Latent Structures in Vote Data
COS 424 Final Project

Shun Yamaya
Joe Bartusek

Data Sources

Ballot Image Data from South Carolina

- Population data set of complete vote choices of South Carolina Electorate
- N = 1,569,831

L2 Voter File Data

- Precinct Aggregates of proportion of senior, non-white, female, college educated, citizens and logged median income

Methods

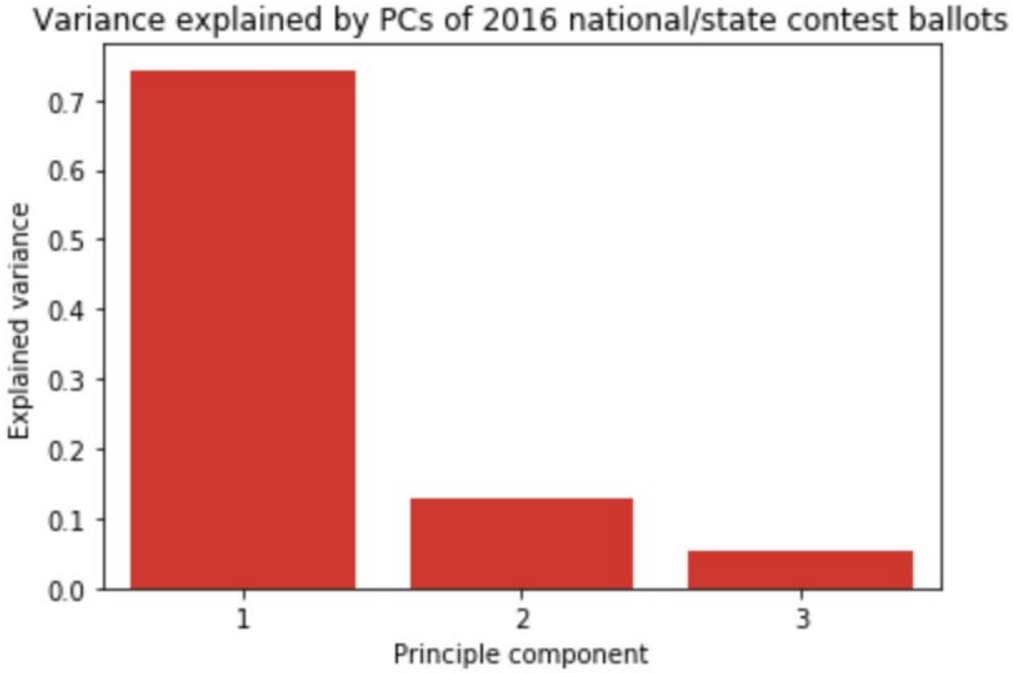
Unsupervised methods

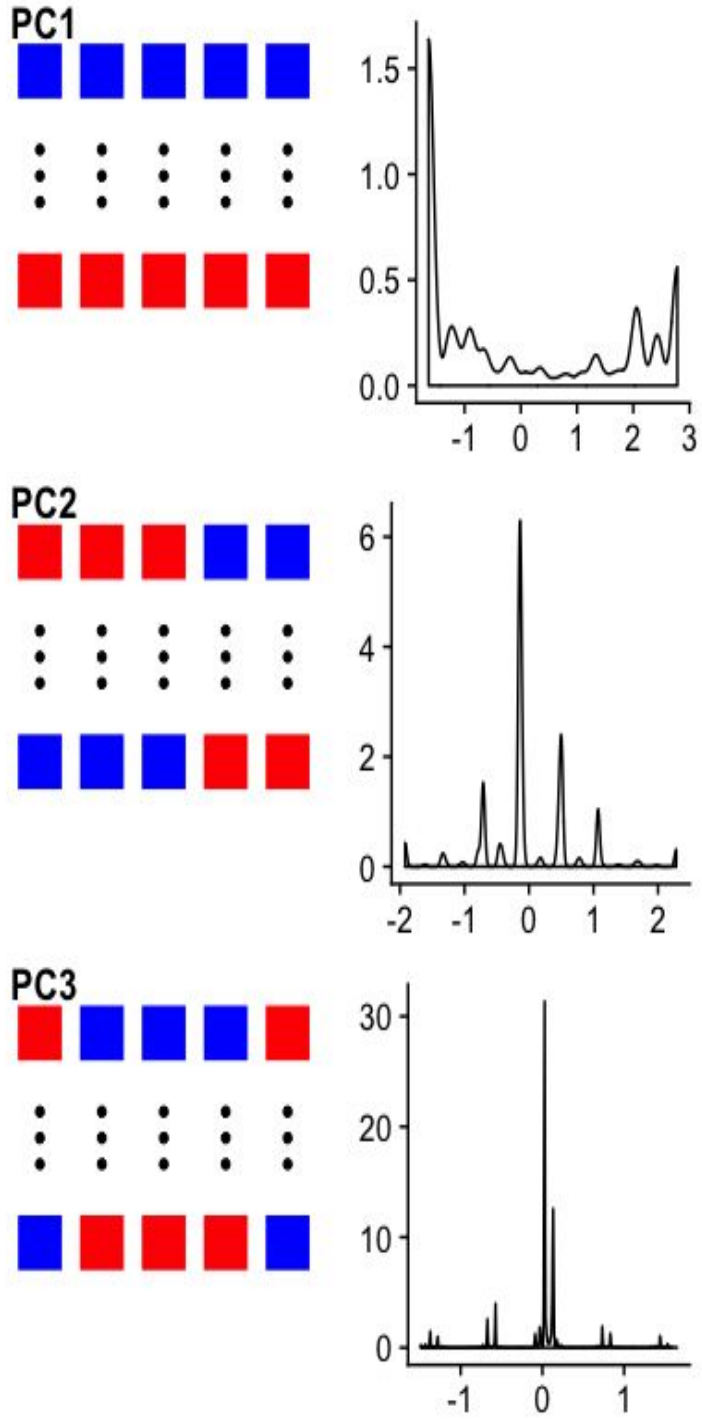
- Principal Component Analysis (main results)
- Latent Dirichlet Allocation
- Non-Negative Matrix Factorization

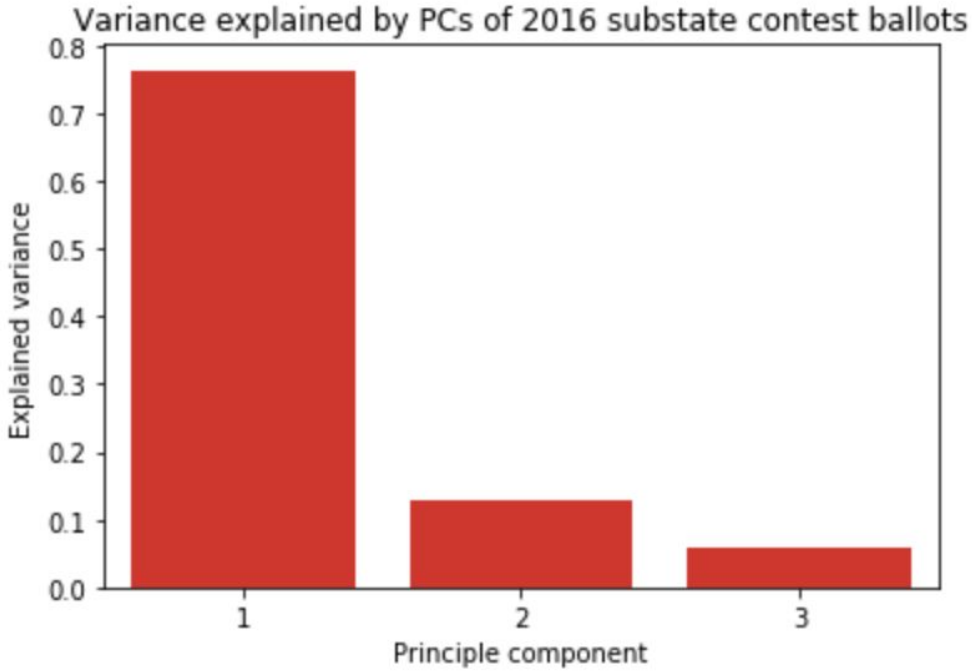
Prediction

- Multivariate OLS

Results


Variance explained by PCs of 2016 national/state contest ballots
Explained variance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
Principle component


PC1
1.5
1.0
0.5
0.0
-1
0
1
2
3
PC2
6
4
2
0
-2
-1
0
1
2
PC3
30
20
10
0
-1
0
1


Variance explained by PCs of 2016 substate contest ballots
Explained variance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
2
3
Principle component

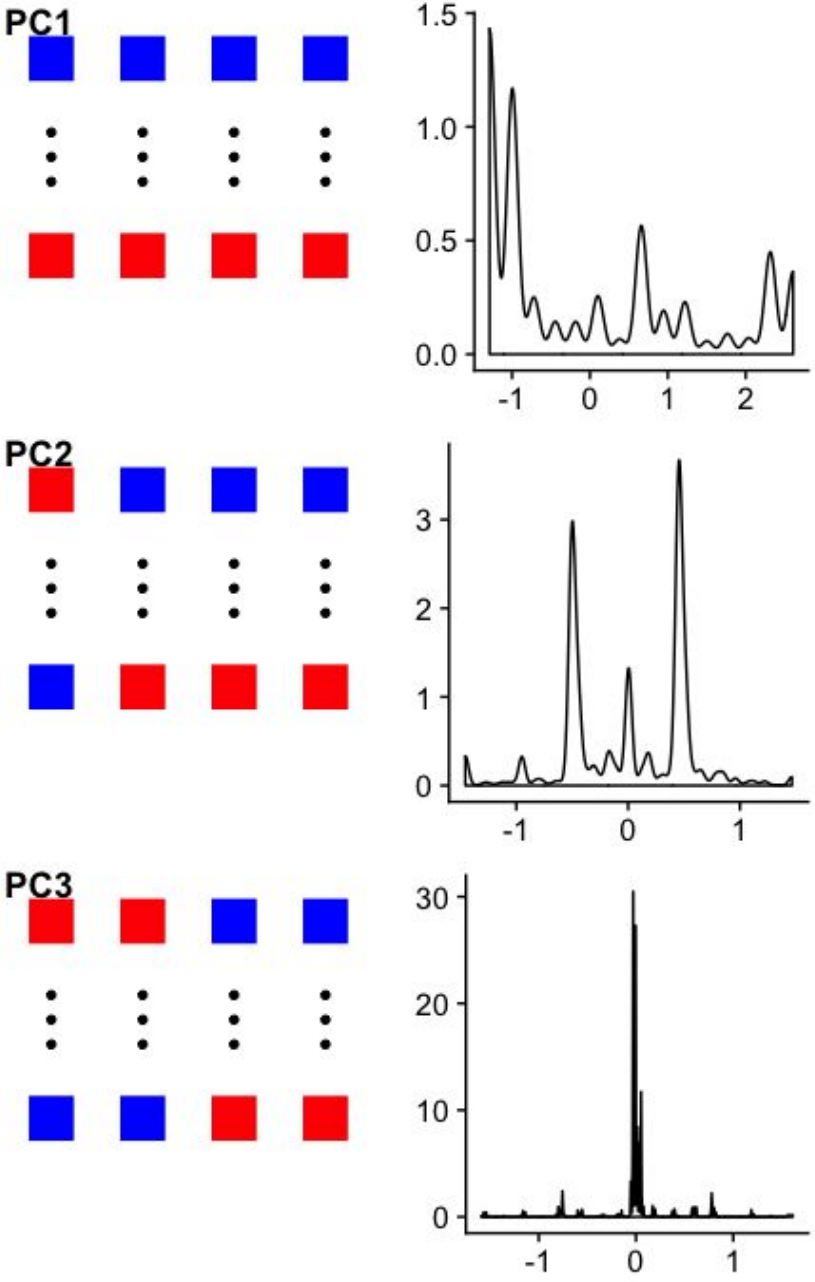

PC1
PC2
PC3
1.5
1.0
0.5
0.0
-1
0
1
2
3
30
20
10

TABLE 1. *PCA Loadings on Federal and State Races*

	Dependent variable:		
	PC1	PC2	PC3
	(1)	(2)	(3)
Percent Senior	0.256*	0.445***	−0.007
	(0.148)	(0.131)	(0.090)
Percent Female	1.175**	−1.576***	0.201
	(0.549)	(0.487)	(0.334)
Percent Non-White	3.568***	0.845***	−0.014
	(0.073)	(0.065)	(0.045)
Percent College Educated	0.096	−0.820***	0.031
	(0.205)	(0.182)	(0.124)
Median Income	−0.123*	−0.136**	−0.072*
	(0.071)	(0.063)	(0.043)
Constant	−0.627	2.321***	0.670
	(0.823)	(0.730)	(0.500)
Observations	2,022	2,022	2,022
R^2	0.690	0.228	0.004
Adjusted R^2	0.689	0.226	0.002
Residual Std. Error (df = 2016)	0.560	0.496	0.340
F Statistic (df = 5; 2016)	897.855***	118.987***	1.615
Note:			*p<0.1; **p<0.05; ***p<0.01

TABLE 2. *PCA Loadings on Local Races*

	Dependent variable:		
	PC1	PC2	PC3
	(1)	(2)	(3)
Percent Senior	0.147	−0.124	0.028
	(0.320)	(0.110)	(0.049)
Percent Female	−0.196	−0.194	0.051
	(1.212)	(0.418)	(0.187)
Percent Non-White	0.269	0.022	0.048*
	(0.164)	(0.057)	(0.025)
Percent College Educated	−0.079	−0.154	−0.079
	(0.453)	(0.156)	(0.070)
Median Income	0.228	0.012	0.082***
	(0.155)	(0.053)	(0.024)
Constant	−2.475	0.045	−0.923***
	(1.805)	(0.623)	(0.278)
Observations	1,729	1,729	1,729
R^2	0.003	0.003	0.011
Adjusted R^2	0.0002	0.0003	0.008
Residual Std. Error (df = 1723)	1.157	0.399	0.178
F Statistic (df = 5; 1723)	1.077	1.100	3.662***
Note:			*p<0.1; **p<0.05; ***p<0.01

Conclusion

- Largest source of variation comes from left-right ideology that aligns with partisanship. This is consistent in both Federal and Local races.

- In the context of South Carolina, Liberal-Conservative ideology at the federal level is largely explained by race, with gender also explaining some variance. However, these conclusions do not seem to hold at the local level.

Thank You for Listening!

We would <3 to hear any comments

jfb4@princeton.edu
syamaya@princeton.edu