A Clustering Approach for Characterizing Voter Types: An Application to High-Dimensional Ballot and Survey Data

Shiro Kuriwaki (Harvard University)

July 2020

Paper URL: https://osf.io/v3rhz
Author Homepage: https://www.shirokuriwaki.com

Motivation for Clustering

Social scientists and political journalists want to group the electorate into **clusters**:

- by vote choice: e.g. Swing voters
- by demographics: e.g. Soccer moms

But vote data hard to summarize due to large number of combinations

Data on Vote Choice

- ☑ Individual-level
- ☑ Unordered outcomes
  (e.g. "D", "R", "Green", "Abstain")
- ☑ High-dimensional
  (e.g., Congress, Governor, Sheriff)

Existing approaches

- Focus on one pair of offices at a time
- Show all pair-wise correlations
- Impose an ideal point model

```
VOTER A
-------------------------------------------------------------
 10 Hillary Rodham Clinton        President and Vice President
 18 Thomas Dixon                  U.S. Senate
 27 Dimitri Cherny                CON0001 U.S. House of Representatives
 53 Peter J Tecklenburg           Auditor
 61 Andrew C Smith.               County Treasurer
104 No                            Question 1
107 No                            Question 2

VOTER B
-------------------------------------------------------------
 14 Donald J Trump                President and Vice President
 22 Tim Scott                     U.S. Senate
 30 Mark Sanford                  CON0001 U.S. House of Representatives
 35 Stephen Goldfinch             SEN0034 State Senate
 38 Lee Hewitt                    HOU0108 State House of Representatives
 41 Scarlett Wilson               Solicitor Circuit 9
 44 Al Cannon                     Sheriff
 47 Julie J Armstrong             Clerk of Court
 50 Rae H Wooten                  Coroner
 54 Elizabeth Moffly              Auditor
 59 Mary E Tinkler.               County Treasurer
 64 John H Smoak                  Soil and Water District Commission
 68 Tony E Lewis                  CSB City of Charleston
 71 Chris Collins                 CSB North Area
 74 Louis L Smith                 CSB North Area
 78 Priscilla Jeffery             CSB West Ashley
 80 Michael Miller                CSB West Ashley
 84 Anita Renee Alston-Gore       SCH0001 Con SB St James Santee District
 88 Juanita M Middleton           SCH0001 Con SB St James Santee District
 89 Samuel N Robinson             SCH0001 Con SB St James Santee District
104 No                            Question 1
107 No                            Question 2

VOTER C
-------------------------------------------------------------
  1 Democratic                    STRAIGHT PARTY
 10 Hillary Rodham Clinton        President and Vice President
 18 Thomas Dixon                  U.S. Senate
 27 Dimitri Cherny                CON0001 U.S. House of Representatives
 44 Al Cannon                     Sheriff
 47 Julie J Armstrong             Clerk of Court
 53 Peter J Tecklenburg           Auditor
 58 Mary Tinkler                  County Treasurer
 84 Anita Renee Alston-Gore       SCH0001 Con SB St James Santee District
 86 Thomas L Colleton Jr          SCH0001 Con SB St James Santee District
 87 Erma J Harrell                SCH0001 Con SB St James Santee District
103 Yes                           Question 1
106 Yes                           Question 2
```

Pictured: Cast vote records from South Carolina

EM Algorithm for Clustering Categorical Data

Suppose voter i belongs to an unobserved cluster $Z_i \in \{1, ..., K\}$

We only observe their vote $Y_{ij} \in \{\text{abstain}, \text{split}, \text{straight}\}$ in office j.

Quantities of interest:

- How large is cluster k?

$$\pi_k \equiv \Pr\left(Z_i = k\right)$$

- How likely is cluster k to split in office j?

$$\mu_{(k,j,\text{split})} \equiv \Pr\left(Y_{ij} = \text{split} \mid Z_i = k\right)$$

- Given their vote pattern, is voter i in cluster k?

$$\zeta_{ik} = \Pr\left(Z_i = k \mid \mathbf{Y}_i, \boldsymbol{\pi}, \boldsymbol{\mu}\right)$$

Model of vote choice: given cluster, offices are independent

$$\text{Likelihood (Voter in cluster } k \text{ votes } \ell \text{ in office } j) = \prod_{j=1}^{J} \prod_{\ell} \left(\mu_{kj\ell}\right)^{\mathbf{1}\left(Y_{ij}=\ell\right)}$$

where $\ell \in \{\text{abstain}, \text{split}, \text{straight}\}$

Estimation Challenge:

Find values of $\boldsymbol{\mu}, \boldsymbol{\pi}$ that are most likely given the model and data

... with a multinomial logit IIA assumption for varying choice sets due to uncontested races

$$\Pr\left(Y_{ij} = \ell \mid Z_i = k\right) = \frac{\exp\left(\psi_{kj\ell}\right)}{\sum_{\ell' \in \mathcal{Y}_{ij}} \exp\left(\psi_{kj\ell'}\right)}$$

... with allowing covariates to predict cluster classification

$$\pi_{ik} = \frac{\exp\left(\boldsymbol{X}_i^\top \gamma_k\right)}{\sum_{k'=1}^{K} \exp\left(\boldsymbol{X}_i^\top \gamma_{k'}\right)},$$

where $\boldsymbol{X}$ is a $N \times P$ matrix of voter-level covariates.

Expectation Maximization (EM) algorithm (via new pacakge, `clusterCVR`)

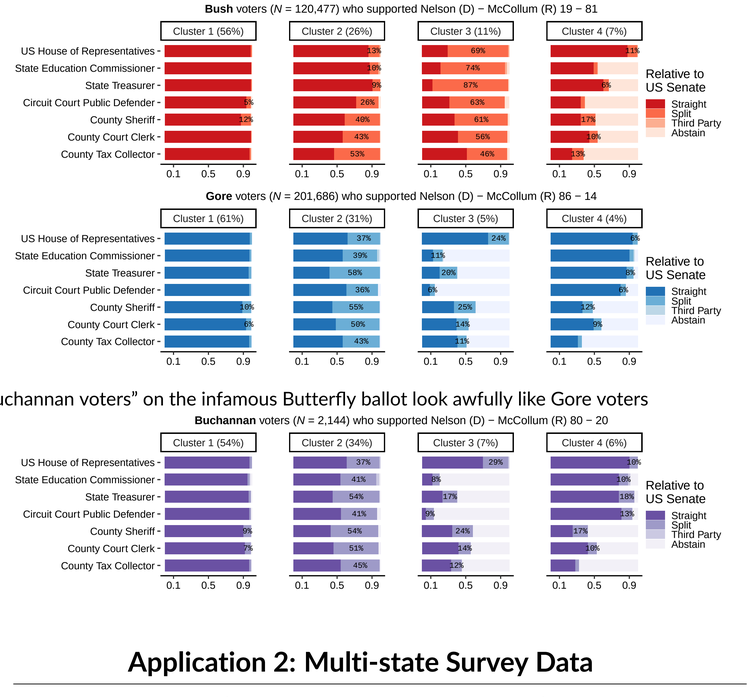
Guess $Z_i \rightarrow$ update parameters by MLE $\rightarrow$ update $\mathbf{E}[Z_i = k \mid \mathbf{Y}_i]$, ..., ⟳ until convergence.

Application 1: Cast Vote Records (CVRs)

Re-analyze Herron and Lewis (2007) ballot data in Florida's 2000 Election.

Data: Palm Beach County, recoded "Straight party", "Split ticket", "Third party", or "Abstain"

Cluster analysis of CVRs show more ticket splitting down-ballot



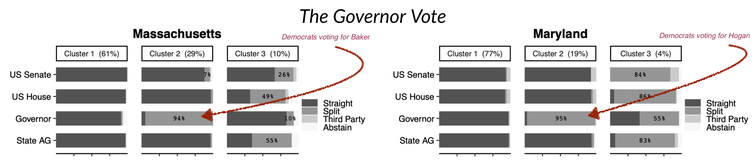
"Buchannan voters" on the infamous Butterfly ballot look awfully like Gore voters

Application 2: Multi-state Survey Data

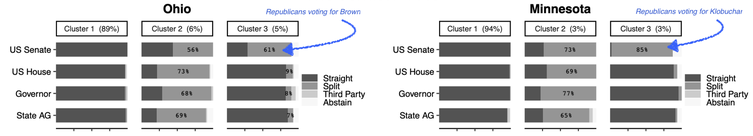
Data: 2018 CCES in states with US Senate, US House, Governor, and State Attorney General

Alhough most voters vote the party line, popular Governors and Senators have a personal vote

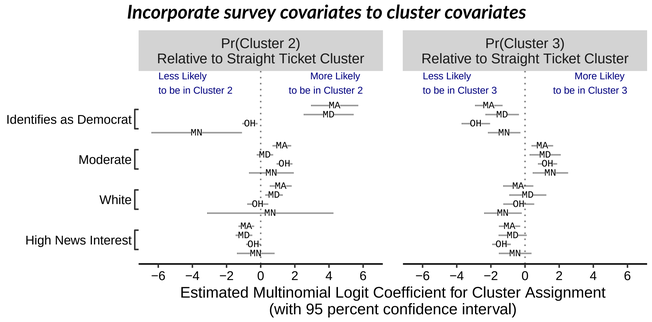
The Governor Vote



The Senator Vote



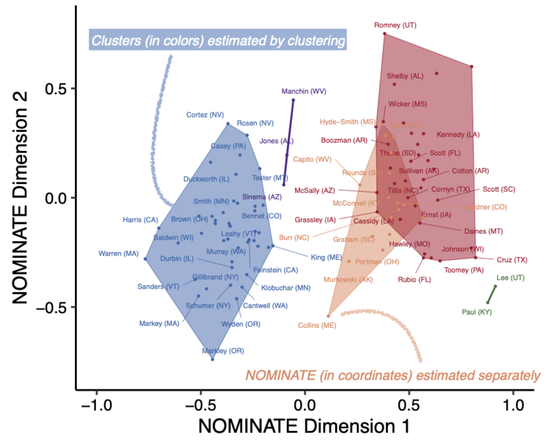
Incorporate survey covariates to cluster covariates



Comparison with Existing IRT Methods

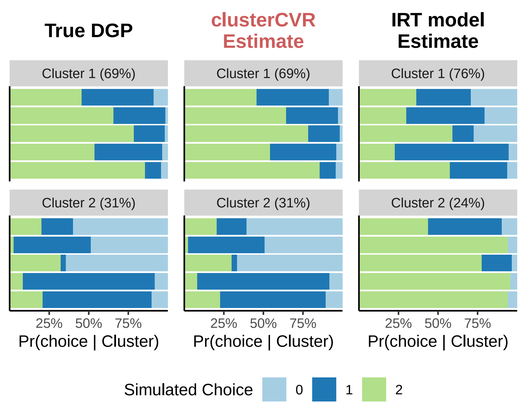
Most existing work summarizes high-dimensional vote choice data with **ideal points** (IRT).

Clustering recovers reasonable clusters in rollcall votes, similar to NOMINATE



Source: Voteview.
Covers 406 votes in 2019.
clusterCVR estimated with 5 clusters.

In simulated data, clustering model recovers true DGP more reliably



Run IRT with

- One dimensional `emIRT` on binary data (with abstain)
- Then cluster ideal points by k-means.
- Characterize choice probabilities by sample means (as in clustering)

Cautions on Interpretation

- Interpretation depends on *number* of clusters users to pick
- Interpretation and labelling requires substantive knowledge

Selected References

- Kuriwaki, Shiro (2020), "Party Loyalty on the Long Ballot: Is Ticket Splitting More Prevalent in State and Local Elections?" (10.31235/osf.io/bvgz3)
- Linzer, Drew, and Jeff Lewis (2011). "`poLCA`: An R Package for Polytomous Variable Latent Class Analysis". *JSS*
- Yamauchi, Soichiro (2020). `emlogit`: A ECM algorithm for the Multinomial Logit Model.