Democracy vs Dictatorship or Something More?: Using Unsupervised Learning to Cluster Regimes

Simon Hoellerbauer

University *of* North Carolina *at* Chapel Hill

Research Question

- Can unsupervised learning help us find regime types?
- How accurately do current regime categorizations reflect underlying structure in the data?

Motivation

- Political scientists categorize regimes because they believe there are important descriptive and causal differences between them, otherwise there would be no point to the exercise. [1] There are, however, numerous—at times conflicting, at times overlapping—categorizations of regimes used in the political science literature
  - Often rely on subjective coding that is also very time-consuming
  - It is not clear which categorizations are more "important," in the sense that they reflect intrinsically different regime types—researchers can make subjective decisions about which "aspect" of a regime may be more important, although this may not be reflected in execution.
- There are also a variety of continuous measures used, such as Polity IV and V-Dem
  - Regardless of aggregation strategy, some of the heterogeneity of indicators can be obscured
  - In addition, states often cluster in clear bins
- Why not let data tell us how many groups there are?

Data and Methods

- Using 63 *Mid-Level* and *Other* Indices from V-Dem 8 [2] for 15015 post-1900 country-years
  - Results in 15015 × 63 feature matrix
- Use K-Means clustering and Gaussian Mixture Models (GMM) to find regime types
  - First, compare $k = 2$ clustering performed by the these two methods to compare against Cheibub et al. (2010)'s dichotomous measure of regime type [3]
  - Second, find the optimal K within the data
  - Third, analyze clusters produced
- All algorithms and models fit using the `scikit-learn` module in `Python`

2 Clusters: Democracy vs Dictatorship

Correlation Between Label Assignments

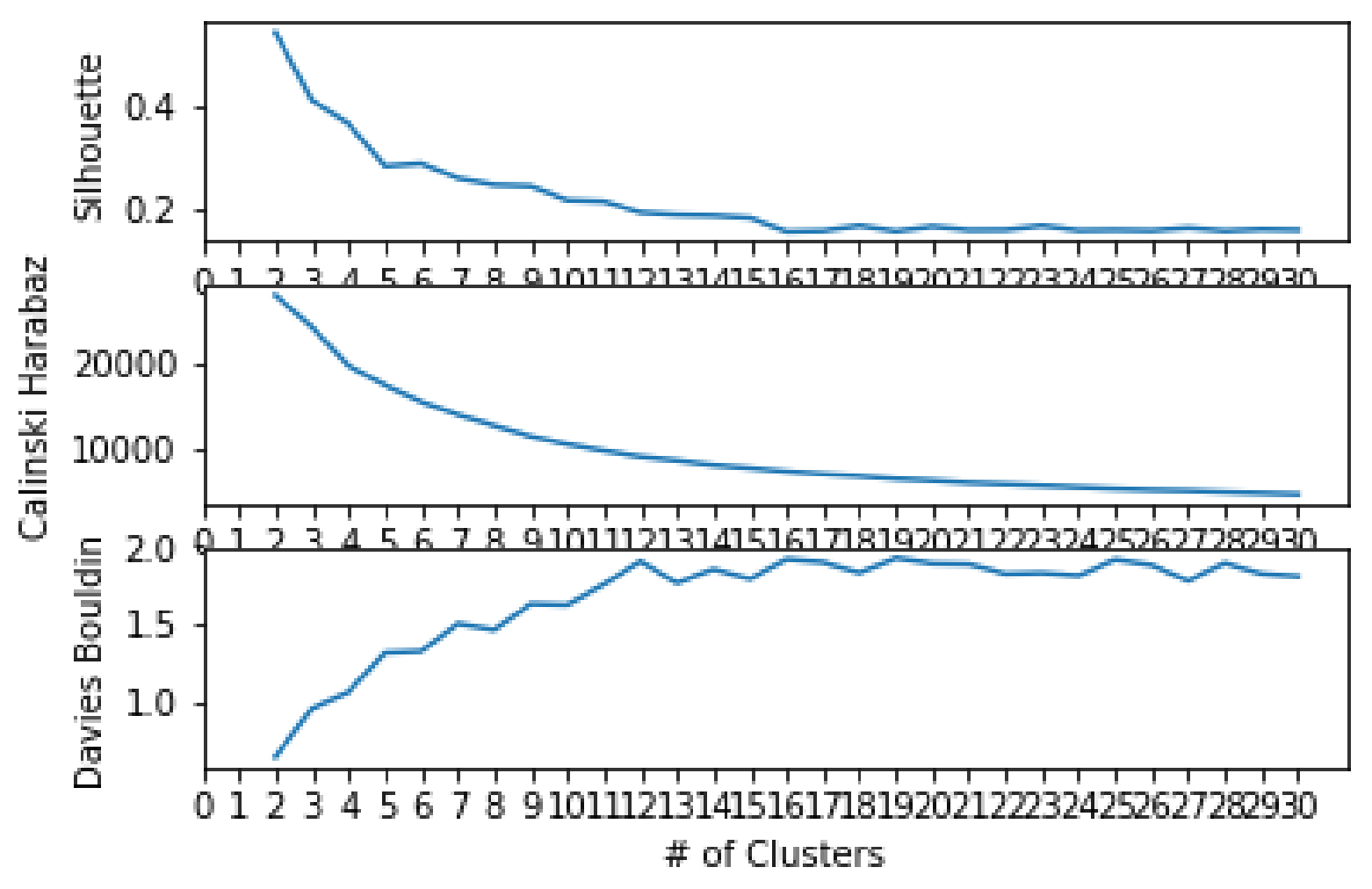
	Cheibub et al.	K-Means	GMM
Cheibub et al.	1.00		
K-Means	.78	1.00	
GMM	.37	.45	1.00

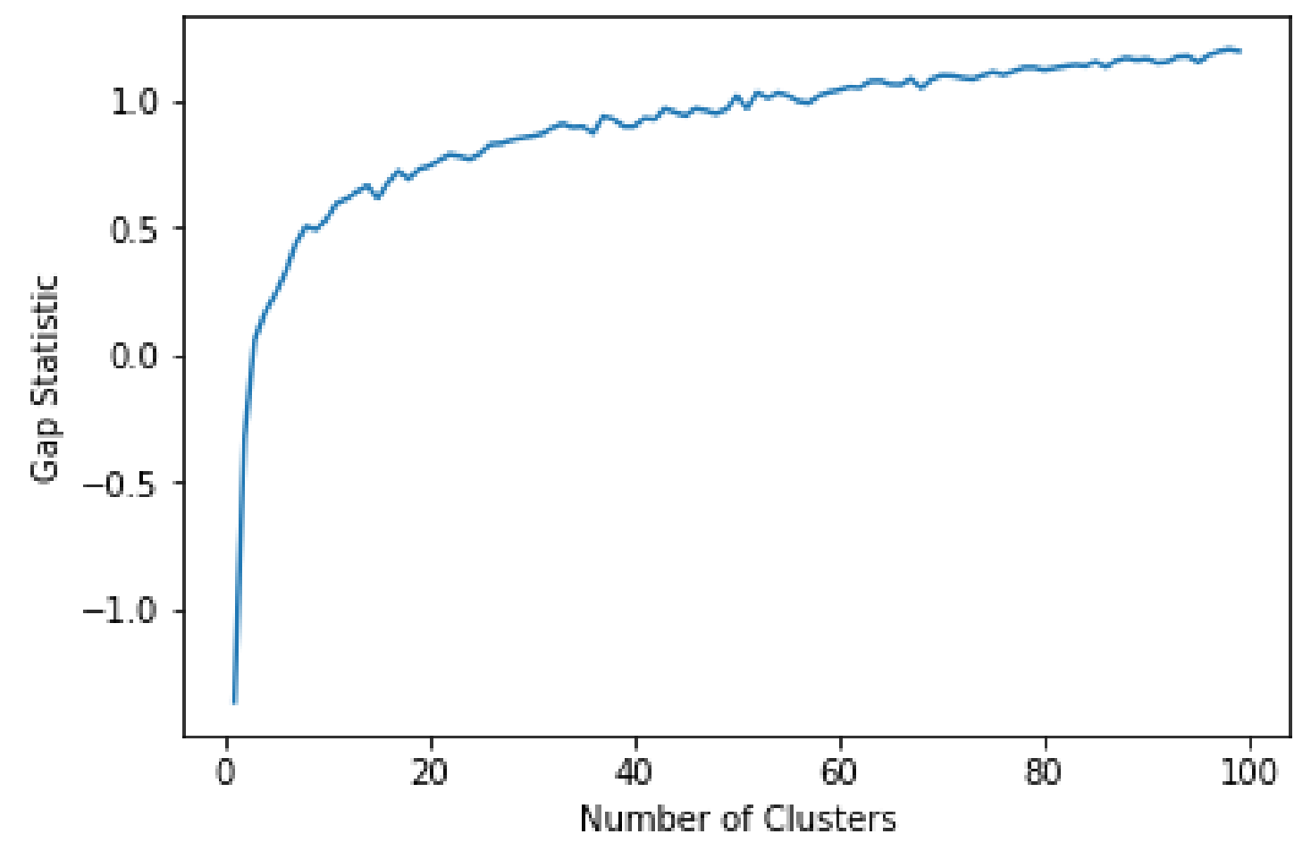
Sample of Disagreements

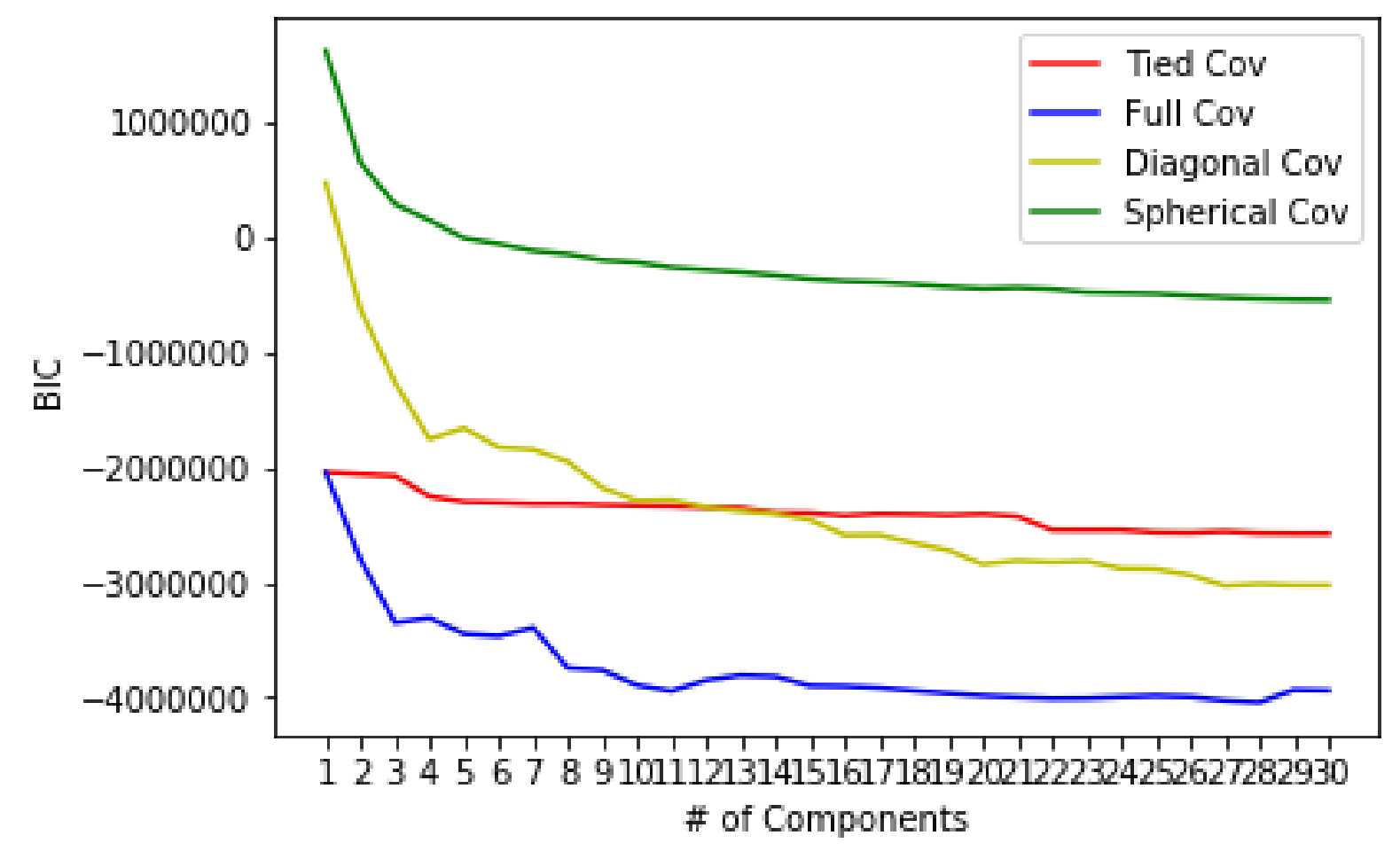
Country, Year	Cheibub et al.	K-Means	GMM
South Korea, 1981	0	0	1
Greece, 1950	1	0	1
Cyprus, 1981	0	1	1
South Africa, 2002	0	1	1
Nicaragua, 1975	0	0	1

Finding K

K-Means vs GMM







- Inconsistent results: K-Means scoring methods indicate 2 clusters; Gap Statistic indicates 98 (but this may grow with higher number of clusters); GMM indicates 11 clusters
- A Bayesian Guassian Mixture Model with Dirichlet Process prior was fit, but kept adding clusters until max number of clusters was hit (max used was 100; required by `scikit-learn`'s approximation method)

$K = 11$

Number of Country-Years within Clusters

Cluster	Country-Years
0	2636
1	1241
2	2185
3	693
4	3140
5	535
6	1360
7	631
8	560
9	1487
10	547

Mean Levels of Different Aspects of Democracy Within Clusters, Ranked from Lowest to Highest by Electoral Democracy

Cluster	Electoral	Liberal	Part.	Delib.	Egal.
Cluster_4	0.0636	0.0675	0.0414	0.0456	0.0662
Cluster_8	0.0836	0.0466	0.0441	0.0569	0.1229
Cluster_1	0.1504	0.0704	0.0598	0.0583	0.1494
Cluster_10	0.1616	0.1167	0.0959	0.1225	0.1207
Cluster_9	0.182	0.144	0.0899	0.1206	0.1234
Cluster_5	0.2529	0.1439	0.1378	0.1647	0.1679
Cluster_7	0.2704	0.1675	0.1249	0.1568	0.1205
Cluster_0	0.3154	0.1942	0.1636	0.2022	0.1923
Cluster_6	0.6472	0.5028	0.3951	0.508	0.4388
Cluster_3	0.6701	0.5106	0.4255	0.5128	0.4808
Cluster_2	0.8237	0.7501	0.5851	0.7231	0.7064

Differences Greater than .15 between Cluster 3 and Cluster 6

Index	Cluster_3	Cluster_6	Difference
Women Pol. Part.	0.8315	0.6495	0.1821
Executive Elec. Regime	0.906	0.4712	0.4348
Plebiscite	0.8676	0.0415	0.826

Discussion

- K-Means matches very well against Cheibub et al. (2010), GMM does not
- Both the K-Means result and the clear grouping of the means of the component distributions in the GMM with $K = 11$ unsurprisingly confirm that there are differences between democracies and dictatorships
- The results of $GMM_{K=11}$ seem to show that there is more variation within dictatorships than democracies, something that work of scholars like Linz[4] point out in their work.
- At the same time, the democracy scores at left show that there are some clusters for which the overall mean democracy scores are very similar but that should not be grouped together.

Next Steps

- Use discovered clusters as categorical predictor in place of established categorical measures and rerun analyses using new clusters
- Perform $K = k$ clustering, where k is the categories in different categorical measures of democracy, for comparative purposes
- Train GMM using existent categorical measures, which enables assessing accuracy, but assumes knowledge of categories, and then predict categories for all V-Dem country-years

Email: hoellers@unc.edu

References

[1] Robert A. Dahl.
Polyarchy: Participation and Opposition.
Yale University Press, New York, NY, 1971.

[2] Michael Coppedge, John Gerring, Carl Henrik Knutsen, et al.
V-Dem [Country-Year/Country-Date] Dataset v8.
Varieties of Democracy (V-Dem) Project.
`https://doi.org/10.23696/vdemcy18`, 2018.

[3] José Antonio Cheibub, Jennifer Gandhi, and James Raymond Vreeland.
Democracy and dictatorship revisited.
Public Choice, 143(1-2):67–101, 2010.

[4] Juan Linz.
Totalitarian and Authoritarian Regimes.