Progressives and Never-Trumpers: Contrastive Principal Component Analysis as an Alternative for Public Opinion Analysis

Sam Fuller*[1] and Tzu-Ping Liu†[2]

[1]University of California, Davis
[2]University of Taipei

April 20, 2022

Abstract

This paper exploits an emerging machine learning approach to scaling, contrastive principal component analysis (cPCA), to explore latent political subgroups among survey respondents from the 2018 Public Policy Institute of California's Statewide California Survey. While standard scaling methods, like PCA and factor analysis, often recover ideology as their first dimension when analyzing public opinion data, these methods almost necessarily gloss over important subgroup variation (e.g., moderate versus conservative Republicans). cPCA, on the other hand, is incredibly suited for subgroup analysis in that it identifies the dimensions on which one subset of the data (the target group) varies the most and the other subset of the data (the background group) varies the least. We find that we are able to distinguish between both the progressive versus moderate wings of the Democratic party and the pro- versus anti-Trump wings of the Republican Party, using opposing parties' respondents as the background groups, respectively. These results not only directly derive important subgroup variation but also highlight the general usefulness of cPCA for identifying both subgroups and latent divisions in public opinion research.

*Ph.D. Candidate, Department of Political Science. E-mail: sjfuller@ucdavis.edu.

†Assistant Professor, Department of Social and Public Affairs, University of Taipei. E-mail: tpliu@utaipei.edu.tw.

1 Introduction

Recent public discourse and academic research have both discussed and attempted to examine divides within the American political and ideological landscape. Specifically, there has been much discussion surrounding the progressive and moderate divide in the Democratic Party, see Bernie Sanders and "The Squad's" supporters versus Hillary Clinton and Joe Biden's (e.g., Forgey 2020; Clarke 2020), and the more moderate anti-Trump and more extreme pro-Trump divisions within the Republican Party, see Mitt Romney and Pat Toomey's supporters versus Ted Cruz and Lindsey Graham's (e.g., Bump 2020; Barber and Pope 2019; Rapoport, Reilly, and Stone 2019). This research has found mixed evidence of intraparty differences and diversity, with Rapoport, Reilly, and Stone (2019) finding that "never-Trumper" voters have over time began to support Trump more, whereas Barber and Pope (2019) and Clarke (2020) find significant intraparty differences in both the Democratic and Republican parties, at the voter and elite levels, respectively.

While two of these studies examine intraparty differences at the voter level, they utilize tailor-made surveys to explore potential internal variation. In fact, locating "subgrouped" respondents/voters in standard survey data has often been difficult, if not impossible, using standard unsupervised machine learning and scaling methods like principal component analysis (PCA). This lack of subgroup discrimination is driven by the fact that traditional methods only derive *general* patterns and do not provide direct analyses for latent patterns *within* groups instead of *across* groups.

These methods, while successful in identifying important ideological differences among entire samples, consistently fail to identify latent patterns that involve subgroups like never-Trumpers or progressives in both the Republican and Democratic parties' supporters, respectively. While these methods' results seem to suggest that these groups do not meaningfully exist, we instead argue that this is due to the primary goals of these methods and thus their limitations. Specifically, this arises from the fact that traditional methods either focus on identifying dimensions that explain the most variation (PCA) or minimizing classifica-

tion/prediction error (unfolding methods) in an *entire* sample of data. In both scenarios important subgroup variation and/or latent, full-sample differences are obscured and overlooked. Using our own analyses, we explicitly illustrate this phenomenon: Using traditional approaches often leads to interesting subgroup variation being overwhelmed by more common dimensions, such as left-right ideology or demographic differences.

To address these shortcomings, we exploit an emerging analysis approach, contrastive learning, to effectively isolate and identify these subgroups using contrastive PCA (cPCA) (Abid et al. 2018). Contrastive learning, in this scenario, operates by adjusting PCA's standard goal of identifying dimensions on which the data varies the most in the entire sample to identifying dimensions on which the *difference* between two (sub-)datasets is greatest. Specifically, cPCA instead identifies dimensions on which one subset of the data (the target group) varies the most and the other subset of the data (the background group) varies the least, hence the contrastive of contrastive PCA.

Although cPCA and, generally speaking, contrastive scaling require pre-defined groups for comparison, this does not mean that cPCA is a supervised learning methodology. In fact, these pre-defined groups *do not* divide subgroups: principal components (PCs) are still *unknown* and need to be estimated. One can analogize pre-defined groups as an extra constraint on the *definition* of principal directions—cPCA identifies the principal components/directions which vary the most in the pre-defined target group, but the least in the background group. In short, cPCA removes the dimensions of high variance shared across these two groups, often ideology, and instead focuses on what differentiates the target group from the background group. Importantly, this method is agnostic to what these dimensions are: the researcher does not select which variables constitute which dimensions, but only selects the target and background groups.

Applying cPCA to our data, we locate both the progressive versus moderate wings of the Democratic party, using Republicans as the background group, and the pro- versus anti-Trump wings of the Republican Party, using Democrats as the background group. This

analysis not only directly explores these important differences in American political parties but also highlights the general usefulness of cPCA for identifying both subgroups and latent divisions within entire populations. Furthermore, unsupervised methods like cPCA complement existing methodologies for studying public opinion in two major ways: first, it can identify dimensions that can be directly used in analyses as independent variables of interest; and second, cPCA can be used as an exploratory tool to identify hidden differences often overlooked by traditional methods. Overall, this paper's primary contribution is extending the use of cPCA to voter surveys specifically and public opinion more generally and providing a foundation from which future work can build off of.

2 Methodology

2.1 Contrastive Learning

cPCA operates by first splitting a dataset into two groups, the target and background groups, based on predefined classes, such as an individuals' party identification. The target group is the focus of the analysis, the group in which we are attempting to find subgroup variation, whereas the background group's primary use is as a contrast. Once these groups are defined, the method then differentiates the variance-covariance matrices of the target group and the background group to find PC(s) on which the target group varies maximally and the background group varies minimally. Given that cPCA is an extension of PCA, cPCA results can be interpreted as if they were from a PCA. Indeed, the recovered PCs can then be analyzed by looking at how variables included in the analysis load on to said PCs. We can also examine where individual datapoints (or respondents) are located in the recovered space. We can also examine these datapoints by calculating their distances between each other and by exploring which variables may separate individuals into various subgroups/clusters. Finally, the recovered dimensional positions of each datapoint can also be used in standard prediction models, like regressions, as independent variables.

Note that when applying contrastive learning, researchers can adjust a *contrast parame-*

ter, α, to control the trade-off between having high target variance versus low background variance. For instance, when $\alpha = 0$, the results are equivalent to applying standard PCA to only the target dataset. In other words, contrastive learning places greater emphasis on directions that reduce the variance of a background dataset as α increases. Therefore, instead of representing a parameter that determines the "best" solution, α rather represents a *value set* that researchers can manually explore to derive latent spaces that may reveal interesting patterns in the data. As proven by Abid et al. (2018), as long as $0 < \alpha < \infty$, "cPCA computes the subspaces of a list of α's and returns a few subspaces in terms of *principal angles* (Miao and Ben-Israel 1992)." In other words, through different α's or principal angles, cPCA can discover different subgroup structures within different subspaces and researchers can then decide which are related to their research question(s) of interest. Importantly, this aspect of cPCA rules out concerns regarding the possibility of recovering *false* or *inaccurate* dimensions.

Recently, scholars have developed approaches to aid researchers in reducing the time it takes to manually tune and explore the possible values of α. For instance, Fujiwara et al. (2020) and Fujiwara and Liu (2020) have developed auto-selection methods for finding the latent contrast space where a target dataset has the *greatest* variance relative to a background dataset.[1] For another instance, Boileau, Hejazi, and Dudoit (2020) propose to select the α in terms of *clustering*, i.e., selecting the α by maximizing the average silhouette width over clusterings of the reduced-dimension representation of the target data. This latter approach is used in our analyses.

3 PPIC Statewide California Survey 2018

The October 2018 PPIC Statewide Survey of Californians 2021 [cited 15 September 2021] surveyed voters on a wide-ranging set of policy questions and also recorded their personal demographic information. The survey was conducted between October 18–21, 2018 and re-

[1]Note that cPCA only identifies principal directions on which the target data has greater variance than the background data, and thus the greatest variance is just the criterion of this approach.

ceived responses from 1,700 individuals. Detailed coding and question information is located in Appendix B. Using respondents' stated party identification, we examine both Democrats and Republicans as target groups with the other being used as the background group.

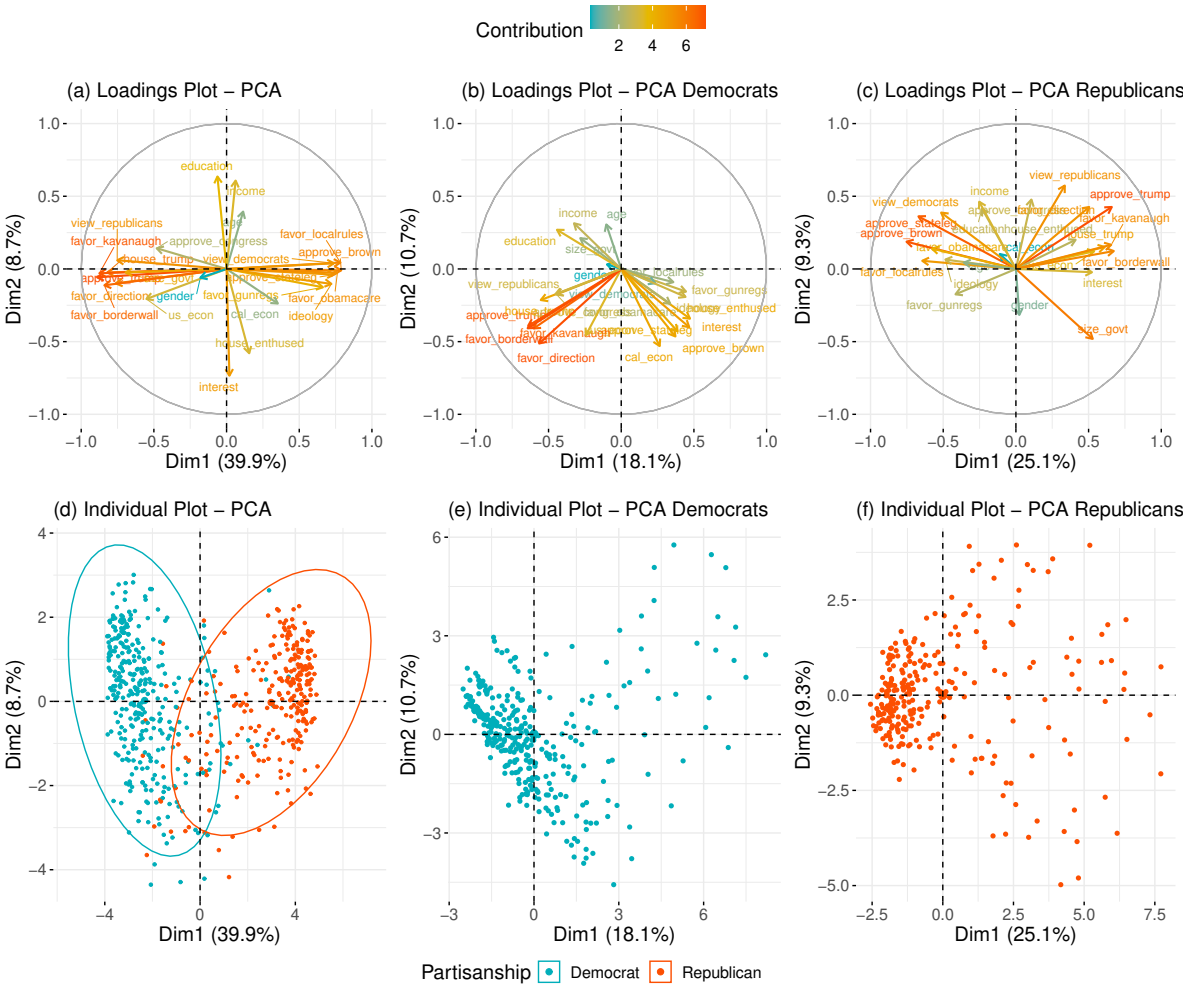
3.1 Data

The October 2018 PPIC Statewide Survey of Californians 2021 [cited 15 September 2021] surveyed voters on a wide-ranging set of policy questions and also recorded their personal demographic information. The survey was conducted between October 18–21, 2018 and received responses from 1,700 individuals. Detailed coding and question information is located in Appendix B. Using respondents' stated party identification, we examine both Democrats and Republicans as target groups with the other being used as the background group.

3.2 Standard PCA Results

First, we report the variable loadings and individual (respondent) positions of three PCA analyses in Figure 1. Percentages on each axis refer to the percent of variance explained by their respective dimension. The first analysis is conducted on only Democrats (Panels a and d), the second conducted on only Republicans (Panels b and e), and the third is conducted on the entire sample (Panels c and f).

Figure 1: PCA Analyses of Full Sample (Left), Displaying Only Democrats (Middle) and of Only Republicans (Right)



In Figure 1, Panel a shows the PCA's variable loadings and their contribution to each dimension, as measured by their direction and length/color, respectively. For instance, the variable `approve_trump` lies almost parallel with PC1 and is longest in length, thus it has the largest impact on PC1 and almost no impact on PC2. PC1 is comprised primarily of policy preferences, party support, and approval of various politicians and institutions. Demographic variables such as education, income, and political interest load instead on PC2.

Panel d shows the individual locations of respondents in this two-dimensional space. Based on Panel d, it is clear that respondents are distributed along the first and second PC

in the newly derived two-dimensional space. Once we color respondents based on their own party identification, the recovered positions demonstrate that respondents' policy preferences divide Democrats and Republicans along the first PC (PC1). Each partisan cluster can be identified using only PC1, with very little overlapping between the parties. Overall, based on both Panels a and d, the first principal component seems to very accurately measure standard left-right ideology, cleanly dividing Democrats and Republicans. Dimension 2, on the other hand, seems to mainly represent general demographic information and political interest.

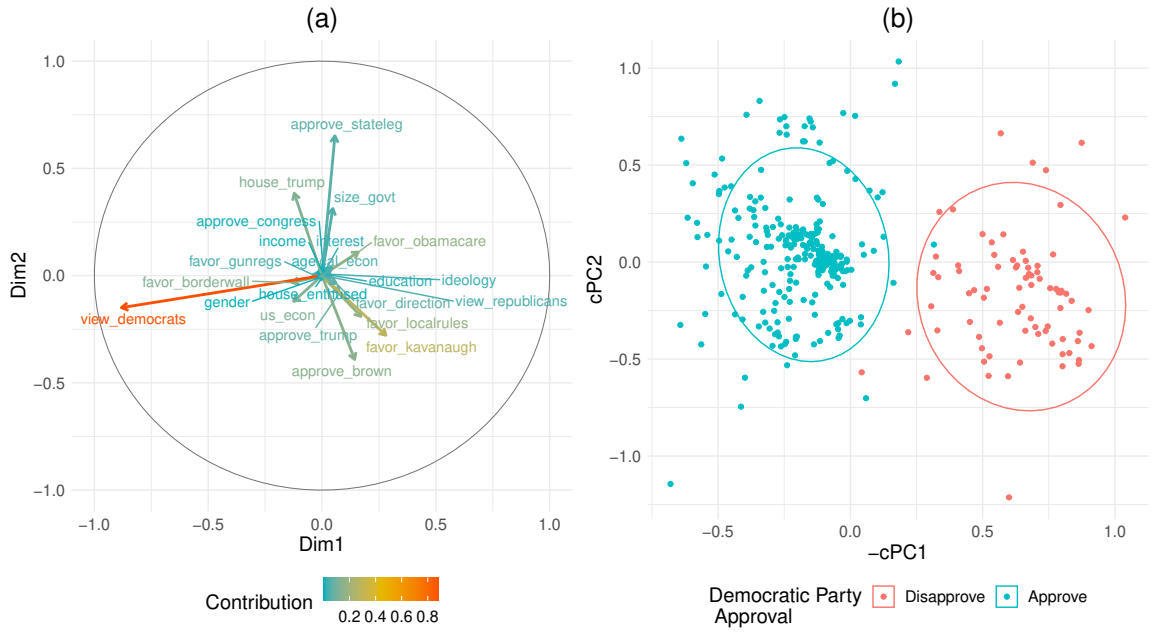
Additionally, we further split the original data by partisanship to apply PCA to the two new datasets separately. The dimensions in the Democratic (Panels b and e of Figure 1) and Republican (Panels c and f of Figure 1) analyses are significantly less clear than those in the combined analysis. Variables are not clearly loading onto a single dimension as before, but it still seems that divisions are mainly ideological along the X axis for Republicans, with similar results for Democrats if rotated (as in factor analysis). Clearly, while these dimensions are useful for standard ideological analyses (at least with the results from the full sample PCA), they do not reveal any particularly interesting subgroup variation. This lack of interesting findings provides us ample motivation to explore the data using cPCA.

3.3 cPCA Results: Democrats

To illustrate the results of the cPCA, we first use Democrats as the target group and Republicans as the background group. Looking at the variable loadings presented in Panel a of Figure 2 we see that unlike the *general pattern* derived by PCA in Figure 1, cPCA clearly finds that the variable `view_democrats` (do you have a (un)favorable view of the Democratic Party) is *the only* variable which loads significantly onto PC1.[2] This loadings plot indicates that Democrats, when contrasted with Republicans, are distributed based on their support of the Democratic Party.

[2]For the sake of simplicity, we will focus our discussion on PC1. Nevertheless, it is important to note that the composition of PC2 is also different than the PC2 derived from PCA.

Figure 2: cPCA Results (Target: Democrats, Background: Republicans)

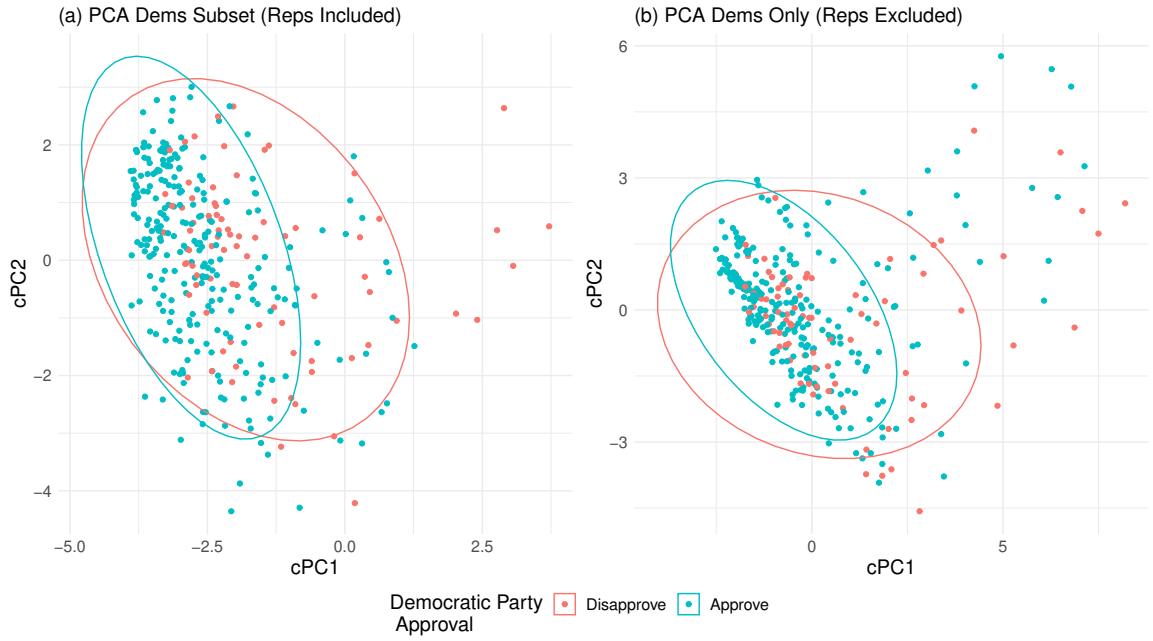


To illustrate the importance of `view_democrats`, we color and group Democrats in the latent space based on their responses. As Panel b of Figure 2 demonstrates, PC1 from the cPCA analysis divides Democrats into two distinct subgroups: those who approve of the Democratic Party and those who do not. Descriptively, we argue that this cPCA analysis identifies the progressive/anti-establishment wing of the Democratic party. Importantly, this difference is wholly obscured in both the full-sample and Democrat-only PCA analyses, presented in Panels a and b of Figure 3.

3.4 cPCA Results: Republicans

We now move on to the results of the cPCA analysis using Republicans as the target group and Democrats as the background group. Similar to the analysis of Democrats, the variable loadings in Panel a of Figure 4 show us that, unlike the PCA results, cPCA identifies `favor_borderwall` (do you favor building a border wall between the U.S. and Mexico) as the variable with the greatest loading onto PC1. The group-colored results of the Republicans are further presented in Panel b in Figure 4. Overall, these results indicate that the

Figure 3: PCA: Individual Coordinate Plots of Democrats



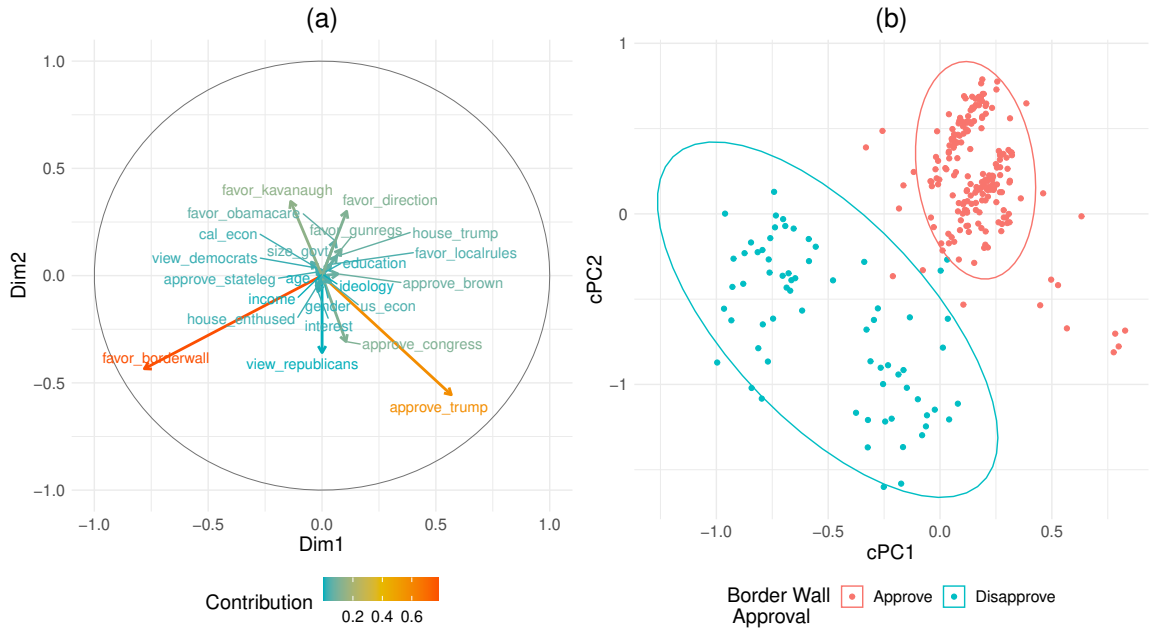
most controversial issue that divides Republicans is highly related to Trump, and separates Republicans into two distinct pro and anti US border-wall camps.

Again, this result highlights how cPCA enables the discovery of interesting differences that exist within one group, but not within another. In other words, this policy divides Republicans significantly, while not mattering almost at all for Democrats. As a validity check, we color and group Republicans by their responses to `view_borderwall` and report the PCA locations from both the full-sample and a Republican-only PCA analysis. As Figure 5 demonstrates, the internal Republican division over building a border wall is concealed in the results from the original PCA analysis. As illustrated in both Panels a and b, the distribution of both the pro- and anti-wall camps overlap almost completely in both latent PCA spaces.

3.5 OIRT & Blackbox Comparisons

To verify that our recovered differences are not merely an artifact of comparing only PCA and cPCA, we use other widely utilized scaling methods with the same dataset as a validity check: blackbox scaling (Poole 1998) and ordinal item response theory (OIRT) model (Quinn 2004).

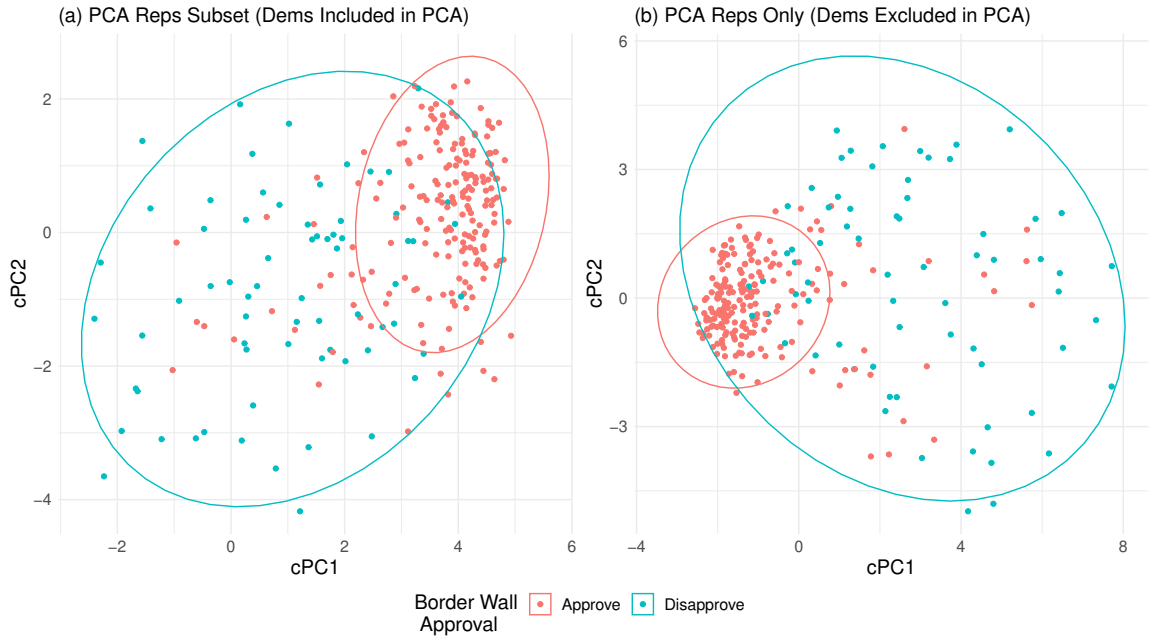
Figure 4: cPCA Results (Target: Republicans, Background: Democrats)



We apply these methods to three different settings and present the results in Appendix A: the full sample (deriving both partisans' positions simultaneously, as in Figure 1); Democrats' or Republicans' subset positions while *including* the other partisans in the analysis (as in Figure 3); and Democrats' or Republicans' positions only, *excluding* the other partisans in the analysis (as in Figure 5). Although the derived results across these three methods are nonidentical, they recover similar patterns. First, every method recovers a space where respondents' ideal points are clearly separated and grouped by their partisanship. Second, every recovered space fails to demonstrate any meaningful or interesting patterns *within* each party. Given that each of these recovered spaces demonstrate similar patterns to our initial PCA results, we argue that our findings are not PCA dependent. Instead, we contend that they are contrastive dependent; that is, our results highlight the distinct differences in recovered spaces between contrastive and standard scaling methods.[3]

[3]Given that OIRT does not derive issue vectors akin to PCA and Blackbox scaling, we align the x-axis with the vectors which posses the highest contribution to the composition of the x-axis in Figure 1, Figure 3, and Figure 5 separately and align the y-axis with education which has the largest impact on the composition of the y-axis in Figure 1.

Figure 5: PCA: Individual Coordinate Plots of Republicans



4 Discussion: cPCA for Subgroup Analysis in Public Opinion Data

Dimensional analysis techniques, like PCA, have been used widely in public opinion research to identify latent dimensions like ideology. Above and beyond identification, researchers can also use these methods to reduce their analyzed dimensions so as to address the curse of dimensionality and increase their model's parsimony. Nevertheless, these methods necessarily focus on the variance in the data across the *entire sample* analyzed. Specifically, these scaling methods operate by identifying dimensions that capture the highest variance in the data. Consequently, they often find dimensions, like ideology, that do capture distinct differences between and across groups within the sample, but that also obscure dimensions that may represent interesting or important subgroup variation. This often occurs, as shown in our analyses, even when a researcher subsets their data to specific groups (like Democratic or Republican respondents). Simply put, this means that these methods are relatively ineffective at analyzing subgroups in public opinion data.

Contrastive learning represents an alternative approach to scaling analyses that addresses many of these issues. As we have shown in our analyses, contrastive PCA is able to recover important subgroup variation and division among individuals within both major political parties. Critically, these dimensions are not recovered from full-sample nor subset analyses using either PCA, blackbox scaling, or OIRT. However, we stress that cPCA and other contrastive learning methods are not meant as replacements to standard dimensional analysis methodologies, but rather as complements geared specifically toward subgroup analyses. Put simply, cPCA is not superior to PCA, or other methods, rather it is an additional tool researchers can use that provides distinct benefits and advantages when performing subgroup analysis.

More precisely, cPCA can help identify how groups are meaningfully different from one another in the distribution of their data. For example, in our analyses of Democrats and Republicans in Figure 2 and Figure 4, respectively, we find that Democrats are distinctly divided by their (dis)approval of the Democratic Party while Republicans are just as divided by their (dis)approval of a border wall with Mexico. Again, these dimensions are wholly obscured in all our analyses using standard methods.

In general, cPCA also benefits from the fact that it is a highly data-driven, mathematical approach modified primarily by only one, contrast parameter (α). This means that few assumptions are necessary for analysis and, in contrast to other subgroup methods, requires no prior information about subgroups (like what variables might be most divisive within a subgroup). Instead, the researcher need only determine which groups should be analyzed as target and background groups.[4] Overall, cPCA, specifically, and contrastive learning, generally, constitute a set of methods with a high degree of potential for subgroup analysis in public opinion data.

[4]While not addressed directly in this paper, this is in contrast to other methods for subgroup analysis like class specific multiple correspondence analysis (CSA) and subgroup multiple correspondence analysis (sMCA). These other methods require much more information about what variables may cause divisions in subgroups, whereas cPCA does not.

References

Abid, Abubakar, Martin J Zhang, Vivek K Bagaria, and James Zou. 2018. "Exploring Patterns Enriched in a Dataset with Contrastive Principal Component Analysis." *Nature Communications* 9 (1): 1–7.

Barber, Michael, and Jeremy C Pope. 2019. "Conservatism in the Era of Trump." *Perspectives on Politics* 17 (3): 719–736.

Boileau, Philippe, Nima S. Hejazi, and Sandrine Dudoit. 2020. "Exploring High-Dimensional Biological Data with Sparse Contrastive Principal Component Analysis." *Bioinformatics* 36 (11): 3422–3430.

Bump, Philip. 2020. *No Senator Ever Voted to Remove A President of His Party from Office. Until Mitt Romney.,* February. https://www.washingtonpost.com/politics/2020/02/05/no-senator-ever-voted-remove-president-his-party-office-until-mitt-romney/.

Clarke, Andrew J. 2020. "Party Sub-Brands and American Party Factions." *American Journal of Political Science* 64 (3): 452–470.

Forgey, Quint. 2020. *AOC: 'In Any Other Country, Joe Biden and I Would Not Be In the Same Party',* January. https://www.politico.com/news/2020/01/06/alexandria-ocasio-cortez-joe-biden-not-same-party-094642.

Fujiwara, Takanori, and Tzu-Ping Liu. 2020. "Contrastive Multiple Correspondence Analysis (cMCA): Using Contrastive Learning to Identify Latent Subgroups in Political Parties." *arXiv preprint arXiv:2007.04540.*

Fujiwara, Takanori, Jian Zhao, Francine Chen, Yaoliang Yu, and Kwan-Liu Ma. 2020. "Interpretable Contrastive Learning for Networks." *arXiv preprint arXiv:2005.12419.*

Miao, Jianming, and Adi Ben-Israel. 1992. "On Principal Angles Between Subspaces in Rn." *Linear Algebra and Its Applications* 171:81–98.

Poole, Keith. 1998. "Recovering a Basic Space from a Set of Issue Scales." *American Journal of Political Science* 42 (3): 954–993.

Public Policy Institute of California. Statewide Survey October 2018. 2021 [cited 15 September 2021]. Database: PPIC [Internet]. Available from: https://www.ppic.org/dataset/ppic-statewide-survey-data-2018/.

Quinn, Kevin. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12 (4): 338–353.

Rapoport, Ronald B., Jack Reilly, and Walter J. Stone. 2019. "It's Trump's Party and I'll Cry if I Want To." *The Forum* 17 (4): 693–709. https://doi.org/doi:10.1515/for-2019-0041. https://doi.org/10.1515/for-2019-0041.