CASE 1: Contrastive Multiple Correspondence Analysis (cMCA): Applying the Contrastive Learning Method to Identify Political Subgroups

Team Members

1. Mario Palomino Viero
2. Sandaru Welikala
3. Rachel Mendosa
4. Komalpreet

Content

Objective 3
Literature Review 4
Explanatory Data Analysis 6
CCES15_Common – Raw dataset 6
Issuevalue_short – Dataset of this study 8
Advanced Techniques for Analyzing and Classifying Complex Categorical and Mixed-Type Data 12
Ordered Optimal Classification (OOC) 13
Multiple Correspondence Analysis (MCA) 15
Basic Space (BS) 17
Bayesian Mixed Factor Analysis (BMFA) 19
Contrastive Multiple Correspondence Analysis (cMCA) 21
Data Imputation Process 23
Implementation of code for cMCA 25
Result of cMCA 27
Findings of cMCA Study 33
Discussion 34
Bibliography 35

Objective

The objective of this study is to replicate the analysis performed in the original study on Contrastive Multiple Correspondence Analysis (cMCA) by exclusively applying a set of dimensional reduction and classification techniques namely Ordered Optimal Classification (OOC), Multiple Correspondence Analysis (MCA), Basic Space (BS), Bayesian Mixed Factor Analysis (BMFA), and Contrastive Multiple Correspondence Analysis (cMCA) specifically to the Cooperative Congressional Election Study (CCES) 2015 dataset. By employing these methodologies, this replication effort aims to identify and delineate political subgroups within the dataset, focusing particularly on uncovering patterns of ideological diversity and polarization among American voters. The study seeks not only to validate the efficacy of cMCA and related methods in political subgroup identification but also to explore the depth of ideological distinctions present in the CCES 2015 data, contributing to the broader understanding of voter behavior and political landscape polarization.

Literature Review

Ideal point estimation and dimensionality reduction play a critical role in political science by simplifying and clustering complex, high-dimensional political data for analysis and visualization. Traditional methods such as Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) often uncover the latent left-right ideological spectrum. However, these methods frequently face challenges when the left-right scale is uninformative, particularly among highly moderate respondents. Scholars have argued that these methods struggle to differentiate subgroups when data clusters centrally or when some variables dominate the source of variation (Coombs 1964; Armstrong et al. 2014; Fujiwara, Kwon, and Ma 2020).

To address these limitations, contrastive learning approaches, like contrastive PCA (cPCA) developed by Abid et al. (2018), have emerged. Contrastive learning is effective in identifying hidden patterns by comparing predefined groups and deriving PCs on which the target group shows the greatest variation relative to the background group. This approach can reveal biases that traditional methods overlook, often providing insights into binary, ordinal, and nominal variables common in social science (Harris and Harris 2010).

The contrast between traditional MCA and contrastive MCA (cMCA) is significant. cMCA improves dimensional analysis by contrasting subsets of data to highlight specific patterns that differ from one dataset to another, thus offering an enhanced ability to delineate meaningful subgroups and influencing factors (Greenacre 2017; Abid et al. 2018).

Applications of cMCA have demonstrated its utility across various cultural and political contexts. For instance, analyses of datasets such as the 2015 Cooperative Congressional Election Study (CCES), 2012 UTokyo-Asahi Elite Survey (UTAS), and 2018 European Social Survey (ESS) illustrate the method's capability to detect underlying political cleavages. For example, in the CCES, cMCA identified ideological polarization within American voters beyond what traditional methods revealed (Fiorina and Abrams 2008; Iyengar and Westwood 2015). Similar applications in Japan’s UTAS and the UK's ESS reflected cMCA’s strength in identifying intra-group divisions even amid low polarization, highlighting the method’s flexibility and power in revealing latent ideological clusters (Endo and Jou 2014; Adams, Green, and Milazzo 2012).

Overall, this paper suggests that cMCA offers an advanced methodological framework capable of enhancing existing political data analysis techniques by capturing dimensions and subgroup patterns overlooked by traditional PC methods. This makes it an invaluable tool in contemporary political science research, fostering a deeper understanding of complex social and cultural dynamics. (Fujiwara, Takonari and Liu 2020)

Explanatory Data Analysis

CCES15_Common – Raw dataset

The CCES15_Common_OUTPUT_Jan2016 dataset is part of the Cooperative Congressional Election Study (CCES), conducted in 2015. It is a large, nationally representative survey of U.S. voters that collects data on political attitudes, behaviors, and demographic characteristics.

Key features of the dataset include:

- Demographics: Information on respondents' age, gender, race, education, and income.
- Political Behavior: Voting patterns, party identification, and political ideology.
- Policy Opinions: Views on various policy issues such as healthcare, taxation, and climate change.
- Geography: Location data at different levels (state, district, etc.).
- Survey Design Variables: Weights and other variables to adjust for sample representation.

The raw dataset comprises 253 features (survey questions) and 14,248 responses, containing both numerical and categorical data.

To achieve the cMCA objective of this study, we needed to preprocess the dataset to reduce its dimensions and transform the data into numerical format. Therefore, we decided to follow the same mechanism as employed by the original authors of the study. This involved assigning meaningful feature names to the columns and transforming the data into the suggested numerical ranges, as outlined in their R code.

The transformation process included recoding categorical variables into numerical values, standardizing the scales of continuous variables, and grouping certain variables based on predefined ranges (as specified in the authors' methodology). Additionally, variables that were originally binary were assigned 0 or 1 values, and factors with multiple levels were mapped to corresponding numerical values based on their significance.

Example of Recoding:

The variable partyid had values indicating party affiliation (1 to 7). Recoding transformed this into two groups:

- Values 1, 2, and 3 were recoded to 1 (Democratic).
- Values 4, 5, and 6 were recoded to 2 (Republican).
- Any other values were set to NA:

During this process, we discovered that the raw dataset lacked 12 important features containing information on military-diplomacy and unequal rights factors, which had the highest loadings in differentiating the two main political parties (Democrats and Republicans). (Fujiwara, Takonari and Liu 2020)

As a result, we opted to use issuevalue_short, a transformed subset of the raw data, for our cMCA analysis.

Issuevalue_short – Dataset of this study

The issuevalue dataset contains 1000 records and 53 columns. It is a cleaned and processed version of survey responses from the CCES15_Common_OUTPUT_Jan2016.dta raw data file. Each column represents a survey question, demographic variable, or derived attribute related to political preferences, ideology, socio-economic status, and religiosity. Transformations primarily ensure consistency, handle out-of-range values, and recode variables into meaningful categories, enabling structured analysis and interpretation.

Column Name	Description	Data Type	Missing Values
ideology	Respondents' self-reported political viewpoint (e.g., Very liberal, Liberal, Moderate, Conservative)	Numerical	64
socialideology	Describe self political viewpoint	Numerical	101
economicideology	Viewpoint on National Economic condition	Numerical	142
natsecurityideology	Knowledge of which party has majority seats in state senate	Numerical	115
repealACA	Support or opposition to repealing the Affordable Care Act (Obamacare)	Numerical	14
approvekeystone	Support or opposition to the Keystone XL pipeline	Numerical	53
iransanctions	Support or opposition to sanctions on Iran	Numerical	22
transpacificpartnershipact	Support or opposition to the Trans-Pacific Partnership trade agreement	Numerical	44
normalizecubarelations	Support or opposition to normalizing relations with Cuba	Numerical	25

renewpatriotact	Support or opposition to renewing the USA Patriot Act	Numerical	523
usafreedomact	Support or opposition to the USA Freedom Act	Numerical	508
tradeadjustmentact	Support or opposition to the Trade Adjustment Assistance program	Numerical	23
VAWA	Support or opposition to reauthorizing the Violence Against Women Act	Numerical	13
cutmedicaremedicaid	Support or opposition to cutting funding for Medicare/Medicaid	Numerical	24
budgetcutsacrossboard	Support or opposition to across-the-board budget cuts	Numerical	29
middleclasstaxcut	Support or opposition to tax cuts for the middle class	Numerical	31
taxhikepreventionact	Support or opposition to the Tax Hike Prevention Act	Numerical	38
balancebudget.firstchoice.cutdefensespending	First-choice preference for balancing the budget: cut defense spending	Numerical	20
balancebudget.firstchoice.cutdomesticspending	First-choice preference for balancing the budget: cut domestic spending	Numerical	20
balancebudget.firstchoice.raisetaxes	First-choice preference for balancing the budget: raise taxes	Numerical	20
environment_regulateCO2	Support or opposition to regulating carbon dioxide emissions	Numerical	14
environment_raisefuelefficiency	Support or opposition to raising fuel efficiency standards	Numerical	17
environment_requirerenewablefuels	Support or opposition to requiring the use of renewable fuels	Numerical	17
environment_strengthenenforcementcleanair	Support or opposition to strengthening enforcement of the Clean Air Act	Numerical	16

guncontrol_backchecks	Support or opposition to background checks for gun purchases	Numerical	9
guncontrol_prohibitpublishingnames	Support or opposition to prohibiting publishing gun owners' names	Numerical	16
guncontrol_banassaultrifles	Support or opposition to banning assault rifles	Numerical	12
guncontrol_easierconcealcarry	Support or opposition to making it easier to obtain concealed carry permits	Numerical	11
immigration_grantlegalstatus	Support or opposition to granting legal status to undocumented immigrants	Numerical	0
immigration_increasepatrols	Support or opposition to increasing border patrols	Numerical	0
immigration_allowpolicequestion	Support or opposition to allowing police to question individuals about immigration status	Numerical	0
immigration_finebusinesseshire	Support or opposition to fining businesses that hire undocumented immigrants	Numerical	0
immigration_deportillegalimmigrants	Support or opposition to deporting undocumented immigrants	Numerical	0
alwaysallowabortion	Support or opposition to always allowing abortion	Numerical	15
prohibitabortiontwentyweeks	Support or opposition to prohibiting abortions after 20 weeks	Numerical	14
allowemployersdeclineabortion	Support or opposition to allowing employers to decline abortion coverage in health plans	Numerical	12
prohibitfederalfundingabortions	Support or opposition to prohibiting federal funding for abortions	Numerical	18
gaymarriage	Support or opposition to legalizing gay marriage	Numerical	15

affirmativeaction	Support or opposition to affirmative action policies	Numerical	20
syria_significantforce	Support or opposition to using significant military force in Syria	Numerical	0
moraltrad.adjust	Support or opposition to adjusting traditional moral values	Numerical	31
moraltrad.lifestyle	Views on the importance of traditional lifestyles	Numerical	60
moraltrad.tolerant	Support or opposition to a more tolerant society	Numerical	38
moraltrad.familyvalues	Support or opposition to policies promoting family values	Numerical	58
egalitarianism.equal	Support for policies promoting equality	Numerical	32
egalitarianism.toofar	Concern that egalitarian policies go too far	Numerical	42
egalitarianism.bigprob	Belief that inequality is a big problem	Numerical	47
egalitarianism.worryless	Belief that inequality is less of a concern	Numerical	64
egalitarianism.notbigprob	Belief that inequality is not a big problem	Numerical	58
egalitarianism.fewerprobs	Belief that equality-related problems have decreased	Numerical	51
militarism.strength	Support for prioritizing military strength	Numerical	87
militarism.diplomacy	Support for prioritizing diplomacy over military strength	Numerical	90
partyid	Self-identified party affiliation (e.g., Democrat, Republican, Independent)	Numerical	38

Table 1: Issuevalue_short dataset description

Advanced Techniques for Analyzing and Classifying Complex Categorical and Mixed-Type Data

The survey examined in this study is the 2015 Cooperative Congressional Election Study (Ansolabehere and Schaner 2017). Prior to applying the correspondence analysis (cMCA) objective, we utilized techniques such as Ordered Optimal Classification (OOC), Multiple Correspondence Analysis (MCA), Basic-Space Procedure (BS), and Bayesian Mixed Factor Analysis (BMFA) on the issuevalue_short dataset. These methods helped in data preparation and dimensionality reduction, providing important insights that enhanced the quality and interpretability of the final cMCA analysis.

This study also explores how these techniques, particularly OOC, MCA, BS, and BMFA, can be applied to handle and analyze complex datasets, especially those involving categorical, ordinal, and mixed-type variables. The methods are relevant to fields such as social science research, survey data analysis, psychometrics, and market research, offering valuable insights into revealing patterns, relationships, and latent factors within data.

Ordered Optimal Classification (OOC)

Ordered Optimal Classification (OOC) is a technique used to map ideological structures and polarization in datasets, particularly those with ordered categorical data. In political science, it helps analyze ideological positioning based on survey responses related to political attitudes or preferences.

OOC uncovers underlying dimensions of beliefs that traditional methods like factor analysis may miss. It is especially useful for survey data with ordered categorical responses, such as Likert-scale items (e.g., strongly agree to strongly disagree). By revealing how beliefs are structured and aligned, OOC identifies cleavages and polarization within a population.

The method optimizes the assignment of respondents to categories based on their ordered responses, aiming to find the most meaningful classification that reflects ideological dimensions. It maximizes internal consistency while preserving response order. Unlike traditional classification techniques, OOC does not assume predefined clusters but searches for informative divisions within the data, making it valuable for analyzing complex ideological spectra like left-right or liberal-conservative continuums. (Hare, Liu, and Lupton 2018)

To implement Ordered Optimal Classification (OOC) in Python, we followed these steps:

1. **Handling Missing Data**: We used Support Vector Machine (SVM) regression to impute missing values. For each feature with missing data, we trained an SVM model based on the available (non-missing) data and used it to predict and fill in the missing values.

2. **Standardization**: The imputed dataset was standardized using the Standard Scaler to ensure all features had zero mean and unit variance. This step is crucial for preventing features with larger scales from dominating the analysis.

3. **Dimension Reduction**: We applied Principal Component Analysis (PCA) to reduce the dataset from its original dimensions to two principal components. This allows for visualizing the data in a two-dimensional space while retaining as much variance as possible from the original dataset.

4. **Visualization**: The results were plotted using a scatter plot, where the data points were color-coded based on political party affiliation (Democrats vs. Republicans). This visualization helped to reveal the ideological structure and any possible polarization in the dataset.

By applying these steps, we ensure that the data is preprocessed, reduced to a manageable form, and visualized in a way that highlights the underlying ideological structure of the dataset, as shown in Fig. 1. The result of OOC visualization reveals a clear ideological separation between the two groups, with Democrats clustering on the left side and Republicans on the right, indicating potential polarization along these dimensions.

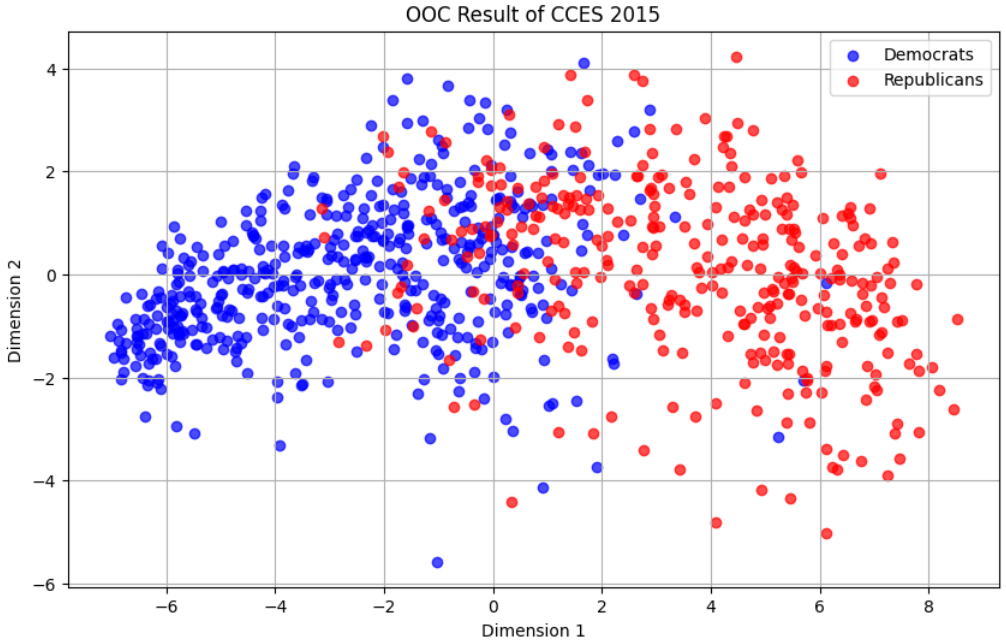


Fig. 1: OOC

Multiple Correspondence Analysis (MCA)

Multiple Correspondence Analysis (MCA) is a dimensionality reduction technique for categorical data, extending Correspondence Analysis (CA) to more than two variables. It uncovers relationships and patterns in large datasets by reducing dimensions while preserving significant associations. MCA is particularly useful for survey data, market research, or any dataset with multiple categorical variables.

MCA simplifies complex datasets by identifying underlying structures among categorical variables. It highlights key factors driving relationships and provides a reduced-dimensional visualization, making interpretation easier. Researchers can use MCA to group similar responses or observations, revealing trends or clusters.

The methodology involves creating a contingency table of categorical variables, calculating correspondences, and applying Singular Value Decomposition (SVD) to reduce dimensionality. The resulting coordinates for each category or observation can be plotted to explore relationships or identify patterns. (Greenacre 2007).

To implement Multiple Correspondence Analysis (MCA) in Python, we followed these steps:

1. **Data Preprocessing**: We loaded the dataset and split it into two groups based on party affiliation (Democrats and Republicans). We then handled missing values by filling categorical columns with 'na' and numerical columns with 99.

2. **MCA Execution**: We applied the MCA algorithm to the preprocessed data, excluding the party affiliation column, and reduced the dimensionality to two components for visualization.

3. **Visualization**: The MCA results were visualized using a scatter plot, where each data point was color-coded based on party affiliation. This allowed us to observe the ideological structure and potential polarization between the two groups.

These steps ensure that the data is cleaned, reduced to a lower-dimensional form, and visualized, revealing the underlying ideological patterns in the dataset, as shown in Fig. 2. The MCA visualization illustrates the distribution of different party affiliations across the reduced dimensions, highlighting ideological patterns and potential polarization. The result shows some overlap and distinct clusters, similar to the findings in Fig. 1.

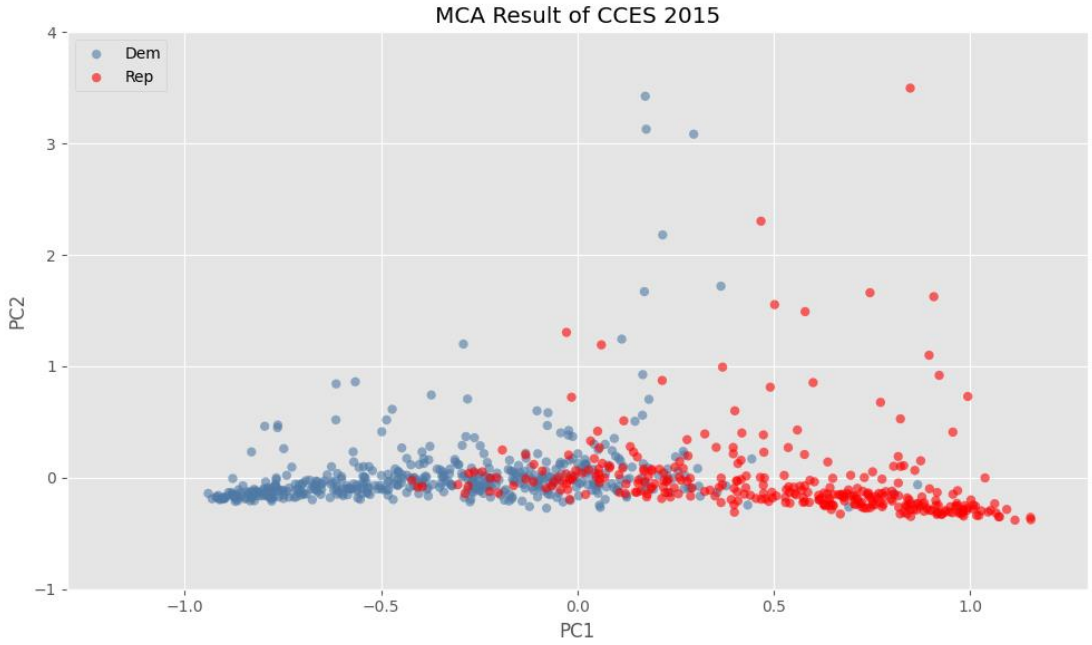


Fig. 2: MCA

Basic Space (BS)

The Basic-Space (BS) procedure, introduced by Poole in 1998, is a mathematical method for analyzing political ideologies and alignments, particularly in political science and survey data. It simplifies multidimensional data into a lower-dimensional space to reveal underlying ideological structures and is commonly used to study voting patterns, party affiliation, and polarization.

BS helps analyze complex political data where relationships between variables, such as party affiliation and ideological positions, are not straightforward. It identifies latent factors, reduces dimensionality, and visualizes ideological divisions and alignments within a dataset.

The methodology applies techniques from multidimensional scaling and factor analysis to organize individuals or entities (e.g., voters or candidates) along axes reflecting their ideological positions. By measuring similarity or distance, it uncovers latent dimensions of political preference, offering a clear, interpretable representation of political space. (Poole 1998)

To implement Basic Space (Poole 1998) in R, we followed these steps:

1. **Handling Missing Data**: We handled missing values in the dataset by replacing missing values with a placeholder value (99), which is commonly used for missing data.
2. **Data Conversion**: All columns were converted to numeric values to ensure compatibility for analysis, as Basic Space requires numerical data for dimensionality reduction.
3. **Dimension Reduction**: The BlackBox method was applied to reduce the data into two dimensions (PC1 and PC2), creating a lower-dimensional representation of the dataset that highlights the ideological space.
4. **Visualization**: A scatter plot was created, where the points were color-coded by party affiliation (Democrats vs. Republicans), showing how the different party groups are distributed in the ideological space.

These steps ensure that the data is cleaned, reduced to two dimensions, and visualized, revealing the ideological patterns in the dataset as shown in Fig. 3. The scatter plot illustrates the distribution of respondents across the two reduced dimensions, PC1 and PC2, with party affiliation color-coded (blue for Democrats and red for Republicans). This visualization highlights the ideological

structures within the dataset, showing potential polarization, with the overlap and separation of colors indicating the degree of ideological alignment and division between the party groups.

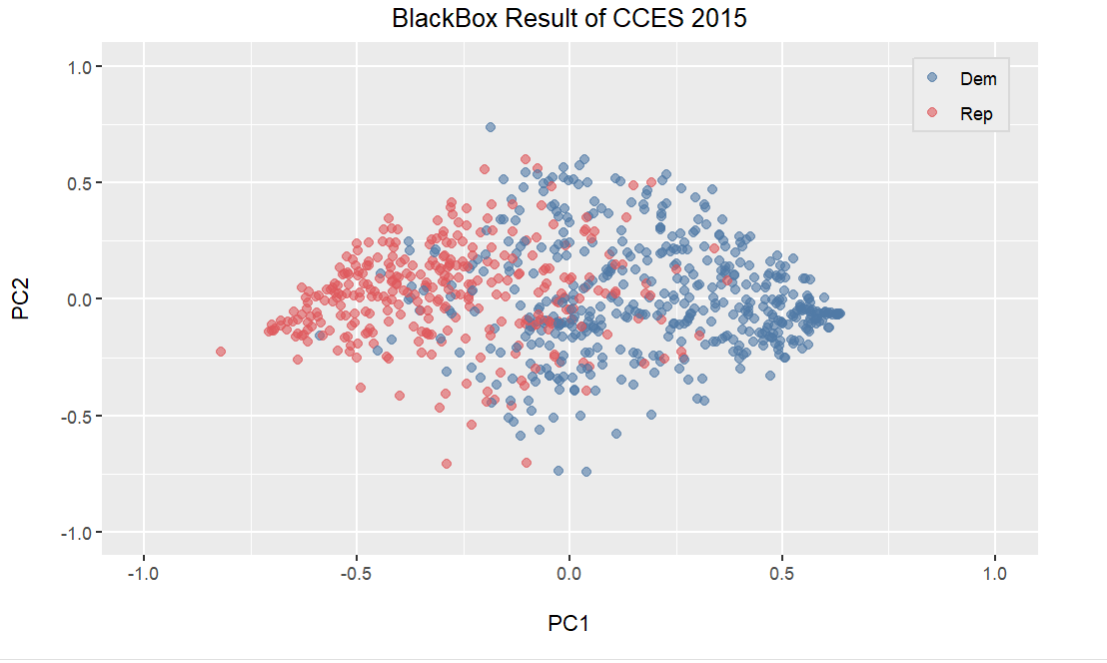


Fig. 3: BS

Bayesian Mixed Factor Analysis (BMFA)

Bayesian Mixed Factor Analysis (BMFA) is a statistical technique for uncovering latent dimensions in datasets containing both continuous and categorical variables. By extending traditional factor analysis into the Bayesian framework, BMFA provides a probabilistic approach to modeling and estimation, offering robust handling of uncertainty and measurement error. It is particularly useful for high-dimensional datasets in social sciences, political studies, and other fields with diverse response types, including ordinal, nominal, and continuous variables.

BMFA's strength lies in its ability to model mixed data types while accounting for uncertainty using Bayesian priors. This flexibility helps identify underlying structures, manage missing data, and handle high-dimensional spaces more effectively than traditional factor analysis. Likelihood functions tailored to observed data types ensure accurate estimation across diverse variable formats.

The methodology estimates latent dimensions through Bayesian inference, specifying priors for factor loadings and scores. Parameters are estimated using Markov Chain Monte Carlo (MCMC) methods to obtain posterior distributions of latent traits. These dimensions can then be visualized to reveal clustering, polarization, or patterns across groups, providing insights into the dataset's structure. BMFA has been widely applied in political analysis to study ideological patterns. (Martin, Quinn, and Park 2011; Quinn 2004)

To implement Bayesian Mixed Factor Analysis (BMFA) in Python, we followed these steps:

1. **Data Preparation**: The dataset was processed to map party affiliations to "DEM" (Democrats) and "REP" (Republicans), with other values treated as missing. The partyid variable was converted to a categorical variable, and the original column was removed for clean modeling.
2. **Model Specification**: Using PyMC:

- Two latent factors (factors) and a matrix of factor loadings (loadings) were defined with normal priors. Loadings for ideology were constrained to ensure positive impact.
- Predictions were modeled as the dot product of factors and loadings, with observed data modeled as a normal distribution.
- The PyMC sampler generated 1,000 samples with a tuning period of 1,000 iterations.

3. **Visualization**: Posterior means of latent factors (coord1D and coord2D) were paired with party affiliations for visualization. A scatter plot color-coded by party (blue for Democrats, red for Republicans) highlighted ideological positions and polarization.

The result plot, shown in Figure 4, displays data points representing respondents, color-coded by party affiliation: blue for Democrats and red for Republicans. The two principal components, PC1 and PC2, represent the latent factors identified through the analysis. This visualization reveals a clear ideological gradient, with distinct clustering of party affiliations that highlights patterns of ideological positions and polarization within the dataset.

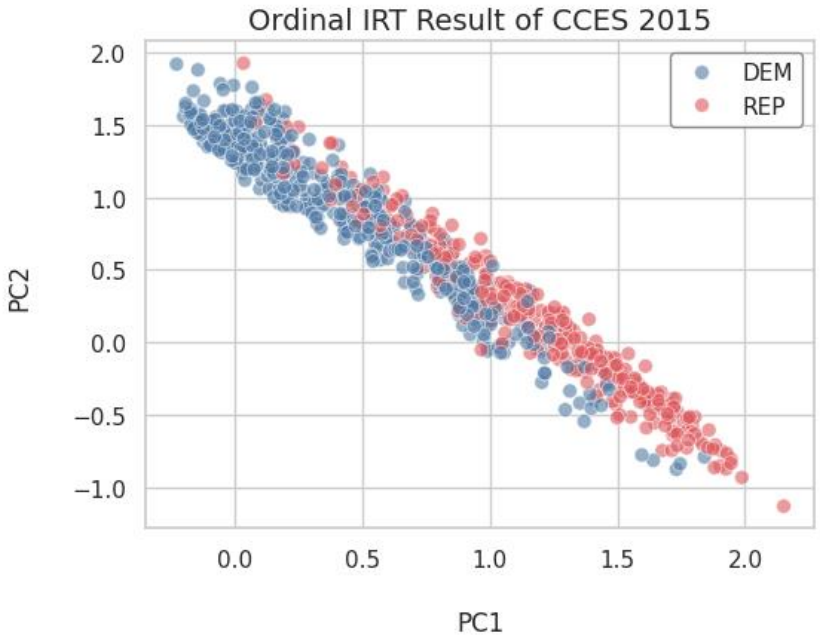


Fig. 4: BMFA

The four methods of dimensional reduction and estimation reveal a consistent pattern among American voters, highlighting a clear separation between Democrats and Republicans on the first principal component or left-right scale, based on self-reported values. (see Fig.1, Fig. 2, Fig. 3, and Fig. 4) This suggests that voters' preferences, closely tied to party affiliation and vote choices, show significant ideological polarization. This aligns with earlier research. (Fujiwara and Liu 2020) Additionally, we are going to use cMCA to assess its effectiveness in distinguishing subgroups, where traditional methods also perform well.

Contrastive Multiple Correspondence Analysis (cMCA)

Contrastive learning (CL) is an emerging machine learning approach, which analyzes high dimensional data to capture patterns that are specific to, or enriched in, one dataset relative to another (Abid et al. 2018). Unlike ordinary approaches, such as PCA and MCA, that usually aim to capture the characteristics of one entire dataset, CL compares subsets of that dataset relative to one another (target versus background group). The logic behind CL is to explore unique components or directions that contain more salient latent patterns in one dataset (the target group) than the other (background group). (Fujiwara and Liu 2020)

Dataset

The data mainly comes from 2015 Cooperative Congressional Election Study (CCES 2015) and is combined the UGA2015. The variables in the data are categorical, representing each of the answers on the surveys.

In this study we found two datasets.

1. Issuevalue_short: This dataset is a sample of the original data provided used by the authors.
    - Observations: 1000
    - Variables: 53
2. Raw data of the survey:

    CCES 2015:

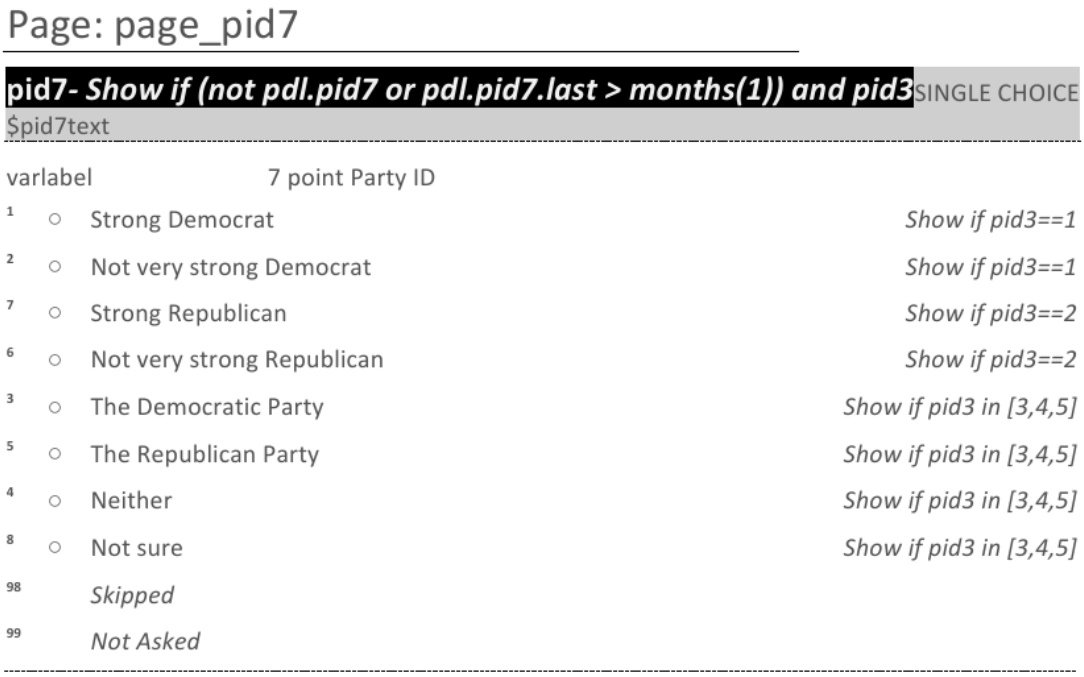
    https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/SWMWX8
        - Observations: 14248
        - Variables: 254

    UGA 2015: not available

Target Variable

The variables in the issuevalue_short dataset consist of categorical responses to survey questions. The most important variable, which serves as the target variable, is "partyid," originally labeled as "pid7" in the dataset.

The variable *partyid/pid7* is a categorical ordinal variable that reflects survey responses indicating the degree of affiliation with either the Democratic or Republican party. It also includes categories for neutral responses, skipped answers, and cases where the question was not asked. Fig. 5


Page: page_pid7

pid7- *Show if (not pdl.pid7 or pdl.pid7.last > months(1)) and pid3* SINGLE CHOICE

$pid7text

varlabel		7 point Party ID	
1	○	Strong Democrat	*Show if pid3==1*
2	○	Not very strong Democrat	*Show if pid3==1*
7	○	Strong Republican	*Show if pid3==2*
6	○	Not very strong Republican	*Show if pid3==2*
3	○	The Democratic Party	*Show if pid3 in [3,4,5]*
5	○	The Republican Party	*Show if pid3 in [3,4,5]*
4	○	Neither	*Show if pid3 in [3,4,5]*
8	○	Not sure	*Show if pid3 in [3,4,5]*
98		*Skipped*	
99		*Not Asked*	

Fig. 5: 7-point party id (source: CCES15_Common_questionnaire)

Data Imputation Process

Due to the unavailability of the raw dataset, we used issuevalue_short, a transformed subset of the original data, for our cMCA analysis in this study. This dataset, provided as a sample in the paper, was selected because key variables for target sub setting were derived from UGA2015, which was not available to be found in raw dataset.

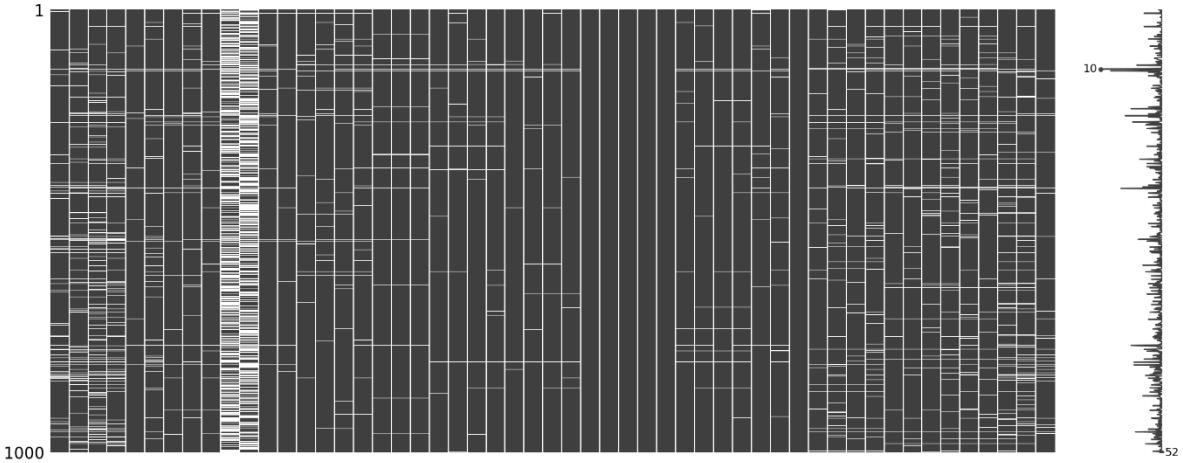


Fig. 6: Missing value chart of issuevalue_short dataset

To apply the cMCA, the dataset could not contain missing values, given that the target variable is partyid and the data is categorical. First, two columns were dropped because they had more than 50% missing values. Second, missing values were imputed using a logic that involved using the mode of each variable based on partyid. This imputation method did not affect the results of answers by political affinity.

Dropped columns:

- renewpatriotact: 523 missing values in the dataset
- usafreedomact: 508 missing values in the dataset

Imputation Logic

Next, an example of the imputation logic applied to the economics variable is provided; the same logic was applied to all variables.

1. Get a heatmap of the variable by *partyid,* shown in Fig. 7.

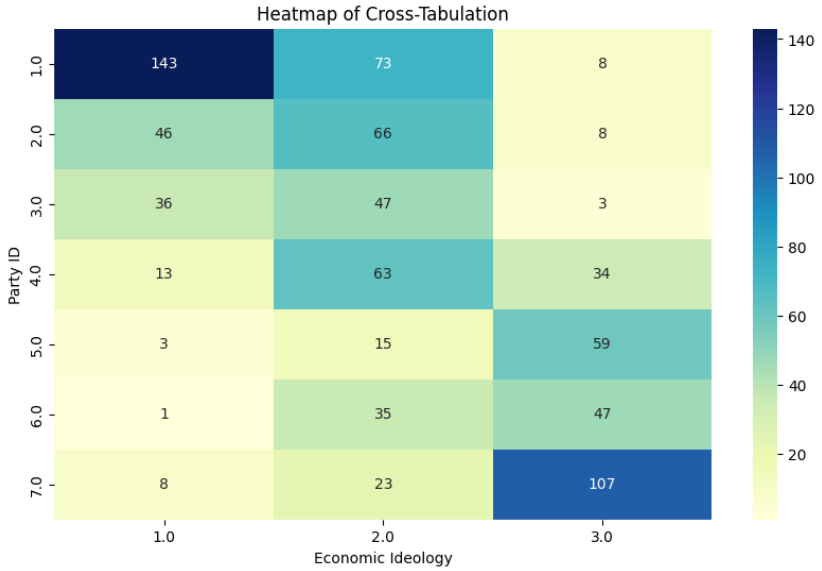


Fig. 7: Heatmap of the variable by partyid

2. Assign imputation to missing values depending on mode/most frequent values by partyid.

The heatmap of the variable by partyid Fig. 7 illustrates,

- For Party ID 1, mode of economic ideology is 1,
- For Party ID 2,3 and 4, the mode/most frequent value is 2.
- For Party ID 5,6,7 the, the mode/most frequent value is 3.

Therefore, we assigned the values using the logic presented, and the same logic was applied to all the variables.

Implementation of code for cMCA

To comprehend the code used for applying cMCA to a dataset, it's crucial to distinguish between selecting a target group and a background group. The target group represents the focus of the study, whereas the background group functions as the reference. This reference group provides a baseline for comparing with the target group.

From this point forward, we will outline the step-by-step process used to implement cMCA.

1. Select the Background and Target.

```python
# Background
group1 = data[data['partyid'] > 4].drop(columns=['partyid']) # Republicans

# Target
group2 = data[data['partyid'] < 4].drop(columns=['partyid']) # Democrats
```

2. Apply PCA the to selected background and PCA transform to the target.

```python
# Apply PCA (proxy for MCA) directly to both groups without one-hot encoding
pca = PCA(n_components=2)

# Fit PCA on group 1 and transform both groups
group1_pca = pca.fit_transform(group1)
group2_pca = pca.transform(group2)
```

3. Compute the contrast difference between Target and background.

```python
# Compute the contrast (difference in coordinates)
contrast = group2_pca - np.mean(group1_pca, axis=0)
```

4. Create the visualization of contrastive.

```python
# Visualization
plt.figure(figsize=(10, 6))
plt.scatter(group1_pca[:, 0], group1_pca[:, 1], alpha=0.6, label="Republicant", color="red")
plt.scatter(group2_pca[:, 0], group2_pca[:, 1], alpha=0.6, label="Democrats", color="blue")
plt.quiver(
    np.mean(group1_pca[:, 0]),  # X-coordinate of the arrow starting point
    np.mean(group1_pca[:, 1]),  # Y-coordinate of the arrow starting point
    np.mean(contrast[:, 0]),  # X-component of the arrow vector (mean contrast for dimension 1)
    np.mean(contrast[:, 1]),  # Y-component of the arrow vector (mean contrast for dimension 2)
    angles='xy', scale_units='xy', scale=1, color='black', alpha=0.7, label='Contrast'
)

plt.title(" Original Contrastive(cMCA) : Target Democrats vs Backgroung Republicants")
plt.xlabel("cPC 1")
plt.ylabel("cPC 2")
```

Result of cMCA

(a) Target: Dem, Background: Rep (Original cMCA Results)

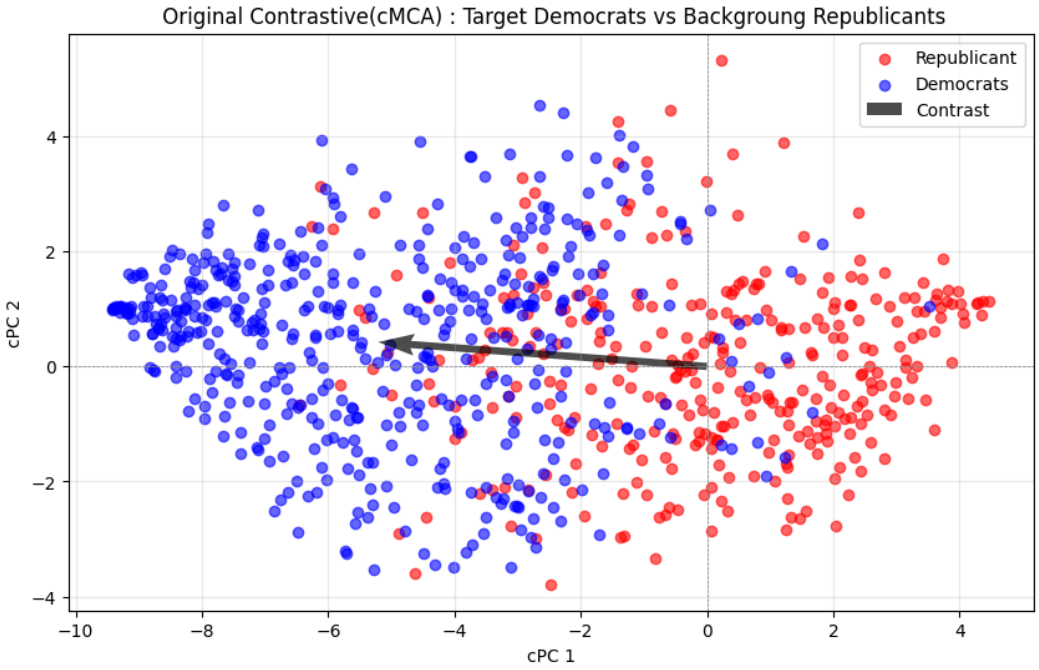


Fig. 8: Target: Dem, Background: Rep original cMCA Results of CCES2015

The Contrastive Multiple Correspondence Analysis (cMCA) chart effectively illustrates the ideological separation between Democrats and Republicans. Democrats, represented by blue dots, are clustered toward the left side of the plot, while Republicans, shown in red, are concentrated on the right. This indicates a significant difference in their underlying attributes, primarily captured along the first principal component (cPC1). The contrast vector, depicted as a black arrow pointing from right to left, further reinforces this distinction by illustrating the systematic shift between the two groups. The second principal component (cPC2) does not show a strong separation, suggesting that the primary differences are captured along cPC1. While the groups are largely distinct, some overlap in the middle of the plot indicates potential centrist positions or shared characteristics among individuals. Overall, the cMCA results highlight a clear differentiation between Democrats and Republicans, with the contrast vector emphasizing the key ideological divergence while acknowledging some degree of convergence in the central region.

(b) Target: Dem, Background: Rep (with Background Rep)

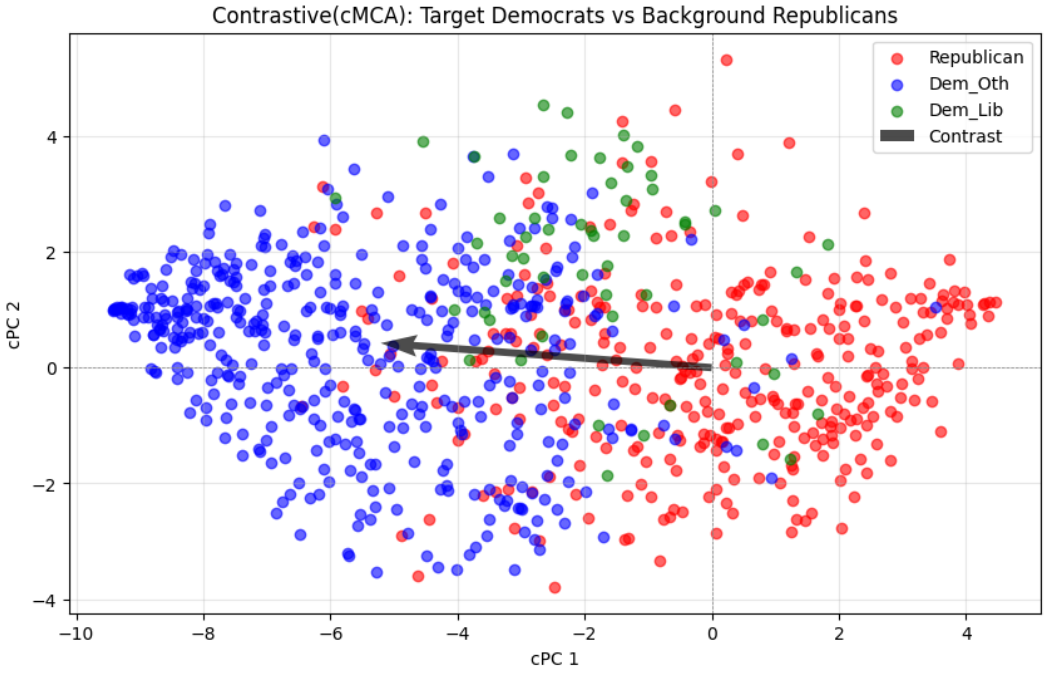


Fig. 9: Target: Dem, Background: Rep (with Background Rep) cMCA Results of CCES2015

The Contrastive Multiple Correspondence Analysis (cMCA) chart visually distinguishes Democrats and Republicans based on their ideological attributes. Republicans, represented by red dots, are predominantly concentrated on the right side of the plot, whereas Democrats are divided into two subgroups: Dem_Oth (blue) and Dem_Lib (green). Both Democratic subgroups are mostly positioned on the left side, indicating a strong separation from Republicans along the first principal component (cPC1), which captures the primary ideological differences.

(c) Target: Dem, Background: Rep (without Background)

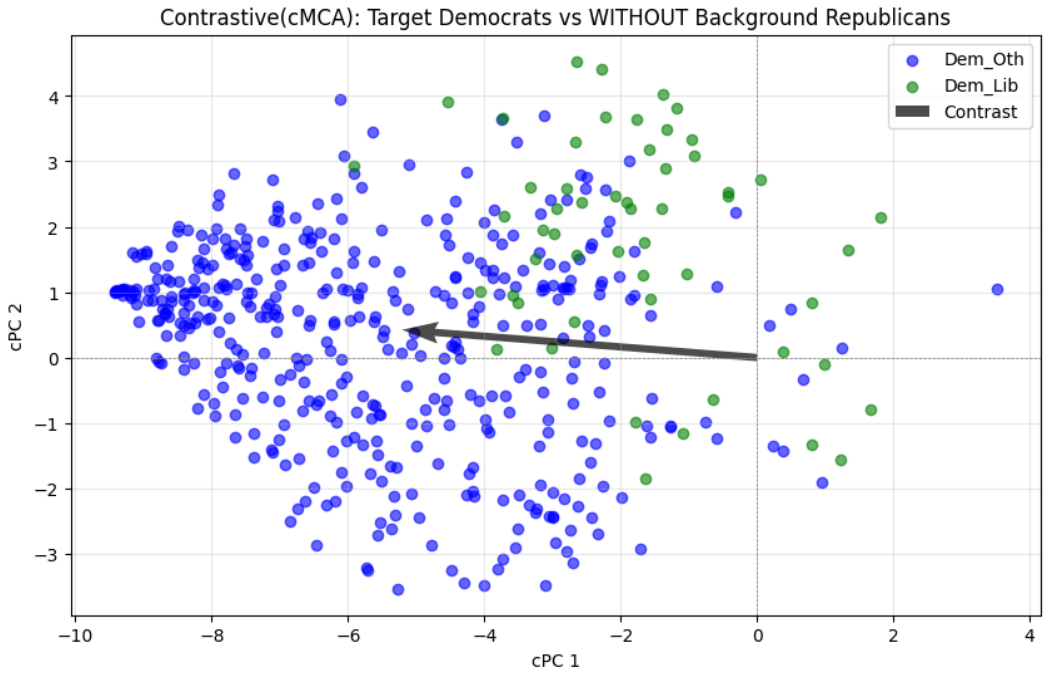


Fig. 10: Target: Dem, Background: Rep (without Background) cMCA Results of CCES2015

The Contrastive Multiple Correspondence Analysis (cMCA) chart visualizes the ideological distribution within the Democratic group, without including Republicans as a background reference. The data points represent two Democratic subcategories: Dem_Oth (blue) and Dem_Lib (green). The chart reveals a notable distinction between these subgroups, with Dem_Oth clustered more tightly on the left and Dem_Lib appearing more dispersed across the right side.

(d) Target: Rep, Background: Dem (Original cMCA Results)

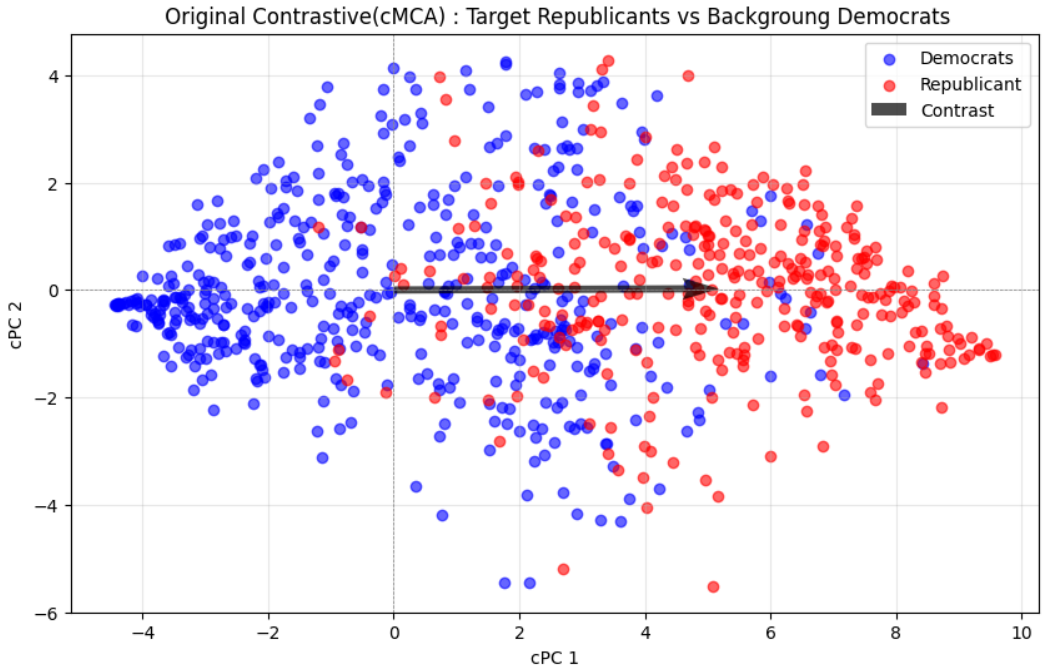


Fig 11: Target: Rep, Background: Dem original cMCA Results of CCES2015

The Contrastive Multiple Correspondence Analysis (cMCA) chart highlights the ideological positioning of Republicans (red) against the background of Democrats (blue). The first principal component (cPC1) serves as the primary axis of separation, with Democrats concentrated on the left side and Republicans occupying the right. This indicates strong categorical distinctions between the two groups along cPC1. The contrast vector, depicted as a black arrow, points from the Democratic cluster to the Republican cluster, emphasizing the systematic differences in categorical features defining Republicans compared to the Democratic baseline. Some overlap near the center of the chart (cPC1 $\approx$ 0) suggests the presence of individuals with centrist or mixed ideological positions. The second principal component (cPC2) does not show significant differentiation, indicating that most of the variance is captured by cPC1. Overall, the cMCA results reveal a clear ideological contrast between Republicans and Democrats, with Republicans systematically deviating from the Democratic baseline while retaining some degree of blending in the central region.

(e) Target: Rep, Background: Dem (with Background)

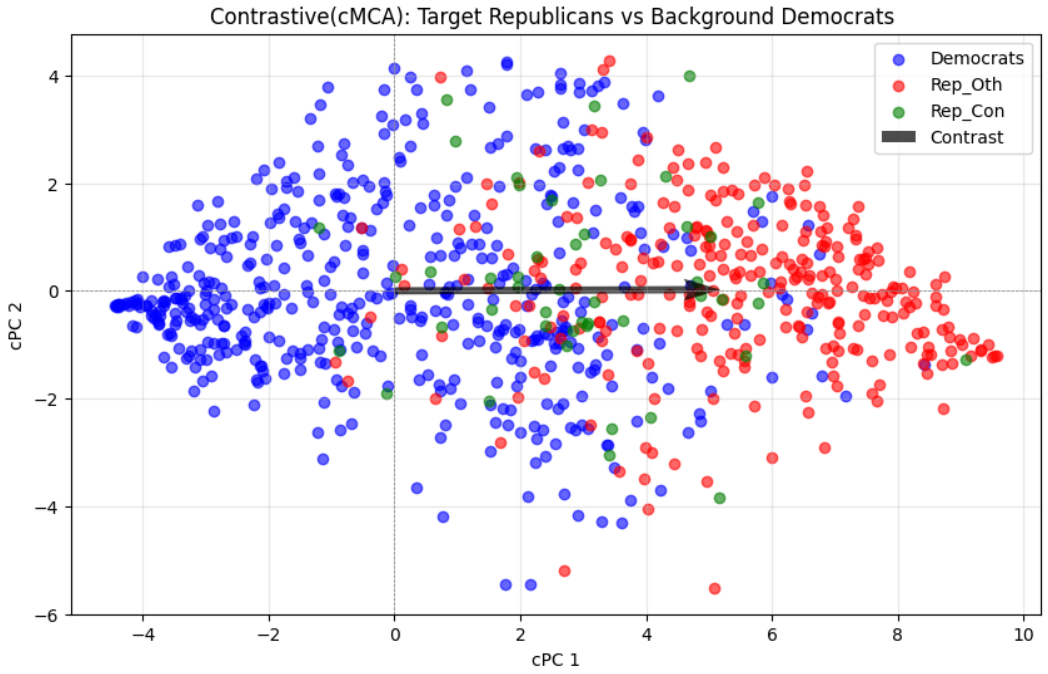


Fig. 12: Target: Rep, Background: Dem (with Background) cMCA Results of CCES2015

The Contrastive Multiple Correspondence Analysis (cMCA) chart illustrates the ideological positioning of Republicans (split into Rep_Oth (red) and Rep_Con (green)) against the background of Democrats (blue). The first principal component (cPC1) serves as the primary axis of separation, with Democrats concentrated on the left and Republicans on the right, reinforcing the stark contrast between the two groups. The presence of Rep_Oth and Rep_Con as subcategories suggests internal variation among Republicans, with Rep_Con (green) appearing more dispersed and possibly representing a distinct ideological subgroup.

(f) Target: Rep, Background: Dem (without Background)

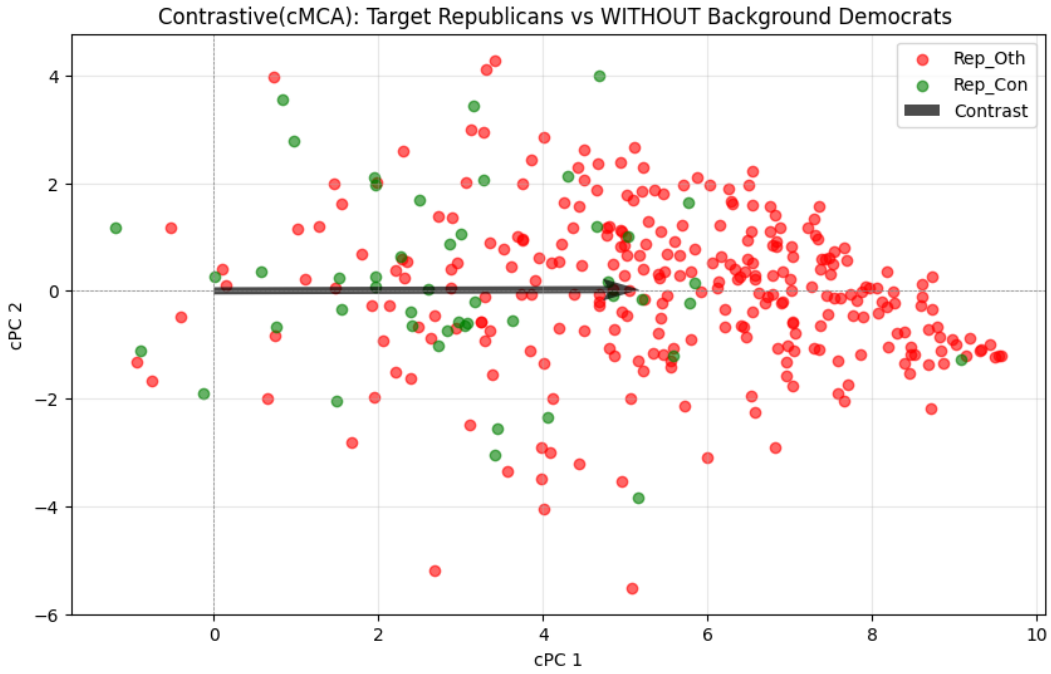


Fig. 13: Target: Rep, Background: Dem (without Background) cMCA Results of CCES2015

The Contrastive Multiple Correspondence Analysis (cMCA) chart presents the ideological distribution within the Republican group, specifically distinguishing between Rep_Oth (red) and Rep_Con (green), without a Democratic background for reference. The first principal component (cPC1) remains the dominant axis of separation, with Rep_Oth and Rep_Con distributed across the right-hand side of the plot. The absence of Democrats means that this visualization focuses exclusively on internal differences within the Republican cohort rather than a partisan contrast.

Findings of cMCA Study

Target: Democrats with Republicans as Background

- The cMCA results show strong differentiation between Democrats and Republicans.
- cPC1 serves as the primary axis of separation, capturing ideological or categorical divergence.
- The contrast vector clearly indicates a direction from Republicans to Democrats, highlighting systematic differences between the two groups.
- Some overlap in the center suggests potential areas of ideological convergence or individual variation.
- Democrats can be sub-grouped by Democrats Liberal and Democrat Orthodox.

This charts effectively illustrates how Democrats and Republicans differ in their categorical attributes, making it a valuable tool for understanding ideological structures.

Target: Republicans with Democrats as Background

- The cMCA results demonstrate a pronounced left-to-right ideological contrast between Republicans and Democrats.
- Republicans are positioned further right on cPC1, underlining their distinct categorical features.
- The contrast vector confirms that Republicans systematically differ from the Democratic baseline.
- While some ideological blending exists in the central region, the primary differences remain stark.
- Republicans can be sub-grouped by Republican Conservatives and Republican Orthodox.

This chart effectively showcases how Republicans distinguish themselves from Democrats through categorical response patterns, reinforcing the power of cMCA in capturing political differentiation.

Discussion

Our study highlights the effectiveness of contrastive Multiple Correspondence Analysis (cMCA) in providing deeper insights into political group differentiation, specifically between Democrats and Republicans. By applying cMCA, we identified distinct ideological structures that traditional principal component (PC) methods might overlook. This approach allows for a nuanced understanding of the factors and issues that are salient within each group.

The original research paper discusses how typical PC methods often fail to provide informative results due to the inherent structure of different datasets. In contrast, cMCA enables researchers to derive contrasted dimensions from categorical data. Our findings are consistent with this, as cMCA successfully identified substantively important dimensions and divisions that offer complementary information even when traditional PC analysis finds meaningful low-dimensional spaces. For instance, in situations where conventional methods revealed overlapping ideological distributions, cMCA was able to pinpoint specific divisions, such as differences in policy priorities between progressive and moderate Democrats.

Ultimately, our use of cMCA underscores its utility as a powerful tool for uncovering hidden patterns and ideological divergences within political data, advancing our ability to understand and interpret complex political landscapes.

Bibliography

Abid, Abubakar, Martin J. Zhang, Vivek K. Bagaria, and James Zou. "Exploring Patterns Enriched in a Dataset with Contrastive Principal Component Analysis." Nature Communications 9, no. 1 (2018): 1-7.

Adams, James, Jane Green, and Caitlin Milazzo. "Has the British Public Depolarized along with Political Elites? An American Perspective on British Public Opinion." Comparative Political Studies 45, no. 4 (2012): 507-530.

Armstrong, David, Ryan Bakker, Royce Carroll, Christopher Hare, Keith Poole, and Howard Rosenthal. “Analyzing Spatial Models of Choice and Judgment with R”. FL: CRC Press, 2014.

Clyde H. “A Theory of Data”. New York: Wiley, 1964.

Endo, Masahisa, and Willy Jou. "How Does Age Affect Perceptions of Parties' Ideological Locations?" Japanese Journal of Electoral Studies 30, no. 1 (2014): 96-112.

Fiorina, Morris, and Samuel Abrams. "Political Polarization in the American Public." Annual Review of Political Science 11 (2008): 563-588.

Fujiwara, Takanori, Oh-Hyun Kwon, and Kwan-Liu Ma. "Supporting Analysis of Dimensionality Reduction Results with Contrastive Learning." IEEE Transactions on Visualization and Computer Graphics 26, no. 1 (2020): 45-55.

Fujiwara, Takanori, and Tzu-Ping Liu. "Contrastive Multiple Correspondence Analysis (cMCA): Applying the Contrastive Learning Method to Identify Political Subgroups." University of California, Davis, July 20, 2020.

Greenacre, Michael. “Correspondence Analysis in Practice”. New York: Chapman & Hall/CRC, 2017.

Hare, Christopher, Tzu-Ping Liu, and Robert N. Lupton. "What Ordered Optimal Classification Reveals About Ideological Structure, Cleavages, and Polarization in the American Mass Public." *Political Science Research and Methods* 6, no. 4 (2018): 681-699. https://doi.org/10.1017/psrm.2018.10.

Harris, David, and Sarah Harris. “Digital Design and Computer Architecture”. CA: Morgan Kaufmann Publishers, 2010.

Martin, Andrew D., Kevin M. Quinn, and Jong Hee Park. "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software* 42, no. 9 (2011): 1–21.

Iyengar, Shanto, and Sean J. Westwood. "Fear and Loathing Across Party Lines: New Evidence on Group Polarization." American Journal of Political Science 59, no. 3 (2015): 690-707.

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

Quinn, Kevin M. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12, no. 4 (2004): 338–353.