets as a Classifier

- If-Then statements to classify data
- Qualitative Comparative Analysis (QCA):

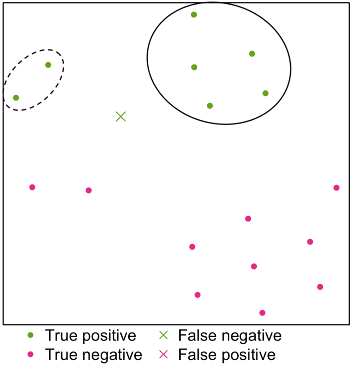
  IF (High Wealth) OR (Medium Wealth AND Low Industrialization) THEN Stable Democracy



Rule Sets as a Classifier

- If-Then statements to classify data
- Qualitative Comparative Analysis (QCA):

IF (High Wealth) OR (Medium Wealth AND Low Industrialization)
THEN Stable Democracy



QCA can't handle errors

- Discards data
- Complex rule sets: uninterpretable, overfitted
- Computationally infeasible

An Alternative Method For Learning Rule Sets

A number of ways, e.g. decision trees (but not Random Forest)

An Alternative Method For Learning Rule Sets

A number of ways, e.g. decision trees (but not Random Forest)
Bayesian Rule Sets (BRS) (Wang et al., 2016)

- Compatible with errors; uses all data
- Maintains sparsity/parsimony
- Computationally Feasible

An Alternative Method For Learning Rule Sets

A number of ways, e.g. decision trees (but not Random Forest)
Bayesian Rule Sets (BRS) (Wang et al., 2016)

- Compatible with errors; uses all data
- Maintains sparsity/parsimony
- Computationally Feasible

Contributions

- Improve BRS
- Uncertainty and stability for rule sets
- Graphical tools

Overview

1. Motivation
2. Method
   - Bayesian Rule Set (BRS)
   - Bootstrapping Rule Sets
   - Graphical Tools
3. Monte Carlo Simulation
4. A Large-N, Large-p Empirical Example: Voter Turnout

BRS: Setup

- Goal: given hyper-parameters H and data S, find rule set A that maximizes posterior (MAP)

BRS: Setup

- Goal: given hyper-parameters H and data S, find rule set A that maximizes posterior (MAP)
- Rule set: e.g. If (A and B) or (C) then Y=1
  - $[(A \cap B) \cup (C)] \subseteq Y^+$

BRS: Setup

- Goal: given hyper-parameters H and data S, find rule set A that maximizes posterior (MAP)
- Rule set: e.g. If (A and B) or (C) then Y=1
  - $[(A \cap B) \cup (C)] \subseteq Y^+$
- Binary outcome, discrete data
- User specifies hyper-parameters

BRS: Setup

- Goal: given hyper-parameters H and data S, find rule set A that maximizes posterior (MAP)
- Rule set: e.g. If (A and B) or (C) then Y=1
  - $[(A \cap B) \cup (C)] \subseteq Y^+$
- Binary outcome, discrete data
- User specifies hyper-parameters
- Prior controls sparsity, likelihood controls performance

BRS: Likelihood

- $\rho_+ \sim \text{Beta}(\alpha_+, \beta_+)$
- $\rho_- \sim \text{Beta}(\alpha_-, \beta_-)$
- $$y_n | x_n, A \sim \begin{cases} \text{Bernoulli}(\rho_+) & \text{if } x_n \in A \\ \text{Bernoulli}(1 - \rho_-) & \text{if } x_n \notin A. \end{cases}.$$

BRS: Likelihood

- $\rho_+ \sim$Beta(α_+, β_+)
- $\rho_- \sim$ Beta(α_-, β_-)
-

$$y_n | x_n, A \sim \begin{cases} \text{Bernoulli}(\rho_+) & \text{if } x_n \in A \\ \text{Bernoulli}(1 - \rho_-) & \text{if } x_n \notin A. \end{cases}.$$

- Choose α_ξ large and β_ξ small so $E[\rho_\xi] = \frac{\alpha_\xi}{\alpha_\xi + \beta_\xi} \approx 1$, $\xi \in \{-, +\}$

BRS-Poisson: Priors

- Modified from Wang et al. (2017)

BRS-Poisson: Priors

- Modified from Wang et al. (2017)
- Pick number of rules $M \sim$Poisson(λ)

BRS-Poisson: Priors

- Modified from Wang et al. (2017)
- Pick number of rules $M \sim$Poisson(λ)
- For $m = 1, 2, ..., M$:
  - Pick length of mth rule $L_m \sim$ Truncated-Poisson(η)

BRS-Poisson: Priors

- Modified from Wang et al. (2017)
- Pick number of rules $M \sim$Poisson(λ)
- For $m = 1, 2, ..., M$:
  - Pick length of mth rule $L_m \sim$ Truncated-Poisson(η)
  - For $j = 1, 2, \ldots, L_m$:
    - Pick variable V_j uniformly at random
    - Pick value w_j of variable uniformly at random

BRS-Poisson: Priors

- Modified from Wang et al. (2017)
- Pick number of rules $M \sim$Poisson(λ)
- For $m = 1, 2, ..., M$:
  - Pick length of mth rule $L_m \sim$ Truncated-Poisson(η)
  - For $j = 1, 2, \ldots, L_m$:
    - Pick variable V_j uniformly at random
    - Pick value w_j of variable uniformly at random
  - rule $a_m = \bigcap_j \{V_j = w_j\}$
- Rule set $A = \bigcup_m a_m$

Hyper-parameters

Well behaved penalties

Hyper-parameters

Well behaved penalties

- Penalty for rule length $\phi(\eta) > 0$ for $\eta < 2$
- Penalty for number of rules $\psi(\lambda, \eta) > 0$ for $\lambda \lesssim 1.47$
- ϕ always strictly decreasing function of η
- ψ strictly decreasing function of η for any λ and for $\eta < 2$

Hyper-parameters

Well behaved penalties

- Penalty for rule length $\phi(\eta) > 0$ for $\eta < 2$
- Penalty for number of rules $\psi(\lambda, \eta) > 0$ for $\lambda \lesssim 1.47$
- ϕ always strictly decreasing function of η
- ψ strictly decreasing function of η for any λ and for $\eta < 2$

Linear search over η: start w/ $\lambda = \eta = 1$, decrease η to penalize complexity more

Hyper-parameters

Well behaved penalties

- Penalty for rule length $\phi(\eta) > 0$ for $\eta < 2$
- Penalty for number of rules $\psi(\lambda, \eta) > 0$ for $\lambda \lesssim 1.47$
- ϕ always strictly decreasing function of η
- ψ strictly decreasing function of η for any λ and for $\eta < 2$

Linear search over η: start w/ $\lambda = \eta = 1$, decrease η to penalize complexity more

If "too" sparse, strengthen likelihood: multiply α_ξ, β_ξ by $c > 1$

Algorithm For Inference

- Enormous search space; bounds to reduce it
- Intuition: can only have a few rules, each has to cover many cases

Algorithm For Inference

- Enormous search space; bounds to reduce it
- Intuition: can only have a few rules, each has to cover many cases
- “Approximate” algorithm: cull rules at beginning w/ arbitrary cutoff

Algorithm For Inference

- Enormous search space; bounds to reduce it
- Intuition: can only have a few rules, each has to cover many cases
- "Approximate" algorithm: cull rules at beginning w/ arbitrary cutoff
- Any search procedure (e.g. simulated annealing – balances greediness w/ exploration, avoid local maxima)

Quantifying Uncertainty

Confidence/credible set/collection infeasible to find, uninterpretable

- Maximum density $\rightarrow$ sort exponentially many rule sets
- Can't summarize using, e.g., end points

Quantifying Uncertainty

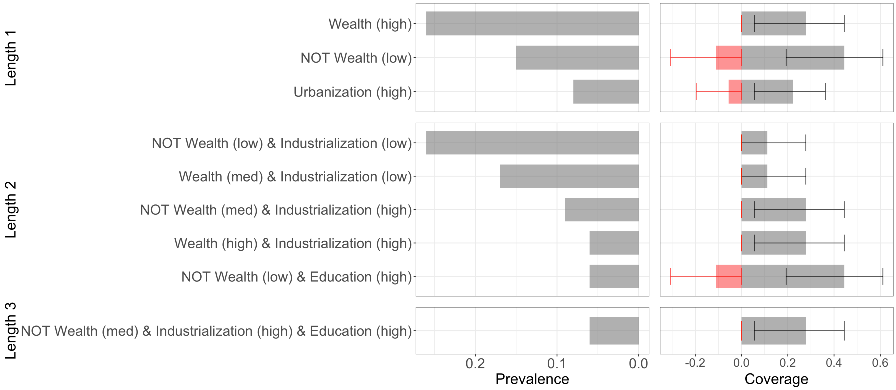
Confidence/credible set/collection infeasible to find, uninterpretable

- Maximum density $\rightarrow$ sort exponentially many rule sets
- Can't summarize using, e.g., end points

Alternative: bootstrapping

- *Prevalence:* proportion of times a rule appears in solution
- *Coverage:* proportion of points covered by rule (bootstrap CI)

Quantifying uncertainty



Stabilizing Results

Small changes in numerical results typically not substantively meaningful

- e.g., $\beta = 1$ vs. $\beta = 1.1$

Small changes in rule sets can be meaningful

- e.g., (A and B and C) vs. (A and B and D)

Stabilizing Results

Small changes in numerical results typically not substantively meaningful

- e.g., $\beta = 1$ vs. $\beta = 1.1$

Small changes in rule sets can be meaningful

- e.g., (A and B and C) vs. (A and B and D)

Instability due to:

- Failure to converge
- Perturbations in data

Stabilizing Results

Small changes in numerical results typically not substantively meaningful

- e.g., $\beta = 1$ vs. $\beta = 1.1$

Small changes in rule sets can be meaningful

- e.g., (A and B and C) vs. (A and B and D)

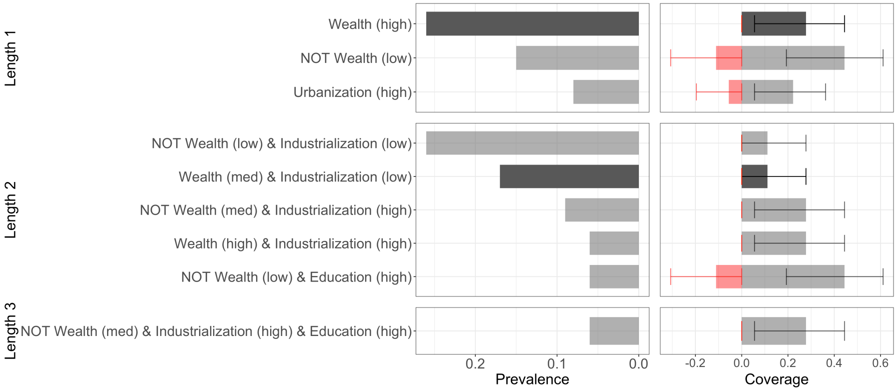
Instability due to:

- Failure to converge
- Perturbations in data

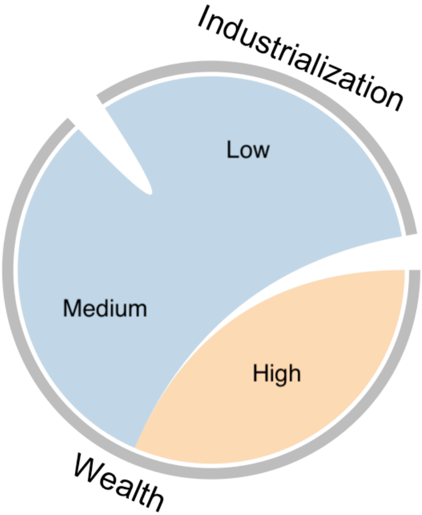
Solution: aggregate high prevalence rules

- Combine rules $\rightarrow$ rule set
- Maximize, e.g., accuracy using at most 3 rules

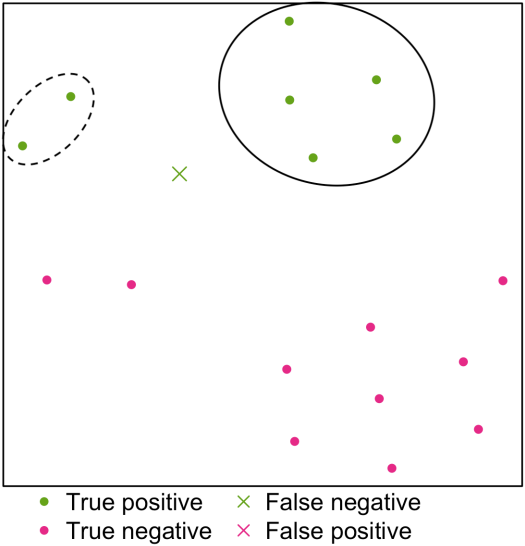
Bar Plots


Length 1
Wealth (high)
NOT Wealth (low)
Urbanization (high)
Length 2
NOT Wealth (low) & Industrialization (low)
Wealth (med) & Industrialization (low)
NOT Wealth (med) & Industrialization (high)
Wealth (high) & Industrialization (high)
NOT Wealth (low) & Education (high)
Length 3
NOT Wealth (med) & Industrialization (high) & Education (high)
0.2
0.1
0.0
Prevalence
-0.2
0.0
0.2
0.4
0.6
Coverage

Chord Diagram



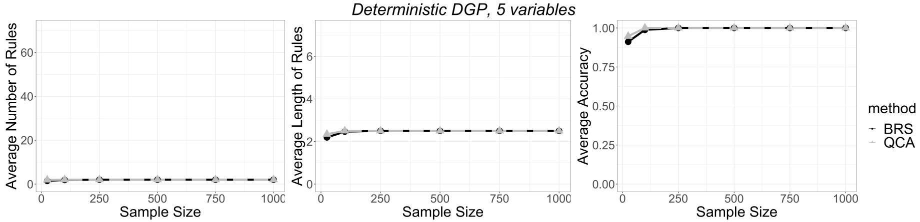
t-SNE Plots


True positive
False negative
True negative
False positive

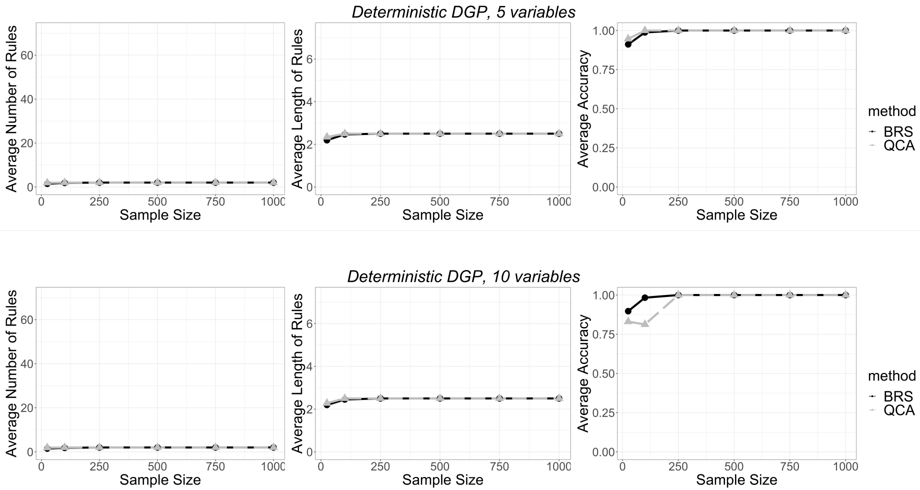
Simulation Setup

- N=25 to 1000
- 5, 10, 20 binary variables
- binary outcome, either deterministic or probabilistic
- True rule set $A^* = (V_1 \cap V_2) \cup (V_3 \cap V_4 \cap V_5^C)$
- $P(y_n = 1 | x_n \in A^*) \in \{1, .75\}$
- $P(y_n = 1 | x_n \notin A^*) \in \{0, .25\}$

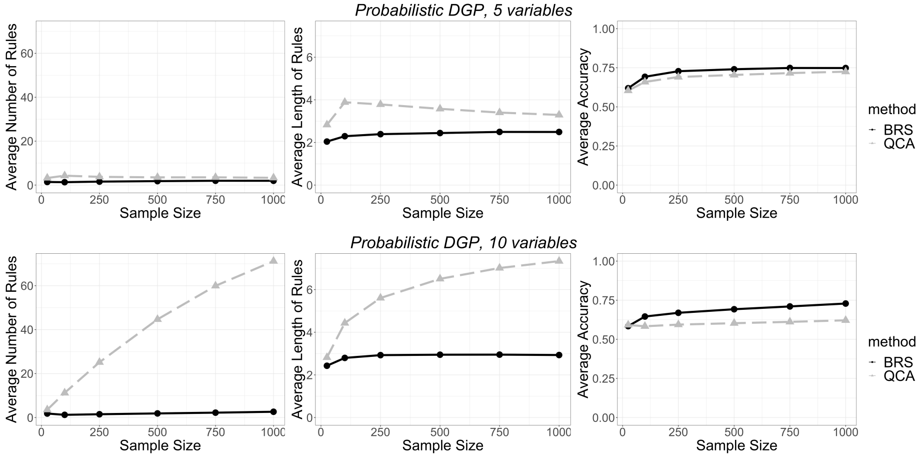
Simulation Results



Simulation Results



Simulation Results



Voter Turnout

Landwehr and Ojeda (2021): regression to estimate the effect of depression on voter turnout

- $N = 1,014$, $p = 13$

Voter Turnout

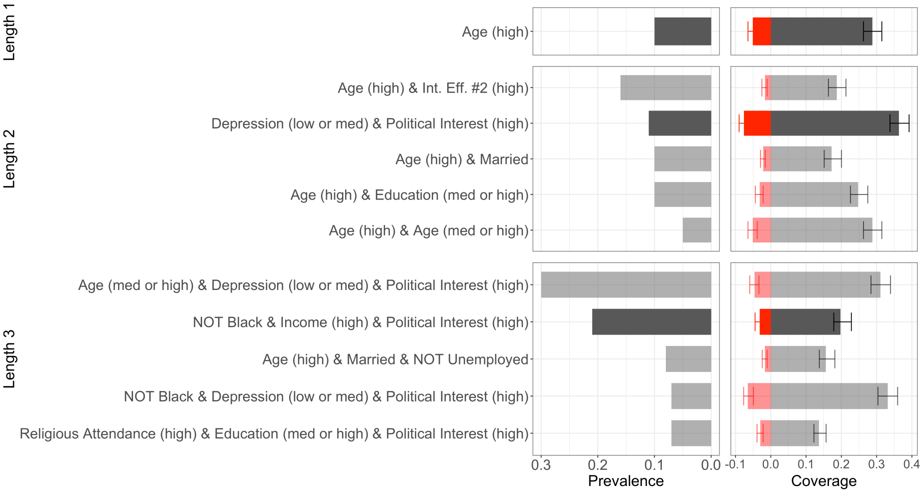
Landwehr and Ojeda (2021): regression to estimate the effect of depression on voter turnout

- $N = 1,014$, $p = 13$

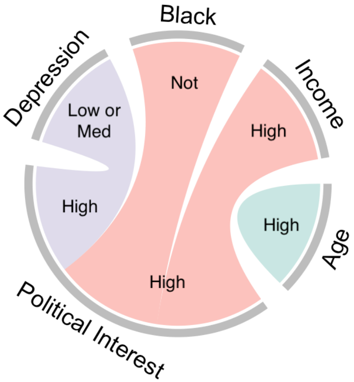
Task of discovery/theory building:

- Who votes
- Which variables are predictive; for whom

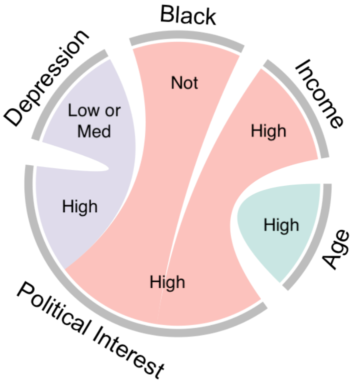
Voter Turnout



Voter Turnout



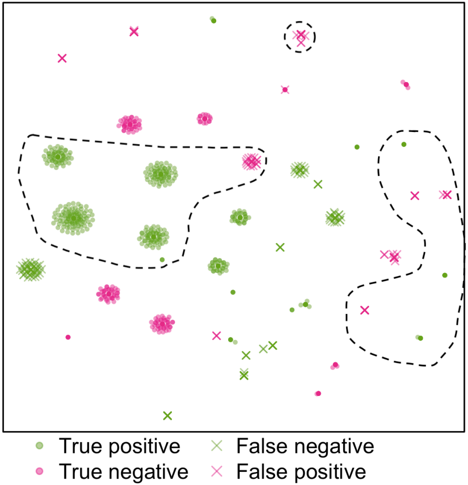
Voter Turnout



One interpretation:

- High age alone is highly predictive; don't need other factors
- Amongst younger, political interest is important but not always enough:
  - Depression
  - Race+class

Voter Turnout



Dashed lines encircle "Depression (low or med) and Political Interest (high)"

Conclusion

- Rule sets can interpretably describe complex relations (better than regression)
- Theory building, data description

Conclusion

- Rule sets can interpretably describe complex relations (better than regression)
- Theory building, data description
- QCA fails when data is large and heterogeneous
- BRS solves some of QCA's problems

Conclusion

- Rule sets can interpretably describe complex relations (better than regression)
- Theory building, data description
- QCA fails when data is large and heterogeneous
- BRS solves some of QCA's problems
- Contributions
    - BRS priors/hyper-parameters: computation, interpretation, ease of use
    - Rule sets: uncertainty and stability
    - Graphical tools

References

Landwehr, Claudia and Christopher Ojeda. 2021. "Democracy and depression: a cross-national study of depressive symptoms and nonparticipation." *American Political Science Review* 115(1):323–330.

Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl and Perry MacNeille. 2017. "A Bayesian framework for learning rule sets for interpretable classification." *The Journal of Machine Learning Research* 18(1):2357–2393.

Wang, Tong, Cynthia Rudin, Finale Velez-Doshi, Yimin Liu, Erica Klampfl and Perry MacNeille. 2016. Bayesian rule sets for interpretable classification. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE pp. 1269–1274.