Mapeando comportamentos com estimação de pontos ideais

Rodrigo Martins - Universidade Federal de Pernambuco

Resumo

Estimação de pontos ideais é uma metodologia de importância crescente na Ciência Política. Desenvolvida desde meados do século passado, seu vínculo com a teoria espacial do voto se mostrou frutífera em diversos campos de estudo, com grandes expectativas sobre seu desenvolvimento futuro. O objeto deste trabalho é apresentar o que é estimação de pontos ideais, e ilustrar didaticamente como aplicar uma das várias técnicas disponíveis, utilizando dados sobre a atuação parlamentar da Câmara dos Deputados no ano de 2019.

Palavras-chave: comportamento; preferências; ideologia; pontos ideais

Abstract

Ideal point estimation is a methodology of growing importance in Political Science. Developed in the mid-20th century, its link with the spatial theory of voting has been fruitful for several fields of study, with great expectations on its future development. The goal of this paper is to present what ideal point estimation is and didactically illustrate how to apply one of the various available techniques, using data on parliamentary behavior from the Chamber of Deputies in 2019.

Keywords: behavior; preferences; ideology; ideal point estimation

1. Introdução

O que estrutura as divisões entre parlamentares, ideologia ou governismo? Mudança de partido afeta o comportamento de legisladores? Qual é o nível de coesão interna dos partidos políticos e como ela oscila ao longo do tempo? A polarização no Congresso Nacional aumentou? Quais políticos possuem comportamento mais imprevisível? Quais tipos de projetos de lei costumam polarizar o comportamento político? Como a capacidade de um presidente estabelecer um governo de coalizão influencia o comportamento parlamentar?

Estas são algumas questões levantadas pela literatura de estudos legislativos para explicar a atividade parlamentar. Em comum a todas elas está a necessidade de acessar preferências, ideologias e comportamentos de representantes eleitos. No entanto, campos diversos da Ciência Política possuem o mesmo interesse com relação a outros tipos de atores políticos e outros tipos de comportamento político. É possível encontrar evidências de comportamento ideológico no poder judiciário? Os discursos de políticos expressam diferenças ideológicas? As preferências ideológicas do eleitorado refletem as preferências de seus representantes? A opção de seguir certas pessoas e não outras no Twitter reflete a preferência política dos usuários da plataforma?

Essas e diversas outras questões são levantadas por estudos que utilizam estimações de pontos ideais para explicar comportamento de atores políticos. A teoria espacial do voto, já teorizada por (Downs, 1957), passou a ser aplicada às mais diversas áreas, não apenas para explicar políticos tradicionais como também para lançar luz sobre as preferências daqueles que estão fora das instituições políticas. A facilidade de se traduzir formalizações teóricas de formação de preferências em fórmulas matemáticas e estatísticas fez com que a teoria espacial do voto obtivesse sucesso para ser implementada de diversas formas.

O objetivo deste trabalho é apresentar, de forma simples e intuitiva, o que é estimação de pontos ideais e ilustrar como utilizá-la didaticamente. Apesar de utilizada para os mais diversos objetivos, o foco aqui será no campo de estudos que mais desenvolveu esta técnica, a área de estudos legislativos. Ilustraremos como estimar pontos ideais utilizando dados sobre da Câmara dos Deputados no ano de 2019, durante o primeiro ano do mandato presidencial de Jair Bolsonaro. Estudos anteriores, que utilizaram estimação de pontos ideais para o caso brasileiro, mostram que a dimensão principal que estrutura o comportamento parlamentar do Senado e Câmara dos Deputados não é a ideológica, mas sim a dimensão governo-oposição (Izumi, 2016; Zucco, 2009), embora em alguns momentos tais dimensões pareçam estar sobrepostas (Leoni, 2002). Diante do diagnóstico de que Bolsonaro alterou a relação entre Executivo e Legislativo ao não compor base de apoio parlamentar por meio de uma coalizão, substituindo "o presidencialismo de coalizão pelo desleixo" (Limongi, 2019), haveria uma mudança no comportamento da Câmara dos Deputados? Haveria uma outra dimensão estruturando as votações nominais em plenário?

2. O que são pontos ideais?

Estimação de pontos ideais deriva do modelo espacial do voto, inicialmente popularizada na Ciência Política por Downs (1957), onde atores politicos podem ser representados como um ponto em um espaço e optam por alternativas políticas, também representadas espacialmente, que estão mais próximas de si. Além de ser um princípio teórico, diversos estudos mostram empiricamente que a maior parte do comportamento político pode ser representado em um espaço unidimensio-

nal, frequentemente tido como uma escala ideológica. A divisão entre esquerda e direita seria uma expressão de como atores políticos, sejam estes eleitores ou representantes, se comportam e manifestam suas preferências de maneira alinhada em assuntos distintos, como política econômica e política social, no mesmo sentido em que Converse (1964) argumenta na direção de um sistema de crença que permeia preferências de assuntos diversos em uma visão unificada de mundo.

A técnica de estimação de pontos ideais nos mostra um mapa de como votantes se distribuem espacialmente, representando a distância entre as preferências de cada um dos indivíduos em análise. De maneira simplificada, o que extraímos é a proximidade entre os votantes quando consideramos o conjunto de todos os indivíduos, e não apenas a semelhança entre eles par a par. Dessa forma é possível, inclusive, estimar a proximidade entre votantes que na prática não participaram juntos de nenhuma votação, pois ao estimar-se a distância destes com relação a todos os outros votantes, saberíamos também a distância entre eles. Um dos pressupostos do modelo é que ao sabermos como um indivíduo decide sobre um tema em específico, seria possível prever seu comportamento em relação a outras decisões, especialmente em temas semelhantes, a partir da derivação da decisão anterior. Dessa forma, ao correlacionar todas as decisões, seria possível estimar coordenadas para os indivíduos, de maneira que aqueles que decidem de forma semelhante fiquem espacialmente próximos, e aqueles que decidem de forma divergente fiquem espacialmente distantes.

Para ilustrar a ideia subjacente à estimação de pontos ideais, podemos recorrer a um exemplo hipotético. Suponhamos que exista um colegiado de três legisladores que precisam tomar uma decisão sobre uma matéria que diz respeito a temática econômica. Considerando que um dos legisladores vota por favorecer mais intervenção do estado na economia e os outros dois votam por menos intervenção, a representação espacial do voto dos legisladores será feita de tal forma que os dois que votaram juntos ocupem o mesmo lado, em oposição ao votante que ficou isolado. Também é possível ilustrar a matéria em votação como uma linha que representa a divisão entre os legisladores em dois lados diferentes. A figura 1 seria a forma de representar espacialmente a divisão neste exemplo. Os círculos vermelho e verde seriam os dois que votaram juntos, o círculo azul o indivíduo que votou isolado e a reta roxa seria a linha da matéria em votação que representa o corte entre os dois lados em questão.

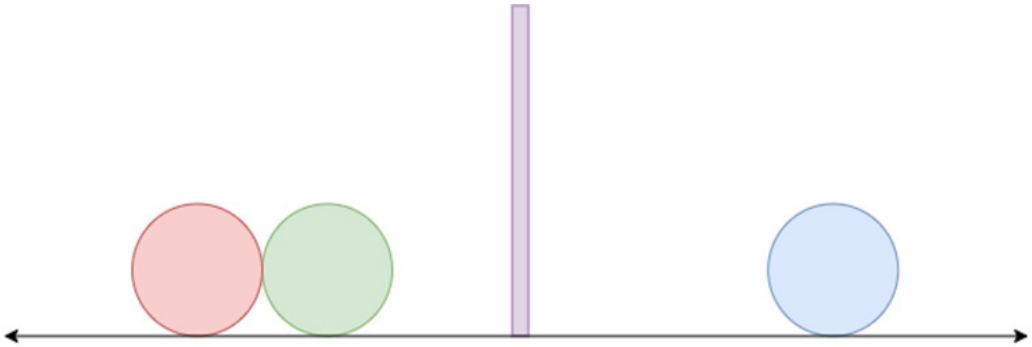


Figura 1: Exemplo de análise espacial com 1 votação e 3 votantes

Suponhamos agora que em uma segunda votação sobre o mesmo tema o legislador representado pela cor verde não vote novamente junto com o votante representado pela cor vermelha, mas sim com o representado pela cor azul. Neste segundo momento, o indivíduo verde deve estar posicionado entre os outros dois de forma equidistante, já que votou uma vez junto com cada um. Ao mesmo tempo, aqueles que não votaram juntos em nenhuma das duas votações devem permanecer em extremos opostos. A linha que representa a primeira votação, em cor roxa, deve manter a divisão entre os votantes da mesma forma que anteriormente. E agora uma segunda linha de votação, que será representada pela cor branca na figura 2, deve deixar isolado o legislador que votou sozinho e manter do mesmo lado aqueles que votaram juntos.

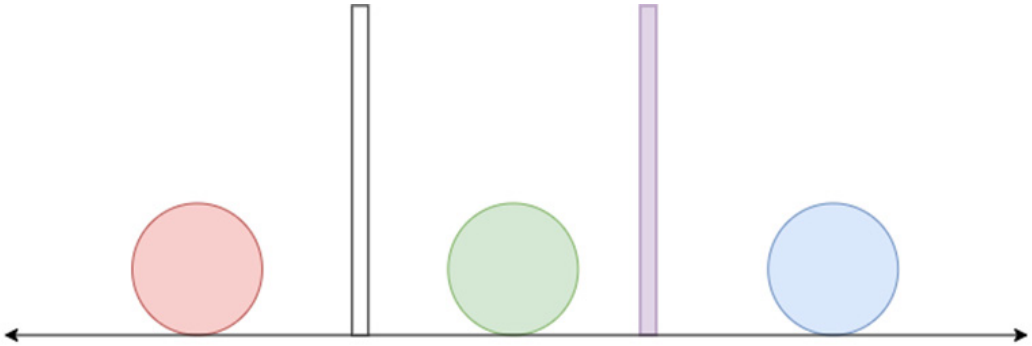


Figura 2: Exemplo de análise espacial com 2 votações e 3 votantes

Como último exemplo, imaginemos que em uma terceira votação sobre outro assunto não relacionado a economia os legisladores que antes se opunham agora votam juntos, deixando o votante que antes estava no centro isolado na minoria. Para que a representação espacial se mantenha mais apropriada a esta situação, o mais adequado é adicionar um eixo vertical onde aquele que foi derrotado se distancie dos outros dois, enquanto os dois vencedores se mantêm próximos um do outro neste segundo eixo. Também podemos representar esta terceira votação como uma linha, representada em amarelo na figura 3, mas desta vez na horizontal.

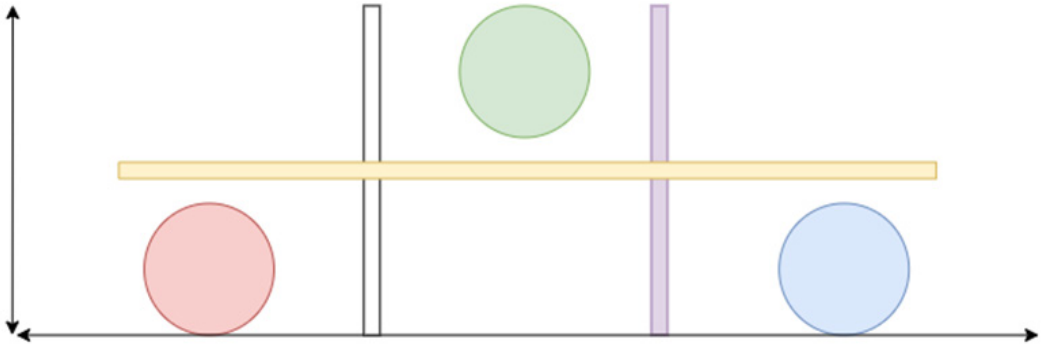


Figura 3: Exemplo de análise espacial com 3 votações e 3 votantes

Esse tipo de modelo projeta as preferências manifestadas através do voto em um espaço (unidimensional, bidimensional e até mesmo além) avaliando as similaridades e divergências entre as decisões de cada um dos votantes. Como não seria factível fazer esse processo manual para muitos votantes em muitas votações, foram desenvolvidas diversas técnicas, com procedimentos estatísticos distintos entre si, com o objetivo de estimar pontos ideais tanto para votantes quanto para votações.

No entanto, é de se imaginar que em algum momento não é mais possível representar perfeitamente a distribuição espacial dos votantes e das votações utilizando-se apenas uma ou duas dimensões. Com o aumento de votos e votações, torna-se inevitável que os pontos de votantes acabem sendo alocados em uma posição que não corresponde ao seu voto em dada matéria. (Poole, 2005) propõe duas medidas para avaliar e comparar o desempenho de modelos de pontos ideais: a porcentagem de classificações corretas e a redução proporcional do erro agregado (RPEA). Para explicar melhor o que são tais medidas e quais as diferenças entre elas, vamos utilizar um exemplo hipotético como ilustração. A figura 4 representa uma votação com 11 votantes identificados por uma letra e uma cor.

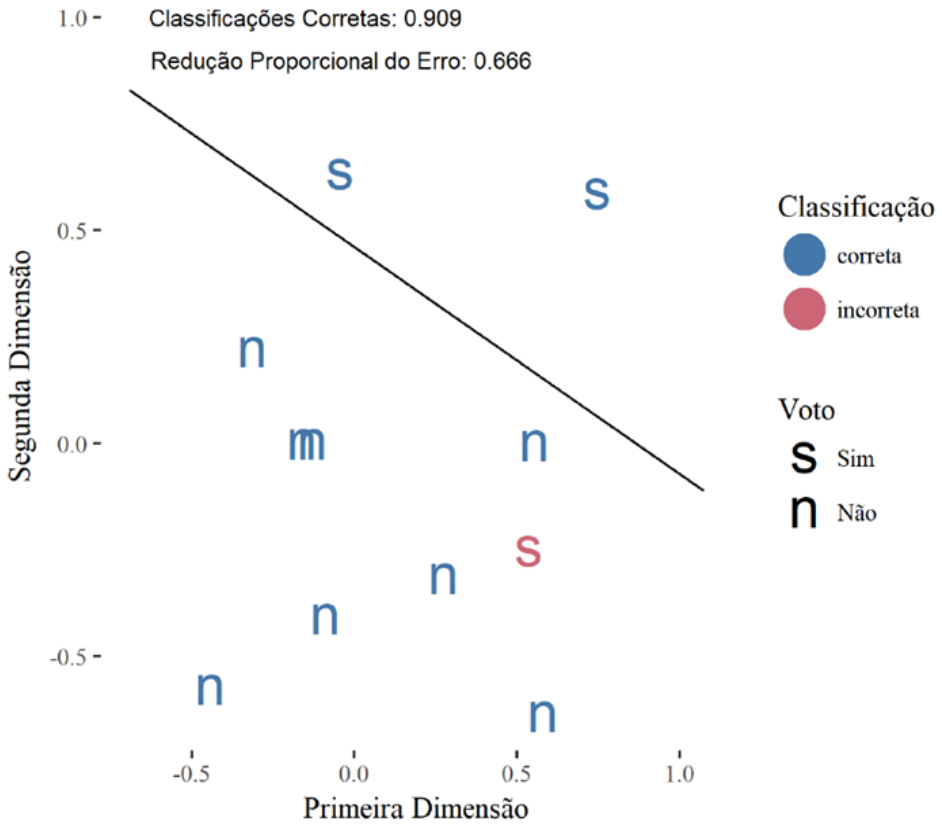


Figura 4: Exemplo para avaliação de medidas de desempenho

Votantes representados pela letra "s" votaram "sim" e votantes representados pela letra "n" votaram "não". O único votante representado pela cor vermelha foi o votante que foi classificado de forma errada pelo modelo, pois votou "sim" enquanto todos os outros representados no mesmo lado da votação votaram "não". Este deveria estar do outro lado da linha, juntamente com os outros que votaram "sim" e foram classificados da forma correta. Dessa forma, a porcentagem de classificação correta desta votação seria de 0.90922[1]. A redução proporcional do erro (RPE) seria de 0.66[2]. Tal medida seria mais adequada e criteriosa para avaliar o desempenho dos modelos. Usando o exemplo de (Armstrong et al., 2014), se em alguma votação houver uma divisão de 65-35 votos a favor de uma proposta, poderíamos classificar corretamente 65% dos votos apenas ao dizer "ingenuamente" que todos os votantes apoiarão a proposta. A redução proporcional do erro mede o quanto o modelo melhora a classificação dos votos em relação a minoria. Voltando ao nosso exemplo hipotético da figura 4, a RPE seria de 0.66 pois classifica corretamente dois dos três votantes da minoria. A RPEA seria uma medida que agrega a RPE de cada uma das votações.

É importante salientar que tal método se utiliza de votações que foram decididas de forma dividida, ou seja, não unânime. Tal critério é importante pois uma vez que se procura identificar como se dá a divergência entre votantes, a matéria julgada deve necessariamente produzir dissenso entre estes. Votações unânimes não têm como nos fornecerem informações sobre agrupamentos e divergências, dessa forma são desconsideradas. Essa característica pode afetar estudos sobre instituições que possuem grande taxa de consenso. No Supremo Tribunal Federal, por exemplo, onde a maior parte das decisões são monocráticas ou unânimes, utilizar esse método nos faria perder capacidade explicativa sobre o tribunal como um todo. No entanto, ao utilizarmos apenas decisões majoritárias, observaríamos casos que tratam de matéria controversa, sujeita então a disputas de interpretações distintas e conflitos políticos. É mais provável que casos importantes e de maior repercussão se enquadrem nesta situação.

É necessário cuidado ao analisar a dispersão dos votantes no espaço, pois o modelo não possui capacidade de inferir o que determina essa dispersão, ou seja, qual é o fator que estrutura o voto nas dimensões espaciais. Este tipo de estimação estatística foi utilizado na Ciência Política brasileira, por exemplo, para analisar a Câmara dos Deputados (Leoni, 2002; Zucco, 2009; Zucco e Lauderdale, 2011), Senado (Izumi, 2016) e Supremo Tribunal Federal (Desposato et al., 2015; Ferreira e Mueller, 2014; Mariano Silva, 2018). No caso da Câmara dos Deputados e Senado, não é a divisão ideológica que acaba sendo evidenciada com as estimações de pontos ideais, mas sim a divisão entre oposição e governo.

O primeiro modelo estatístico de estimação de pontos ideais elaborado na Ciência Política derivado diretamente da teoria espacial do voto foi o NOMINATE (Poole e Rosenthal, 1985, 1991, 1997). Este é um modelo paramétrico aplicado a votações binárias e amplamente utilizado na área de estudos legislativos. Com diversos desenvolvimentos posteriores implementados em R, o mais popular deles é o W-NOMINATE (Poole et al., 2011), enquanto o D-NOMINATE e DW-NOMINATE são utilizados para estimação de pontos ideais dinâmicos, para observar a oscilação do comportamento dos votantes ao longo do tempo.

Com o aumento de performance computacional, novas implementações de estimações de pontos ideais foram elaboradas com a aplicação de estatística bayesiana. Derivados da teoria da resposta ao item, bastante desenvolvida na área de psicometria, e desenvol-

1 número de classificações corretas / número de votantes

2 (votos na minoria - classificações erradas) / votos na minoria

vidos a partir dos trabalhos de Clinton et al. (2004) e Martin e Quinn (2002), a implementação em R do IDEAL no pacote **pscl** é um dos mais populares, junto com o pacote **MCMCpack** que conta com uma implementação do modelo dinâmico bayesiano entre outras. A teoria da resposta ao item é utilizada principalmente em estudos educacionais, onde procuram medir o conhecimento de indivíduos a partir de testes, substituídos na Ciência Política pela mensuração de ideologia a partir de votações.

Com implementações distintas, os modelos da família NOMINATE e os bayesianos acabam por ter pressupostos ligeiramente distintos. Enquanto ambos assumem que os erros são independentes, identicamente distribuídos, com distribuição normal e com menor probabilidade de ocorrer na medida em que o ponto ideal do votante é distante da reta da votação, a forma como cada um dos votantes trata as opções disponíveis ao votarem *sim* ou *não* difere. No NOMINATE, se ambas as alternativas estiverem distantes do ponto ideal do votante, este é mais indiferente as opções oferecidas, pois sua curva de preferência se dá em um formato gaussiano. Nos modelos bayesianos, de maneira distinta, a curva de preferências dos votantes assume um formato quadrático, fazendo com que mantenham a sensibilidade diante de opções distantes de seu ponto ideal. Apesar de, na prática, esse pressuposto não produzir muita diferença de resultado entre os modelos (Carroll et al., 2009), para lidar com este problema foi criado o α-NOMINATE (Carroll et al., 2013). Esta implementação bayesiana do NOMINATE foi feita para substituir as implementações anteriores, e estima se a melhor forma funcional das preferências dos votantes é gaussiana, quadrática ou uma mistura intermediária de ambas.

Neste trabalho, diante de diversas implementações estatísticas do modelo espacial para realizar estimações de pontos ideais, demonstraremos como utilizar o *Optimal Classification* (OC). Esta é uma técnica de estimação de pontos ideais não paramétrica, ao contrário dos modelos anteriores. Seu algorítimo consiste nos seguintes passos (Poole, 2005, p. 82):

1. Gerar valores iniciais para os pontos ideais dos votantes a partir da decomposição em autovalores e autovetores da matriz de concordância entre os pares de votantes

2. Dado os valores dos votantes, encontrar as estimações dos vetores das votações que maximizam o número de classificações corretas.

3. Dados os valores dos vetores de votações, encontrar novos valores dos pontos ideais dos votantes que maximizem o número de classificações corretas.

4. Voltar ao passo 2.

Tal processo se repete até que o número de classificações corretas se estabilize, maximizando o número de acertos. De acordo com Poole, tal técnica é consistente com dados com poucos votantes e poucas votações e uma de suas vantagens seria o fato de possuir menos pressupostos quando comparado aos modelos paramétricos de estimação de ponto ideal, tornando-o menos suscetível a problemas estatísticos de estimação.

Escolhemos esta técnica pois, por ser não-paramétrica, necessita que menos pressupostos estatísticos e sobre o comportamento dos votantes sejam satisfeitos. Rosenthal e Voeten (2004) argumentam que o comportamento estratégico que legislaturas podem apresentar, devido a disciplina partidária, seria menos problemático para o *Optimal Classification* quando comparado com outros modelos paramétricos[3].

3 Este problema não é plenamente resolvido por esta técnica (Spirling e McLean, 2007), pois o pressuposto de monotonicidade ainda persiste. Ver Duck-Mayr e Montgomery ([S.d.]) para um aprofundamento sobre esta questão.

Na seção seguinte, iniciaremos a explicação de como implementar uma estimação de pontos ideais, apresentando o código em R dos comandos e funções necessárias para a utilização deste método.

3. Primeiro passo: criar um objeto *rollcall*

O pacotes que implementam o *Optimal Classification*, *W-NOMINATE*, α*-NOMINATE* e *IDEAL* necessitam que criemos no R um objeto de tipo *rollcall* para que usemos suas respectivas funções. Para isso, é necessário especificar os seguintes elementos na função rollcall():

1. os dados onde cada linha é um votante e cada e coluna é uma votação

2. os valores correspondentes ao voto "sim"

3. os valores correspondentes ao voto "não"

4. os valores correspondentes a *missings*, quando o votante não votou por algum motivo

5. os valores correspondentes a quando o votante não fazia parte daquela legislatura

6. os valores correspondentes aos nomes dos legisladores, na mesma ordem em que aparecem nas linhas do banco de dados

7. os valores correspondentes aos nomes das votações, na mesma ordem em que aparecem nas colunas do banco de dados

8. os valores correspondentes a informações extras sobre os votantes, por exemplo seus partidos, na mesma ordem em que aparecem nas linhas do banco de dados

9. os valores correspondentes a informações extras sobre as votações, por exemplo suas datas, na mesma ordem em que aparecem nas colunas do banco de dados

São necessários obrigatoriamente os 3 primeiros tipos de informações, os outros são opcionais. O código abaixo ilustra o formato de um banco de dados hipotético, e a maneira de se criar um objeto *rollcall* a partir dele:

```
# Observar o banco
banco_hipotetico[1:5,1:13]

##          name      state      party V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
## 1 Legislator1     Alaska   Democrat  1  1  6  6  1  1  6  1  1   1
## 2 Legislator2      Texas   Democrat  1  1  6  1  1  6  6  6  1   1
## 3 Legislator3       Ohio Republican  1  1  6  6  1  6  6  6  6   6
## 4 Legislator4      Texas   Democrat  1  1  1  1  1  1  1  1  1   6
## 5 Legislator5 California   Democrat  6  6  6  1  6  6  6  6  6   1

# Criar um objeto roll call a partir do objeto banco_hipotetico
rollcall_hipotetico <- rollcall(banco_hipotetico[,-c(1:3)],
                                yea = 1,
                                nay = 6,
                                legis.names = banco_hipotetico$name,
                                legis.data = banco_hipotetico[,2:3])

rollcall_hipotetico

## Number of Legislators:    20
## Number of Votes:  100
##
## Using the following codes to represent roll call votes:
## Yea:      1
## Nay:      6
## Abstentions:  NA
## Not In Legislature:   9
##
## Legislator-specific variables:
## [1] “state” “party”
## Detailed information is available via the summary function.
```

Primeiro observamos as primeiras 5 linhas do banco de dados e as 13 primeiras colunas para ter uma ideia de sua estrutura e organização. A primeira coluna representa o nome de cada legislador, a segunda coluna seu estado de origem e a terceira seu partido. Daí em diante as colunas representam votações, onde o valor 1 significa voto favorável e o valor 6 significa voto contrária a matéria. Para criar um objeto de tipo *rollcall*, o primeiro argumento especificado é o banco de dados delimitando apenas as colunas de votação, portanto excluindo-se as colunas de 1 a 3. Em seguida são especificados os argumentos *yea* e *nay* com os valores que correspondem a *sim* e *não*, respectivamente 1 e 6. Depois especificamos os dois argumentos opcionais, *legis.names* indicando a coluna *name* do objeto *banco_hipotetico* como o nome dos legisladores, e as colunas 2 e 3 no argumento *legis.data*, pois estas duas colunas possuem informações extras sobre os legisladores. Caso quiséssemos nomear as votações e incluir informações extras sobre cada uma delas, especificaríamos os argumentos *vote.names* e *vote.data* com tais valores. Ao observamos o objeto criado, *rollcall_hipotetico*, podemos averiguar a estrutura dos dados, indicando o número de legisladores e votações, a codificação dos votos como especificamos, e as variáveis extras sobre os legisladores.

Para analisar o comportamento da Câmara dos Deputados durante o primeiro do governo Bolsonaro, obtemos os dados de votação utilizando o pacote de R *congressbr* (McDonnell et al., 2019), que permite fazer buscas no API[4] da Câmera dos Deputados e Senado para coletar informações diversas sobre o processo legislativo do Congresso Nacional. Quando solicitamos os dados de um determinado ano, o objeto retornado está em formato tidy[5], onde a unidade de análise

4 A sigla API refere-se a *Application Programming Interface*, são interfaces com finalidade de simplificar processos diversos, entre eles o de coletar dados organizados de maneira mais acessível.

5 Para uma explicação sobre os princípios dessa organização de dados, ver (Wickham, 2014)

é o voto de um parlamentar e uma dada votação. O código abaixo ilustra como baixar os dados para deixá-los em formato adequado para a estimação de pontos ideais e convertê-los em um objeto de tipo *rollcall*:

```
# Instalar o pacote tidyverse, necessário para manipulação de dados
install.packages("tidyverse")

# Carregar o pacote tidyverse
library(tidyverse)

# Instalar o pacote congressbr, necessário para obter os dados
# da Câmara dos Deputados
install.packages("congressbr")

# Carregar o pacote congressbr
library(congressbr)

# Instalar o pacote oc
install.packages("oc")

# Carregar o pacote oc
library(oc)

# Baixar os votos de 2019 da Câmara dos Deputados
camara_votos <- cham_votes_year(2019)

# Transformar o camara_votos em formato wide
camara_wide <-
camara_votos %>%
  transmute(legislator_name = str_to_upper(legislator_name),
            legislator_party = str_to_upper(legislator_party),
            rollcall_id, legislator_vote) %>%
  mutate(legislator_vote = recode(legislator_vote,
                                  "Sim" = 1,
                                  "Obstrucao" = 4,
                                  "Art. 17" = 5,
                                  "Nao" = 6,
                                  "Abstencao" = 7)) %>%
  mutate(legislator_id = paste0(legislator_name,
                                  " - ",
                                   legislator_party)) %>%
  pivot_wider(names_from = rollcall_id,
              values_from = legislator_vote)

# Criar um objeto de tipo rollcall
camara_rollcall <- rollcall(camara_wide[,-c(1:3)],
                            yea = 1,
                            nay = 4:7,
                            missing = NA,
                            legis.names = camara_wide$legislator_id,
                            vote.names = colnames(camara_wide)[-c(1:3)],
                            legis.data = camara_wide[,1:3])
```

Primeiro precisamos instalar e carregar os pacotes necessários para manipular e baixar os dados. O *tidyverse* é um conjunto de pacotes de R de utilização bastante difundida para ciência de dados[6]. Em seguida baixamos e instalamos o *congressbr* propriamente dito. A função cham_votes_year() serve para baixarmos todas as votações em plenário de um determinado ano, 2019 no nosso caso.

Precisamos deixar o objeto *camara_votos* em um formato adequado para passar na função *rollcall*. Para isso, precisamos deixar o banco de dados em formato *wide*, onde cada linha é um deputado e cada coluna é uma votação ou informação sobre os deputados. Primeiro selecionamos apenas as colunas que queremos manter, relativas a nome, partido e voto dos parlamentares e a coluna que identifica a votação. Em seguida recodificamos os valores que constam na coluna de votos para valores numéricos. Um passo importante é criar uma nova coluna, que chamamos de *legislador_id*, composta pelo nome do deputado e partido. Assim, caso um parlamentar mude de partido, ele será identificado como outro legislador. Isso é importante pois, como estudos anteriores mostram (Izumi, 2016; McCarty et al., 2001), a mudança partidária e entrada ou saída de parlamentares na coalizão governista faz com que haja uma mudança em seu comportamento. Caso não realizemos esse passo, assumimos que a mudança de partido do legislador não impacta em seu comportamento parlamentar. Ao fazer com que cada ponto seja um deputado em um determinado partido, podemos verificar em que medida seu comportamento se altera com a mudança de sigla. Por fim, convertemos os dados para o formato *wide*, especificando que as novas colunas terão o nome das IDs de cada votação, e seus valores serão os votos dos legisladores.

Finalmente podemos criar o objeto em formato *rollcall*. Primeiro indicamos o objeto que possui os dados, selecionando apenas as colunas que correspondem a votações. Em seguida, estabelecemos quais valores correspondem a votos *sim*, *não*[7] *e missing*. Os últimos argumentos da função se referem a coluna dos nomes dos legisladores (que no caso é a combinação do nome com o partido), o nome das votações, e as informações extras sobre os deputados.

4. Rodando o *Optimal Classification*

Com os dados obtidos e organizados da maneira adequada, podemos utilizá-los para estimar os pontos ideais. O *Optimal Classification* é implementado pelo pacote *oc*. Para rodar a função oc()[8], é necessário especificar o objeto *rollcall* que possuem os dados a serem utilizados, o número de dimensões a serem estimadas, o número mínimo de votos que um legislador precisa ter para ser mantido na análise, a proporção (de 0 a 0.5) mínima que a votação tem que ter como minoria para ser mantida na análise, e a polaridade, que corresponde ao legislador que deve possuir valor positivo em cada uma das dimensões estimadas. Antes de estimar o modelo, cabe algumas explicações sobre esses elementos.

O número de dimensões estimadas, como mencionamos anteriormente, é um problema que todo pesquisador deveria ter em mente ao utilizar qualquer técnica de estimação de pontos ideais. Portanto, cabe ao pesquisador elaborar uma justificativa teórica ou estatística para sua escolha. Uma das formas propostas por Poole (2005) de decidir o número de dimensões, que será a utilizada neste trabalho, é observando o *skree plot*. Amplamente utilizado em estudos psicométricos para avaliar dimensionalidade, a ideia deste tipo de gráfico é obter

6 Para mais detalhes, ver (Wickham et al., 2019)

7 Para simplificar, designamos todos os votos diferentes de "sim" como votos contrários

8 As implementações em R do *W-NOMINATE* e α-NOMINATE também exigem estes mesmos elementos em suas funções

alguma medida de ajuste dos dados variando o número de dimensões estimadas, e observar o momento em que se forma um "cotovelo", fazendo com que a linha fique mais horizontal e menos vertical, indicando que a partir de dada dimensão não existe uma grande melhora ao se adicionar uma dimensão a mais. Este gráfico é gerado pelos pacotes NOMINATE com os valores calculados dos autovalores da matriz de concordância dos votos entre legisladores. Com o exemplo aplicado adiante, tal ideia ficará mais clara. No entanto, tal forma de determinar o número de dimensões não deve ser aplicada sem uma reflexão do sentido substantivo de cada uma delas, pois é frequente que uma melhora na medida de ajuste dos modelos não possua significado algum, sendo resultado apenas de incorporação de ruído aos dados (Armstrong et al., 2014).

O número mínimo de votos que os legisladores devem ter para permanecer na estimação, assim como a proporção mínima que a minoria deve possuir para que dada votação permaneça na estimação, são importantes pois podem influenciar, ainda que ligeiramente por vezes, o resultado final. Votações com minoria muito pequena são pouco úteis para informar a divisão entre seus votantes, e legisladores com poucos votos proporcionam uma confiança baixa sobre a consistência de suas preferências. Tais votações e votantes podem acabar influenciando as estimações por terem características de outliers quando comparados com o conjunto completo de dados. Portanto, é comum excluí-los, mas fica a critério do pesquisador quais critérios serão estabelecidos para tal. O padrão do *Optimum Classification*, caso tais valores não sejam especificados pelos pesquisadores, é que sejam mantidos votantes com no mínimo 20 votos e votações com uma minoria mínima de 2.5%.

A polaridade da votação diz respeito a especificar um votante que possui valor positivo em cada uma das dimensões. Esta parece ser uma escolha crucial para o pesquisador mas, na realidade, ela não afeta o resultado final de maneira significativa. Por exemplo, se acreditamos que a dimensão principal representa uma distribuição ideológica dos parlamentares, e quanto maior o valor mais conservador é o político, deveríamos escolher um votante conservador como referência de polaridade. Caso especifiquemos um progressista, o resultado será apenas um espelhamento do que esperaríamos, progressistas com valores positivos e conservadores com valores negativos. Dessa forma, a posição relativa entre os pontos não se altera, mas apenas o sinal negativo ou positivo de seus valores. Caso não tenhamos certeza ou confiança em quem escolher como polaridade, podemos estimar o *Optimal Classification* com qualquer legislador e verificar o resultado a posteriori.

O código abaixo mostra como rodamos o *Optimal Classification* para a Câmara dos Deputados em 2019, a partir dos dados que obtemos anteriormente:

```
# Roda o algoritmo do Optimal Classification
camara_oc <- oc(camara_rollcall,
                dims = 2,
                polarity = rep("EDUARDO BOLSONARO - PSL", 2))
```

Especificamos o objeto *rollcall* que construímos, pedimos para estimar duas dimensões, e informamos que nas duas dimensões o deputado Eduardo Bolsonaro deve ter valor positivo. Este processo demorou cerca de 20 segundos para rodar[9].

5. Desvendando a dimensionalidade

Podemos ter uma ideia sobre os dados observando o objeto que criamos:

```
# Sumário do objeto gerado pelo Optimal Classification
summary(camara_oc)
##
##
## SUMMARY OF OPTIMAL CLASSIFICATION OBJECT
## ---------------------------
##
## Number of Legislators:    616 (113 legislators deleted)
## Number of Votes:  259 (73 votes deleted)
## Number of Dimensions:     2
## Predicted Yeas:       47153 of 50123 (94.1%) predictions correct
## Predicted Nays:       46840 of 49446 (94.7%) predictions correct
##
## The first 10 legislator estimates are:
##                                coord1D coord2D
## EDIO LOPES - PL                  0.113  -0.360
## HAROLDO CATHEDRAL - PSD          0.143  -0.091
## HIRAN GONCALVES - PP             0.079  -0.380
## JHONATAN DE JESUS - PRB          0.249  -0.075
## JOENIA WAPICHANA - REDE         -0.529   0.703
## NICOLETTI - PSL                  0.283   0.009
## OTACI NASCIMENTO - SOLIDARIED    0.075  -0.126
## SHERIDAN - PSDB                  0.219  -0.025
## ACACIO FAVACHO - PROS            0.136  -0.265
## ALINE GURGEL - PRB              -0.067   0.085
```

Algumas informações úteis são exibidas ao pedirmos um sumário do objeto criado pelo *oc*. O número de legisladores e votação mantidas e excluídas, quantas dimensões foram estimadas, o número e porcentagem de classificações corretas nos votos favoráveis e contra as matérias propostas, e os valores estimados para os 10 primeiros legisladores do banco de dados. É possível obter a porcentagem geral de classificações corretas dos votos e a RPEA com o seguinte comando:

9 Esta é uma vantagem quando comparado com o tempo de estimação de modelos bayesianos, que podem demorar um tempo consideravelmente maior

```
# Obter as medidas de desempenho
camara_oc$fits
## [1] 0.9439986 0.7446653
```

O primeiro valor observado é a porcentagem de classificações correta, seguido da RPEA. Os valores obtidos são relativamente altos quando comparados com os valores apresentados por outros estudos (Leoni, 2002). Cabe observar que quando estimamos apenas uma dimensão, a porcentagem de classificações corretas cai para 92.8% e a RPEA para 0.67. A diferença não é muito grande, nos dando poucos elementos para crer que exista uma segunda dimensão relevante. Uma outra forma de se verificar a relevância de se adicionar mais dimensões é observar o *skree plot* com os autovalores gerados pelo *Optimal Classification*. A lógica desse tipo de gráfico, como mencionado anteriormente, é identificar em qual dimensão se forma um "cotovelo", e considerar apenas o número de dimensões anterior a este ponto. Podemos ver o gráfico com o seguinte código:

```
# Fazer o screeplot
plot.OCskree(camara_oc)
```

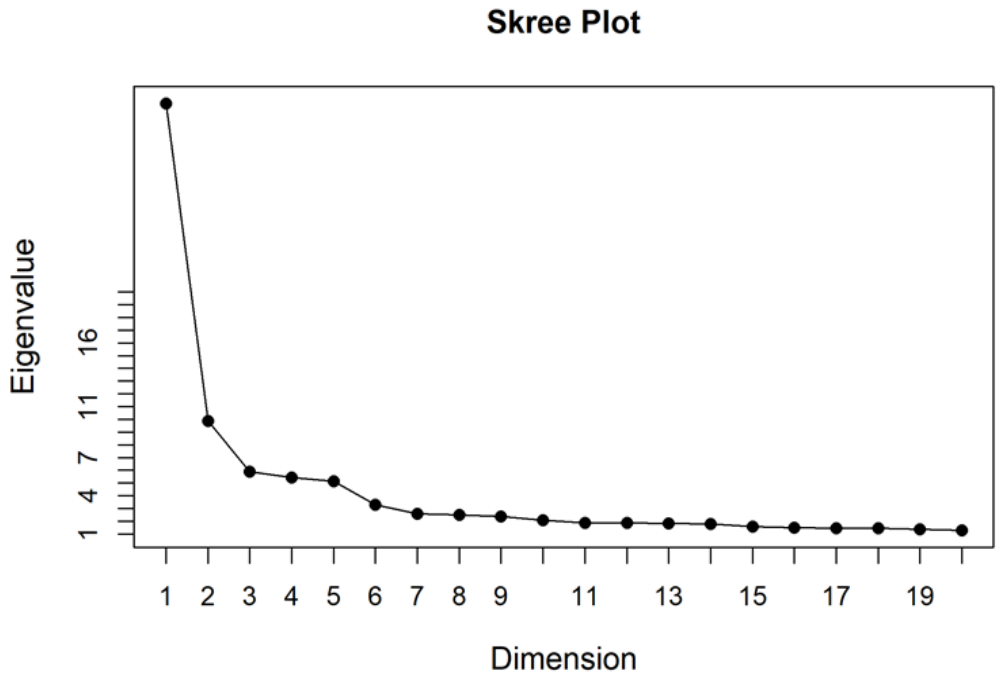


Gráfico 1: Screeplot para avaliação de dimensionalidade

Observamos que na estrutura dos dados predomina uma única dimensão, com uma segunda muito discreta. Esse resultado vai no mesmo sentido de estudos anteriores que identificaram uma única dimensão como relevante na estruturação dos votos da Câmara dos Deputados (Leoni, 2002) e do Senado (Izumi, 2016). Tais estudos também afirmam que essa dimensão diz respeito a divisão entre governo e oposição. Como mencionado anteriormente, a interpretação do significado das dimensões estimadas deve ser feita pelo pesquisador, estas não são dadas ou informadas pelas técnicas de estimação de pontos ideais.

Podemos verificar se em 2019, no primeiro ano de um governo que não estabeleceu um governo de coalizão nos moldes anteriores, essa dimensão permanece caracterizada pelo conflito entre governo e oposição. O gráfico abaixo mostra a correlação entre os pontos estimados para a primeira dimensão e a correlação do voto dos deputados com a posição que o líder de governo orientou[10].

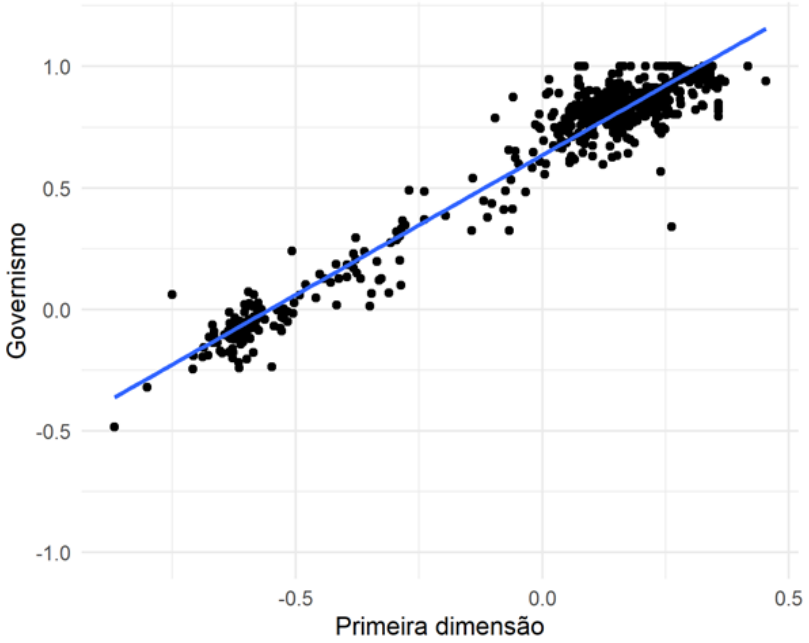


Gráfico 2: Correlação entre pontos ideais e governismo

O gráfico mostra uma alta correlação entre os pontos estimados da primeira dimensão e o voto dos deputados na mesma direção em que o líder do governo indicou. Essa associação é de 0.97. Ou seja, os deputados que possuem costumam ter o voto na mesma direção que o lider do governo estão localizados mais a direita no gráfico, com maiores valores em seus pontos ideais. Portanto, mesmo não havendo um presidencialismo de coalizão nos mesmos moldes que os governos anteriores, a dimensão do conflito político da Câmara dos Deputados se dá de maneira semelhante.

6. Distribuição dos pontos ideais

Os valores dos pontos ideais de cada legislador podem ser verificados no objeto que foi criado pelo *Optimal Classification*, em uma lista chamada *legislators*. Nela consta para cada votante a quantidade de votos corretamente classificados, as coordenadas estimadas e as variáveis sobre os legisladores que especificamos no argumento *legis.data* na função `rollcall()`.

10 Excluímos as votações nas quais o líder do governo liberou a bancada

```
glimpse(camara_oc$legislators)
## Rows: 729
## Columns: 11
## $ legislator_name  <chr> "EDIO LOPES", "HAROLDO CATHEDRAL", "HIRAN GONCALVE...
## $ legislator_party <chr> "PL", "PSD", "PP", "PRB", "REDE", "PSL", "SOLIDARI...
## $ legislator_id    <chr> "EDIO LOPES - PL", "HAROLDO CATHEDRAL - PSD", "HIR...
## $ rank             <dbl> 282.0, 333.0, 220.0, 511.0, 95.0, 548.0, 212.0, 46...
## $ correctYea       <dbl> 69, 106, 79, 44, 67, 104, 100, 105, 95, 33, 97, 80...
## $ wrongYea         <dbl> 1, 8, 4, 4, 6, 0, 5, 2, 2, 5, 0, 1, 4, 0, 5, 0, 0,...
## $ wrongNay         <dbl> 1, 3, 0, 1, 20, 1, 4, 5, 4, 10, 1, 4, 4, 4, 6, 0, ...
## $ correctNay       <dbl> 47, 87, 71, 34, 91, 90, 82, 77, 83, 44, 83, 125, 6...
## $ volume           <dbl> 0.311, 0.019, 0.068, 0.006, 0.003, 0.070, 0.020, 0...
## $ coord1D          <dbl> 0.11288447, 0.14326323, 0.07852262, 0.24926999, -0...
## $ coord2D          <dbl> -3.600102e-01, -9.135407e-02, -3.802511e-01, -7.45...
```

Com essas informações, podemos elaborar um gráfico para mostrar os pontos ideais dos parlamentares. O código a seguir mostra como fazer um gráfico aproveitando a variável partidária para analisar separadamente o comportamento dos partidos na primeira dimensão, utilizando a variável *coord1D*:

```
camara_oc$legislators %>%
  filter(legislator_party %in% c("PSL", "NOVO", "PT", "PSOL",
                                 "PCDOB", "PSB", "PDT", "PSD",
                                 "PSDB", "MDB", "PP", "DEM")) %>%
ggplot(aes(coord1D)) +
  geom_density() +
  facet_wrap(~legislator_party, drop = T, ncol = 3, scales = "free_y") +
  geom_vline(aes(xintercept = mean(camara_oc$legislators$coord1D,
                                   na.rm = T))) +
  theme_minimal() +
  labs(x = "Primeira dimensão", y = "Densidade") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

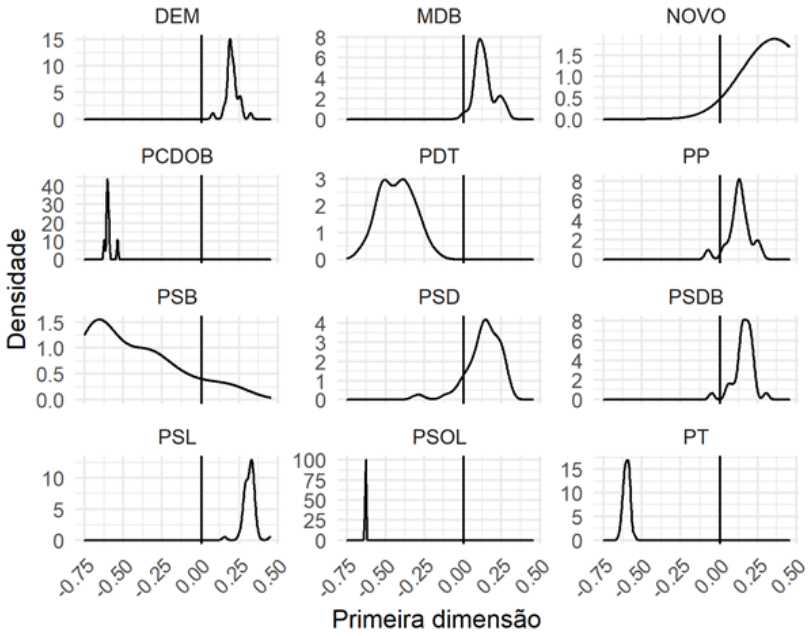


Gráfico 3: Distribuição partidária na primeira dimensão

Optamos por manter no gráfico apenas 12 partidos pois, caso representássemos as 33 diferentes agremiações e aqueles sem partido, o gráfico seria pouco útil para distinguir o comportamento partidário. Representamos também com uma linha vertical o valor médio dos pontos obtidos por todos os parlamentares. Dessa forma é possível verificar partidos com diferentes comportamentos de acordo com a dispersão de seus pontos ideais ao longo da primeira dimensão. PT, PSOL e PC do B concentrados na esquerda com pouca dispersão. PSB e PDT concentrados a esquerda da média, mas bastante dispersos. Os outros partidos, PSL, NOVO, DEM, MDB, PP, PSD e PSDB se concentram na direita.

Além da dispersão dos parlamentares em uma única dimensão, podemos visualizar a dispersão de seus pontos nas duas dimensões estimadas:

```
ggplot(camara_oc$legislators, aes(coord1D, coord2D)) +
  geom_point(alpha = 0.5) +
  theme_minimal() +
  labs(x = "Primeira dimensão", y = "Segunda dimensão")
```

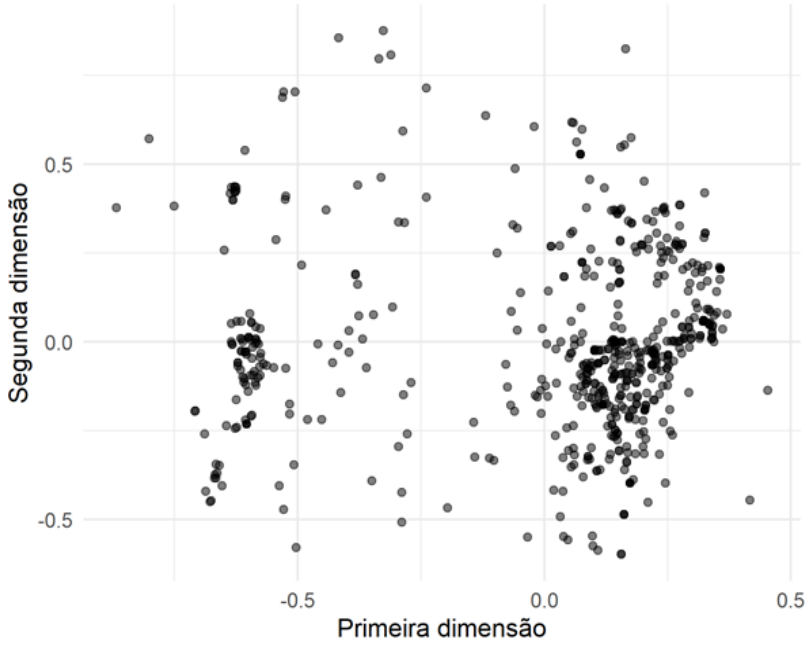


Gráfico 4: Distribuição dos pontos ideais nas duas dimensões

7. Distribuição das retas das votações

Mencionamos anteriormente que as estimações de pontos ideais estimam tanto os pontos dos legisladores quanto as retas das votações que os dividem. Para isso, usamos as informações que constam na lista *rollcalls* que compõe o objeto criado pelo *Optimal Classification*. Podemos verificar, em cada votação, quantos votos foram classificados de maneira correta e incorreta, a redução proporcional do erro, e as informações necessárias para determinarmos as retas que dividem os parlamentares.

```
head(camara_oc$rollcalls)
##      correctYea wrongYea wrongNay correctNay       PRE normVector1D
## [1,]         NA       NA       NA         NA        NA           NA
## [2,]         NA       NA       NA         NA        NA           NA
## [3,]        297       12        1         72 0.8452381    0.9985689
## [4,]         94        4       36        216 0.6923077    0.9254343
## [5,]        121       31       13        244 0.6716418    0.2525467
## [6,]        212       26       20        112 0.6666667    0.2703514
##      normVector2D   midpoints
## [1,]           NA          NA
## [2,]           NA          NA
## [3,]  -0.05347965  0.23146515
## [4,]  -0.37890820 -0.06372496
## [5,]  -0.96758471  0.08072122
## [6,]  -0.96276172  0.08239360
```

A análise de votações específicas, de forma qualitativa, também é uma forma importante para termos evidências do significado substantivo das dimensões estimadas. Não só podemos identificar as coalizões mais frequentes por meio dessa análise, como também podemos saber o ângulo dessas retas e determinar quais delas são predominantemente pertencentes a primeira dimensão, quais possuem influência de uma segunda dimensão e quais não são bem explicadas por nenhuma das dimensões. Para acrescentar as linhas das votações ao gráfico anterior, é necessário o seguinte código:

```
ggplot() +
  geom_point(data = camara_oc$legislators,
             aes(coord1D, coord2D)) +
  geom_segment(data = data.frame(camara_oc$rollcall),
               aes(x = (midpoints*normVector1D)+normVector2D,
                   y = (midpoints*normVector2D)-normVector1D,
                   xend = (midpoints*normVector1D)-normVector2D,
                   yend = (midpoints*normVector2D)+normVector1D),
               color = "grey", alpha = 0.4) +
  coord_cartesian(xlim = c(-1,1), ylim = c(-1,1)) +
  theme_minimal() +
  labs(x = "Primeira dimensão", y = "Segunda dimensão")
```

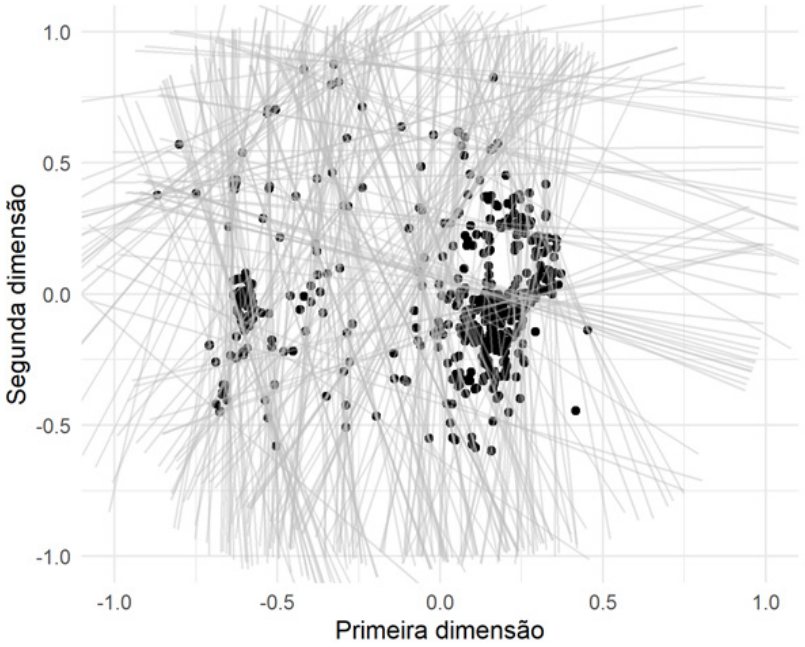


Gráfico 5: Distribuição dos pontos ideais nas duas dimensões e as linhas das votações

Observe que na função geom_point() especificamos os valores dos pontos ideais, que está na lista *legislators* do objeto criado pelo *Optimal Classification*. Na função geom_segment() especificamos os valores necessários para o cálculo dos pontos iniciais e finais das retas nas coordenadas *x* e *y*, utilizando as informações das votações que está na lista *rollcall* do mesmo objeto criado pelo *Optimal Classification*.

Com o gráfico, temos a impressão que a maioria das linhas possuem uma inclinação mais vertical, cruzando o eixo *x* do gráfico, indicando o predomínio da primeira dimensão na estruturação do comportamento dos votantes. No entanto, podemos analisar com um pouco mais de profundidade votações específicas para avaliar quais delas são bem explicadas pela estimação de pontos ideais. A primeira forma de fazer uma seleção das votações, é observar quais tiveram melhor redução proporcional do erro. Essa informação está disponível na lista *legislators* do objeto criado pelo *Optimal Classification*. Com o gráfico abaixo, podemos observar a distribuição desse valor em todas as votações:

```
ggplot(data.frame(camara_oc$rollcalls), aes(PRE)) +
  geom_histogram() +
  geom_vline(aes(xintercept = median(PRE, na.rm = T))) +
  theme_minimal() +
  labs(x = "RPE", y = "Número de votações")
```

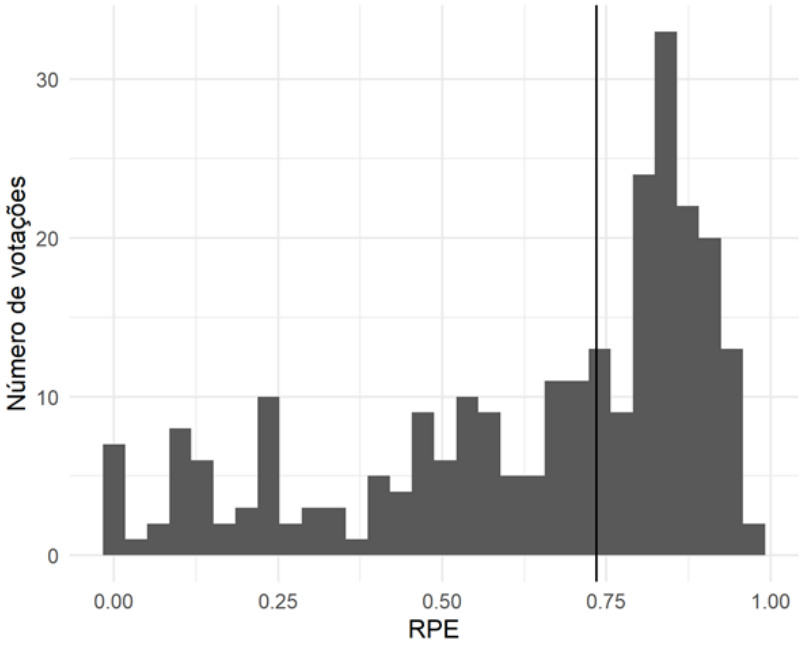


Gráfico 6: Distribuição da redução proporcional do erro

O valor mediano da redução proporcional do erro é próximo de 0.75. Portanto um primeiro critério para averiguar o sentido substantivo da primeira dimensão seria observar as votações com RPE maior do que 0.75. Essa informação sozinha não nos ajuda a saber quais votações dizem respeito a primeira ou segunda dimensão, ou ainda a ambas, mas sim quais votações são bem explicadas pela estimação como um todo. Uma forma de acessarmos quais delas correspondem a primeira dimensão é utilizamos as mesmas variáveis utilizadas para estimar as retas que dividem os parlamentares nas votações. Fazemos isso da seguinte forma:

```
x1 <- (camara_oc$rollcalls[,"midpoints"]*camara_oc$rollcalls[,"normVector1D"])+
  camara_oc$rollcalls[,"normVector2D"]

x2 <- (camara_oc$rollcalls[,"midpoints"]*camara_oc$rollcalls[,"normVector1D"])-
  camara_oc$rollcalls[,"normVector2D"]

x_diff <- abs(x2 - x1)

ggplot(NULL, aes(x_diff)) +
  geom_histogram() +
  geom_vline(aes(xintercept = median(x_diff, na.rm = T))) +
  theme_minimal() +
  labs(y = "Número de votações", x = "Medida de inclinação das retas")
```

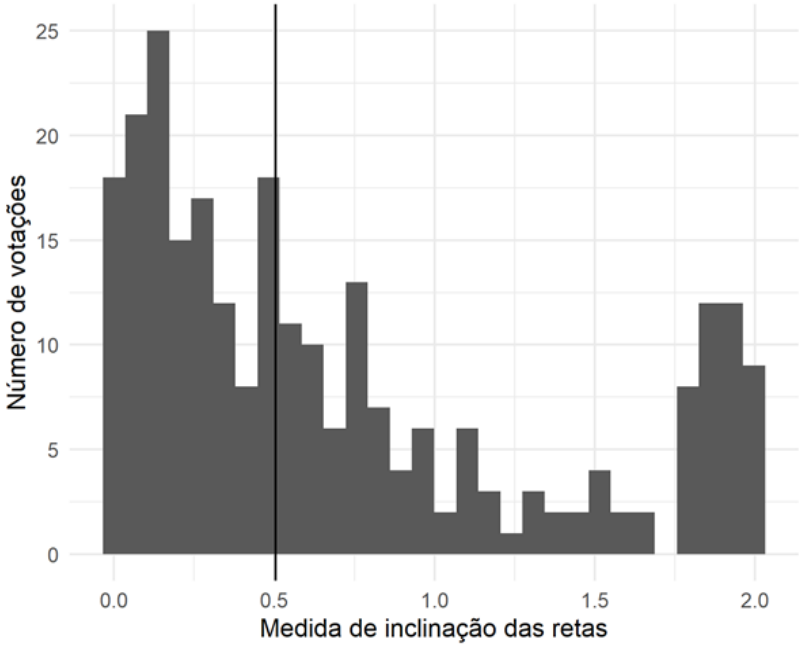


Gráfico 7: Medida de inclinação das retas

A ideia por trás do que fizemos foi: 1) calcular qual é o valor no eixo x, que corresponde a primeira dimensão, do ponto inicial da reta que divide os parlamentares; 2) fazer o mesmo para o ponto final; 3) saber o valor absoluto da diferença entre os dois valores. Se a diferença for baixa, quer dizer que a linha tende a ter uma posição mais vertical, sendo melhor explicada pela primeira dimensão. Com o gráfico, verificamos que a mediana dessa diferença é muito próxima de 0.5. Portanto valores inferiores a este seria um segundo indicador de predominância da primeira dimensão. Por fim, juntamos os dois critérios para selecionar as votações:

```
votacoes_selecionadas <- which(camara_oc$rollcalls[,"PRE"] > 0.75 &
                                 x_diff < 0.5)

colnames(camara_rollcall$votes)[votacoes_selecionadas]
##  [1] "PL-1292-1995-3"  "PL-1292-1995-7"  "PL-4742-2001-1"  "PLP-441-2017-7"
##  [5] "PLP-441-2017-9"  "PLP-441-2017-11" "MPV-852-2018-1"  "MPV-866-2018-1"
##  [9] "MPV-867-2018-3"  "MPV-867-2018-19" "MPV-871-2019-3"  "MPV-871-2019-4"
## [13] "MPV-871-2019-5"  "PEC-6-2019-2"    "PEC-6-2019-3"    "PEC-6-2019-4"
## [17] "PEC-6-2019-5"    "PEC-6-2019-6"    "PEC-6-2019-7"    "PEC-6-2019-8"
## [21] "PEC-6-2019-9"    "PEC-6-2019-13"   "PEC-6-2019-14"   "PEC-6-2019-18"
## [25] "PEC-6-2019-22"   "PEC-6-2019-24"   "PEC-6-2019-31"   "PEC-6-2019-32"
## [29] "PEC-6-2019-33"   "PEC-6-2019-35"   "PEC-6-2019-37"   "PEC-6-2019-38"
## [33] "PEC-6-2019-39"   "PEC-6-2019-40"   "PEC-6-2019-41"   "PEC-6-2019-42"
## [37] "PEC-34-2019-4"   "MPV-881-2019-1"  "MPV-881-2019-5"  "MPV-881-2019-6"
## [41] "MPV-881-2019-7"  "MPV-881-2019-8"  "MPV-881-2019-9"  "MPV-881-2019-11"
## [45] "PL-2787-2019-1"  "PL-2788-2019-2"  "REQ-1464-2019-1" "PL-2999-2019-6"
## [49] "PL-3261-2019-5"  "PL-3261-2019-6"  "PL-3261-2019-7"  "PL-3261-2019-9"
## [53] "PL-3261-2019-10" "PL-3261-2019-12" "PL-3261-2019-13" "MPV-886-2019-1"
## [57] "MPV-886-2019-2"  "PL-3715-2019-2"  "PL-3715-2019-9"  "MPV-890-2019-3"
## [61] "MPV-890-2019-5"  "MPV-890-2019-6"  "MPV-890-2019-7"  "MPV-890-2019-9"
## [65] "PL-4162-2019-1"  "PL-4162-2019-2"  "REQ-2110-2019-1" "REQ-2157-2019-1"
## [69] "REQ-2234-2019-1" "PDL-523-2019-1"  "PDL-523-2019-4"  "PL-5815-2019-1"
## [73] "PL-5815-2019-2"  "PL-5815-2019-3"
```

Nos códigos acima, verificamos quais votações atendem os dois critérios que estabelecemos e visualizamos os nomes destas votações. Podemos observar que diversas votações que dizem respeito a PEC 6/2019, sobre a reforma da previdência, correspondem a primeira dimensão. Podemos checar se essas linhas realmente se comportam da maneira esperada, ou seja, se são verticais com pouca inclinação:

```
ggplot() +
  geom_point(data = camara_oc$legislators,
             aes(coord1D, coord2D)) +
  geom_segment(data = data.frame(camara_oc$rollcall[votacoes_selecionadas,]),
               aes(x = (midpoints*normVector1D)+normVector2D,
                   y = (midpoints*normVector2D)-normVector1D,
                   xend = (midpoints*normVector1D)-normVector2D,
                   yend = (midpoints*normVector2D)+normVector1D),
               color = "grey", alpha = 0.5) +
  theme_minimal() +
  labs(x = "Primeira dimensão", y = "Segunda dimensão")
```

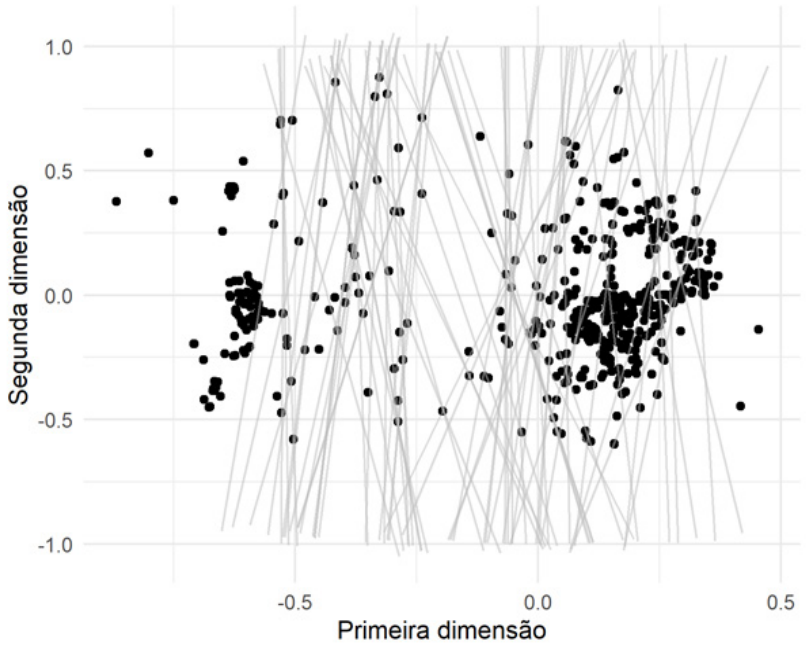


Gráfico 8: Pontos ideais e retas das votações correspondentes a primeira dimensão

Note que a diferença está na especificação da função *geom_segment()*, onde indicamos no objeto *camara_oc$rollcall* que queremos apenas as linhas que correspondem as votações selecionadas. Nosso objetivo foi bem sucedido em selecionar votações bem explicadas pela técnica e que diz respeito a primeira dimensão. Podemos notar que existem votações que acabam dividindo tanto o bloco de parlamentares localizados a esquerda quanto o bloco de deputados localizados a direita.

Com as informações obtidas sobre os pontos ideais de cada votante e as retas de cada votação, é possível analisar mais qualitativamente decisões específicas. Dessa forma é possível se aprofundar em como se dá a estrutura das preferências e do conflito político entre os deputados, contribuindo ainda mais para interpretação do significado substantivo das dimensões estimadas e dos resultados de forma geral. Abaixo mostramos um exemplo com a representação de uma das votações da PEC da reforma da previdência.

```
camara_oc$legislators %>%
  mutate(Voto = camara_rollcall$votes[,147]) %>%
  mutate(Voto = case_when(Voto == 1 ~ "Sim",
                          Voto != 1 ~ "Não")) %>%
  filter(!is.na(Voto)) %>%
ggplot() +
  geom_point(aes(coord1D, coord2D)) +
  geom_segment(data = data.frame(camara_oc$rollcall) %>%
                 slice(147),
               aes(x = (midpoints*normVector1D)+normVector2D,
                   y = (midpoints*normVector2D)-normVector1D,
                   xend = (midpoints*normVector1D)-normVector2D,
                   yend = (midpoints*normVector2D)+normVector1D),
               color = "grey", alpha = 0.5) +
  theme_minimal() +
  facet_wrap(vars(Voto), nrow = 1) +
  labs(x = "Primeira dimensão", y = "Segunda dimensão")
```

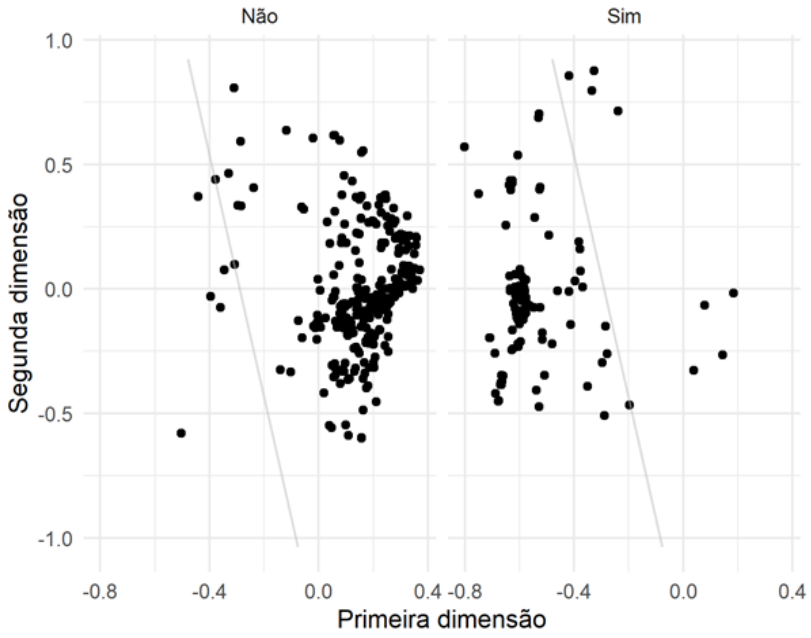


Gráfico 9: Pontos ideais e reta da votação de uma das votações da PEC 6/2019

No código acima, utilizamos o objeto que criamos com o *Optimal Classification* e adicionamos a coluna referente aos votos da coluna 147, que se refere a votação selecionada sobre a reforma da previdência. Em seguida recodificamos os valores para indicar se o voto foi favorável ou contrário a matéria, removendo aqueles que não votaram, e geramos o gráfico. Especificamos que sejam feitos dois painéis separados, de acordo com o voto, e mantemos apenas a reta da votação selecionada. Com o gráfico, podemos visualizar que os governistas, que neste caso votaram contra a proposta, se concentram ao lado direito, com maiores valores de ponto ideal na primeira dimensão, enquanto a oposição votou a favor da matéria. Podemos observar que a estimação erra 5 deputados que votaram *não* e 8 deputados que votaram *sim*, por estarem localizados do lado oposto da reta ao que seria esperado. Essa votação tem 97.1% dos votos corretamente classificados e uma RPE de 0.88. Com este gráfico é possível analisar qualitativamente esta votação específica e votantes específicos. Um exemplo é o com-

portamento da deputada Tabata Amaral. No contexto da reforma da previdência, houve muita crítica, tanto nas redes sociais quanto internamente no PDT, sobre alguns do votos que a deputada fez a favor do texto que foi aprovado. O ponto ideal dela é o que votou *não* localizado mais acima do gráfico. Como observamos, o voto é corretamente classificado pela estimação, colocando ao lado direito da reta que divide a votação. Dessa forma, o voto dela é coerente com seu comportamento em outras votações, fazendo-a ocupar este ponto no espaço. Neste caso, mesmo localizada mais próxima a esquerda, seu voto acompanhando aqueles localizados a direita não seria atípico dada a forma como o conflito político se estrutura na Câmara dos Deputados ao analisarmos espacialmente todos os deputados e votações.

Outra possibilidade de análise do comportamento dos votantes é observar quais deles são mais "imprevisíveis", com maior número de classificações incorretas pois seu comportamento não seguiria a lógica das dimensões encontradas, mas por outras não estimadas. Existem modelos específicos para lidar com tais tipos de votantes de forma mais adequada (Lauderdale, 2010). No objeto criado pelo *Optimal Classification*, são apresentadas colunas com o resultado de quantos *sim* e *não* foram classificados de maneira correta e incorreta. Dessa forma, é possível saber a porcentagem de votos de cada um dos votantes que a estimação de pontos ideais conseguiu alocar de maneira adequada no espaço. Podemos obter os resultados e elaborar um gráfico da seguinte forma:

```
camara_oc$legislators %>%
  mutate(classificacoes_corretas = 100 * (correctYea + correctNay) /
           (correctYea + wrongYea + wrongNay + correctNay)) %>%
  ggplot(aes(classificacoes_corretas)) +
  geom_histogram() +
  geom_vline(aes(xintercept = mean(classificacoes_corretas, na.rm = T))) +
  theme_minimal() +
  labs(x = "Porcentagem de classificações corretas", y = "Quantidade")
```

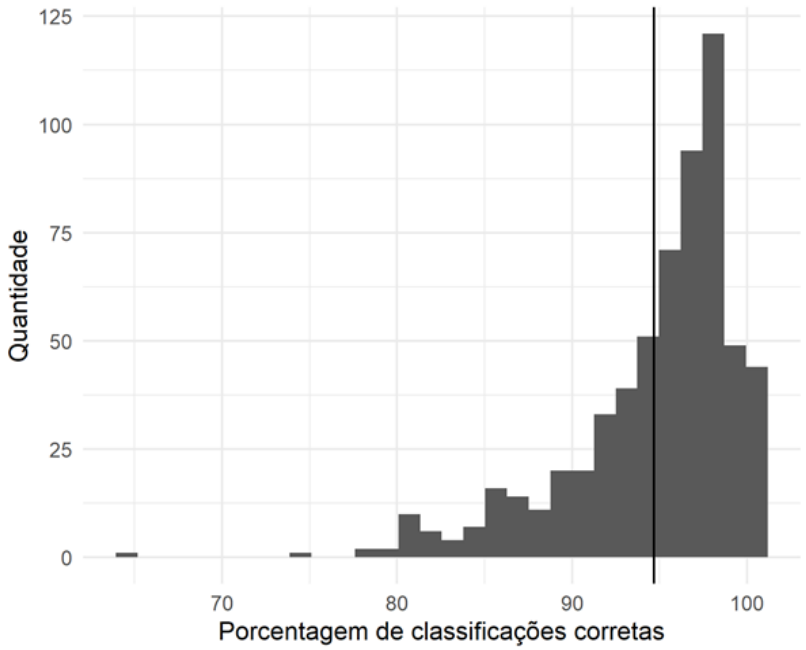


Gráfico 10: Distribuição da porcentagem de classificações corretas dos votantes

Com o histograma, observamos que a média de classificações corretas dos votantes é de 94.6% e a grande maioria dos deputados está acima dos 90%. Para observar quais são os votantes cujos comportamentos são menos explicados pelas duas dimensões estimadas, utilizamos o seguinte código:

```
camara_oc$legislators %>%
  mutate(classificacoes_corretas = 100 * (correctYea + correctNay) /
           (correctYea + wrongYea + wrongNay + correctNay)) %>%
  arrange(classificacoes_corretas) %>%
  slice(1:20) %>%
  mutate(legislator_id = fct_reorder(legislator_id, classificacoes_corretas)) %>%
  ggplot(aes(classificacoes_corretas, legislator_id)) +
  geom_point() +
  theme_minimal() +
  labs(x = "Classificações corretas (%)", y = "Votante")
```

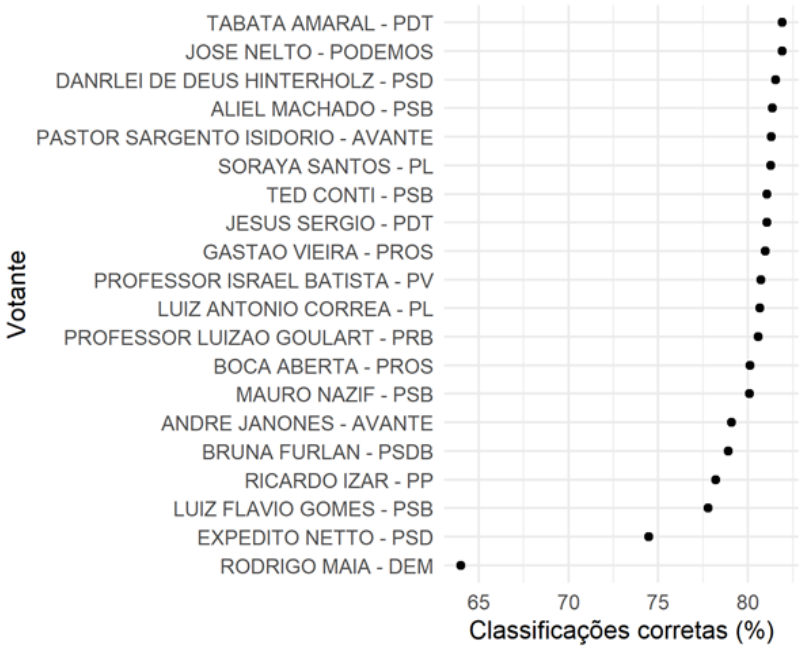


Gráfico 11: Votantes com menor porcentagem de classificações corretas

Mantemos no gráfico apenas os 20 votantes com menor porcentagem de votos corretamente classificados. Observamos que Rodrigo Maia é o que possui menor porcentagem, com menos de 65%, seguido por Expedito Netto com pouco menos que 75%. Curioso notar que a deputada Tabata Amaral, mencionada anteriormente pelas críticas ao seu comportamento, está entre aqueles cujo comportamento não é tão bem explicado pelas duas dimensões estimadas.

8. Considerações finais

O presente trabalho apresentou como estimar pontos ideais e algumas formas de se explorar os resultados de uma das diversas aplicações que a técnica pode ter. Muitos outros pacotes de R foram implementados para explicar o comportamento de atores políticos diversos com tipos de dados variados.

Em estudos que trabalham com textos, ao invés de votos em propostas legislativas, a escolha dos atores estão no uso de palavras. Ao tratarem sobre um determinado tópico, usar certas palavras e omitir outras revelaria uma preferência ideológica sobre como abordar uma questão. Ainda mais, a própria prevalência de certos tópicos em detrimento de outros em um discurso político revelaria um conteúdo ideológico. Implementações como *Wordfish* (Slapin e Proksch, 2008), *Wordscores* (Laver et al., 2003) e *Wordshoal* (Lauderdale e Clark, 2014) procuram lidar com esse tipo de questão, analisando programas partidários, discursos parlamentares, entre outros.

A relativa facilidade de se obter dados do Twitter também tem possibilitado a utilização desse tipo de dado para estimação de pontos ideais. Não apenas o texto dos tweets feitos pelos usuários podem ser analisados com as implementações que acabamos de mencionar, como também a maneira como os usuários interagem e se conectam uns com os outros. O ato de escolher seguir determinadas contas em detrimento de outras relaria uma preferência ideológica dos usuários da rede social em obter certos tipos de conteúdos (Barberá, 2015).

Diversos pacotes de R que surgem a cada dia, além de expandir quais tipos de dados podem ser utilizados para a estimação de pontos ideais, também procuram expandir as possibilidades de uma melhor inferência estatística a partir de dados já utilizados a tempos. A comparação intertemporal de uma mesma instituição, por exemplo, não é trivial. Não podemos estimar pontos ideais, por exemplo, da Câmara dos Deputados em 2019 e comparar diretamente os resultados com uma estimação relativa ao ano anterior. Para isso são necessários pressupostos teóricos e estatísticos que são implementados por outros algorítimos (Bailey et al., 2017; Martin e Quinn, 2002; Poole e Rosenthal, 2001). Além da comparação intertemporal, é possível fazer uma comparação de pontos ideais entre instituições e atores distintos, com os dados e métodos adequados (Bailey, 2007; Bonica, 2014; Tausanovitch e Warshaw, 2013).

REFERÊNCIAS BIBLIOGRÁFICAS

ARMSTRONG, David A et al. (2014), *Analyzing spatial models of choice and judgment with R*. [S.l.] CRC Press.

BAILEY, Michael A. (2007), "Comparable preference estimates across time and institutions for the court, congress, and presidency". *American Journal of Political Science*, vol. 51,no. 3: 433–448.

BAILEY, Michael A & STREZHNEV, Anton & VOETEN, Erik. (2017), "Estimating dynamic state preferences from United Nations voting data". *Journal of Conflict Resolution*, vol. 61,no. 2: 430–456.

BARBERÁ, Pablo. (2015), "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". *Political analysis*, vol. 23,no. 1: 76–91.

BONICA, Adam. (2014), "Mapping the ideological marketplace". *American Journal of Political Science*, vol. 58,no. 2: 367–386.

CARROLL, Royce et al. (2009), "Comparing NOMINATE and IDEAL: Points of difference and Monte Carlo tests". *Legislative Studies Quarterly*, vol. 34,no. 4: 555–591.

CARROLL, Royce et al. (2013), "The structure of utility in spatial models of voting". *American Journal of Political Science*, vol. 57,no. 4: 1008–1028.

CLINTON, Joshua & JACKMAN, Simon & RIVERS, Douglas. (2004), "The statistical analysis of roll call data". *American Political Science Review*, vol. 98,no. 2: 355–370.

CONVERSE, Philip E. (1964), "The nature of belief systems in mass publics. Ideology and discontent". *Ideology and Discontent*, 206–261.

DESPOSATO, Scott W. & INGRAM, Matthew C. & LANNES JR., Osmar P. (2015), "Power, Composition, and Decision Making: The Behavioral Consequences of Institutional Reform on Brazil's Supremo Tribunal Federal". *Journal of Law, Economics, and Organization*, vol. 31,no. 3: 534–567.

DOWNS, Anthony. (1957), *An economic theory of democracy*. New York, Harper; Row.

DUCK-MAYR, JBrandon & MONTGOMERY, Jacob M. *Ends Against the Middle: Scaling Votes When Ideological Opposites*. [S.d.].

FERREIRA, Pedro Fernando Almeida Nery & MUELLER, Bernardo. (2014), "How judges think in the Brazilian Supreme Court: Estimating ideal points and identifying dimensions". *Economia*, vol. 15,no. 3: 275–293.

IZUMI, Mauricio. (2016), "Governo e Oposição no Senado Brasileiro (1989-2010)". *Dados*, vol. 59,no. 1: 91–138.

LAUDERDALE, Benjamin. (2010), "Unpredictable voters in ideal point estimation". *Political Analysis*, no. 18: 151–171.

LAUDERDALE, Benjamin E. & CLARK, Tom S. (2014), "Scaling politically meaningful dimensions using texts and votes". *American Journal of Political Science*, vol. 58,no. 3: 754–771.

LAVER, Michael & BENOIT, Kenneth & GARRY, John. (2003), "Extracting policy positions from political texts using words as data". *American Political Science Review*, vol. 97,no. 02: 311–331.

LEONI, Eduardo. (2002), "Ideologia, Democracia e Comportamento Parlamentar: A Camara dos Deputados (1991-1998)". *Dados*, vol. 45,no. 3: 361–386.

LIMONGI, Fernando. (2019), "Presidencialismo do Desleixo". *Piauí*, no. 158: Disponível em: <https://piaui.folha.uol.com.br/materia/presidencialismo-do-desleixo/>.

MARIANO SILVA, Jeferson. (2018), "Mapeando o Supremo. As posições dos ministros do STF na jurisdição constitucional (2012-2017)". *Novos Estudos - CEBRAP*, vol. 37,no. 1: 35–54.

MARTIN, Andrew D. & QUINN, Kevin M. (2002), "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis*, vol. 10,no. 2: 134–153.

MCCARTY, Nolan & POOLE, Keith T. & ROSENTHAL, Howard. (2001), "The Hunt for PartyDiscipline in Congress". *American Political Science Review*, vol. 95,no. 3: 673–87.

MCDONNELL, Robert Myles & DUARTE, Guilherme Jardim & FREIRE, Danilo. (2019), "congressbr: An R Package for Analyzing Data from Brazil's Chamber of Deputies and Federal Senate". *Latin American Research Review*, vol. 54,no. 4:

POOLE, Keith et al. (2011), "Scaling Roll Call Votes with wnominate in R". *Journal of Statistical Software*, vol. 42,no. 14: 1–21.

POOLE, Keith T. (2005), *Spatial models of parliamentary voting*. [S.l.] Cambridge University Press.

POOLE, Keith T & ROSENTHAL, Howard. (1985), "A spatial model for legislative roll call analysis". *American Journal of Political Science*, 357–384.

POOLE, Keith T & ROSENTHAL, Howard. (1997), *Congress: A political-economic history of roll call voting*. [S.l.] Oxford University Press on Demand.

POOLE, Keith T & ROSENTHAL, Howard. (2001), "D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting". *Legislative Studies Quarterly*, 5–29.

POOLE, Keith T & ROSENTHAL, Howard. (1991), "Patterns of congressional voting". *American Journal of Political Science*, 228–278.

ROSENTHAL, Howard & VOETEN, Erik. (2004), "Analyzing roll calls with perfect spatial voting: France 1946–1958". *American Journal of Political Science*, vol. 48,no. 3: 620–632.

SLAPIN, Jonathan B & PROKSCH, Sven-Oliver. (2008), "A scaling model for estimating time-series party positions from texts". *American Journal of Political Science*, vol. 52,no. 3: 705–722.

SPIRLING, Arthur & MCLEAN, Iain. (2007), "UK OC OK? Interpreting optimal classification scores for the UK House of Commons". *Political Analysis*, 85–96.

TAUSANOVITCH, Chris & WARSHAW, Christopher. (2013), "Measuring constituent policy preferences in congress, state legislatures, and cities". *The Journal of Politics*, vol. 75,no. 2: 330–342.

WICKHAM, Hadley. (2014), "Tidy data". *Journal of Statistical Software*, vol. 59,no. 10: 1–23.

WICKHAM, Hadley et al. (2019), "Welcome to the Tidyverse". *Journal of Open Source Software*, vol. 4,no. 43: 1686.

ZUCCO, Cesar. (2009), "Ideology or What? Legislative Behavior in Multiparty Presidential Settings". *The Journal of Politics*, vol. 71,no. 3: 1076–1092.

ZUCCO, Cesar & LAUDERDALE, Benjamin E. (2011), "Distinguishing Between Influences on Brazilian Legislative Behavior". *Legislative Studies Quarterly*, vol. 3,no. 36: 363–396.

Mapping behavior with ideal point estimation

Rodrigo Martins - Universidade Federal de Pernambuco

Ideal point estimation is a methodology of growing importance in Political Science. Developed in the mid-20th century, its link with the spatial theory of voting has been fruitful for several fields of study, with great expectations on its future development. The goal of this paper is to present what ideal point estimation is and didactically illustrate how to apply one of the various available techniques, using data on parliamentary behavior from the Chamber of Deputies in 2019.

1. Introduction

What structures divisions among parliamentarians, ideology or government support? Does change in party affect the behavior of legislators? What is the level of internal cohesion of political parties and how does it oscillate over time? Has polarization in the National Congress increased? Which political parties have the most unpredictable behavior? What kinds of bills tend to polarize political behavior? How does a president's ability to establish a coalition government influence parliamentary behavior?

These are some of the questions raised by the legislative studies literature to explain parliamentary behavior. In common among them is the need to access preferences, ideologies, and behaviors of elected representatives. However, several fields in Political Science have the same interest regarding other types of political actors and political behavior. Is it possible to find evidence of ideological behavior in the judiciary branch? Do politicians' speeches express ideological differences? Do the electorate's ideological preferences reflect those of their representatives? Does the option to follow certain people and not others on Twitter reflect the political preference of the platform's users?

These and several other questions are raised by studies that use ideal point estimation to explain the behavior of political actors. The spatial theory of voting, theorized by Downs (1957), began being applied to the most diverse areas, not only to explain traditional politicians, but also to shed light on the preferences of those outside political institutions. The ease in translating theoretical formalizations of preference formation into mathematical and statistical formulas made the spatial theory of voting find success in various manners of implementation.

The goal of this paper is to paper, in a simple and intuitive way, what is ideal point estimation and to illustrate how to use it didactically. Despite being used for varied purposes, the focus here will be on the field of studies that most developed the technique, the area of legislative studies. I will illustrate how to estimate ideal points using data from the Chamber Deputies in 2019, the first year of Jair Bolsonaro's presidential term. Previous studies that used ideal point estimation for the Brazilian case show that the main dimension that structures parliamentary behavior in the Senate and Chamber of Deputies is not ideological, but the government-opposition dimension (Izumi, 2016; Zucco, 2009), although at some moments these dimensions seem to be overlayed (Leoni, 2002). Faced with the analysis that Bolsonaro altered the relationship between the executive and the legislative by not forming a parliamentary support base via a coalition, replacing "coalition presidentialism with negligence" (Limongi, 2019), would there be a change in the Chamber of Deputies' behavior? Would there be another dimension structuring nominal votes in the plenary?

2. What are ideal points?

Ideal point estimation is derived from the spatial model of voting, initially made popular in Political Science by Downs (1975), in which political actors may be represented as a point in a space and choose between political alternatives. These also represented spatially, which are closer together. In addition to being a theoretical principle, several studies empirically demonstrate that the majority of political behavior can be represented by a unidimensional space, frequently assumed to be an ideological scale. The division between left and right then is an expression of how political actors, be they voters or representatives, behave and manifest their preferences in ways

that are aligned on distinct subjects, such as economic and social policies. This is similar to Converse's (1964) argument regarding a belief system that permeates preferences on varied issues in a unified worldview.

The ideal point estimation technique shows us a map of how voters are distributed spatially, representing the distance among the preferences of each individual under analysis. Simply put, what we extract is the proximity among voters when we consider all individuals, and not just their similarities two-by-two. In this way, it is even possible to estimate the proximity among voters that in practice did not participate together in any vote since, by estimating their distance in relation to all other voters, we would also know the distance among them. One of the model's assumptions is that when knowing how an individual decides on a specific theme, it would be possible to predict their behavior in relation to other decisions, especially on similar themes, by deriving from the previous decision. Thus, when correlating all decisions, it would be possible to estimate individuals' coordinates, so that those who decide in similar ways are spatially closer and those who diverge are spatially distant.

To illustrate the idea underlying ideal point estimation, we can turn to a hypothetical example. Let's suppose there is a board of three legislators that needs to make a decision on an economic issue. Considering that one of the legislators votes for more state intervention in the economy and the other two who vote for less intervention, the legislators' spatial representation will be done in such a way that the two who voted together occupy the same side, in opposition to the voter who remained isolated. It is also possible to illustrate the issue being voted as a line that represents the division among legislators on two different sides. Figure 1 is the way to spatially represent the division in this example. The red and green circle are the two who voted together, the blue circle the individual who voted alone and the purple line is the issue being voted on which represents the split between the two sides in question.

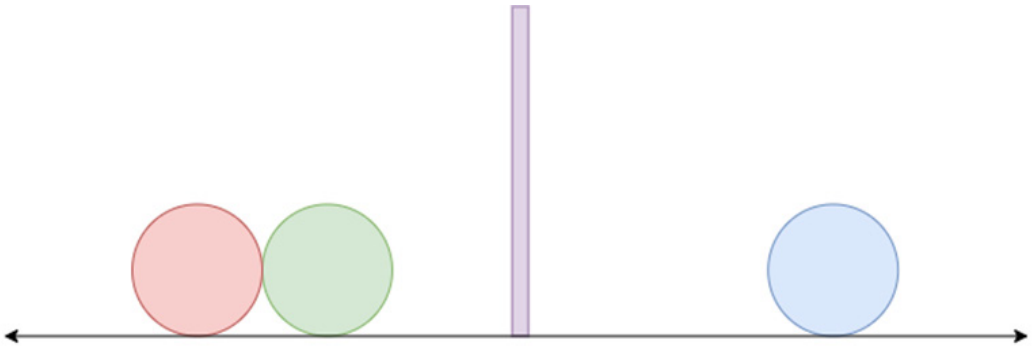


Figure 1: Example of spatial analysis with 1 vote and 3 voters

Let's suppose now that on a second vote on the same issue, the legislator represented by green does not vote again with the voter represented by red, but with the one represented by blue. In this second moment, the green individual must be placed between the others equidistantly, since they voted together with each of them. At the same time, those who did not vote together on either vote must remain at extreme opposition. The line that represents the first vote, in purple, must keep the division among voters as it did previously. And now a second vote line, which will be represented in white in figure 2, must isolate the legislator who voted alone and keep together those who voted together.

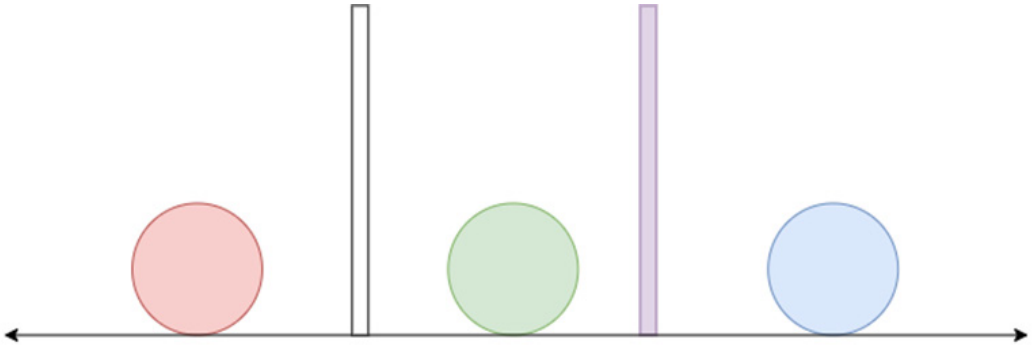


Figure 2: Example of spatial analysis with 2 votes and 3 voters

As a last example, let's imagine that on a third vote on matter unrelated to the economy, the legislators who before were opposed now vote together, leaving the voter who was once in the center, isolated in the minority. So that the spatial representation stays appropriate in this situation, it is best to add a vertical axis where the one who was defeated is distanced from the other two, while the victors are kept close to each other in this second axis. We can also represent this third vote as a line, represented in yellow in figure 3, this time horizontally.

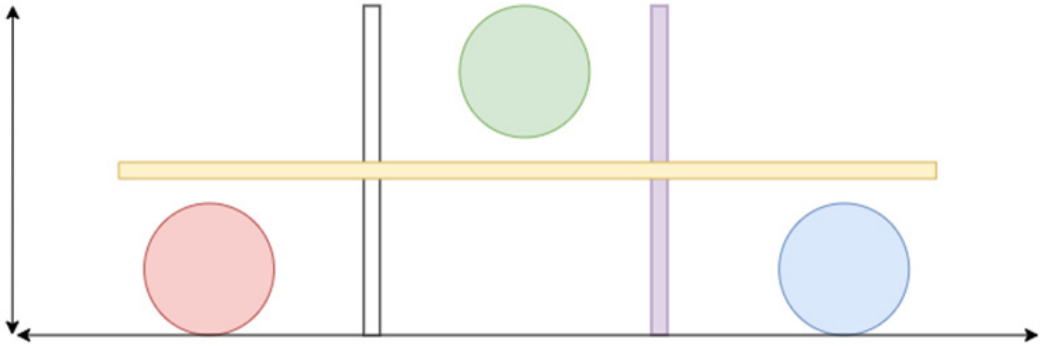


Figure 3: Example of spatial analysis with 3 votes and 3 voters

This type of model projects the preferences manifested through voting in a space (unidimensional, bidimensional, and even beyond), evaluating similarities and divergences in the decisions of each voter. As it would not be practical to conduct this manual process for many voters over many votes, several techniques were developed, with differing statistical procedures, with the goal of estimating ideal points both for voters and votes.

However, one must wonder that at any point it will no longer possible to perfectly represent the spatial distribution of voters and votes using only two dimensions. With an increase in voters and votes, it is inevitable that voters' points end up allocated in a position that does not correspond with their vote on a give issue. Poole (2005) proposes two measures to evaluate and compare the performance of ideal point models: the percentage of correct classifications and the aggregate proportional reduction in error (APRE). To better explain what these measures are and the difference between them, I will use a hypothetical example as illustration. Figure 4 represents a vote with 11 voters identified by a letter and a color.

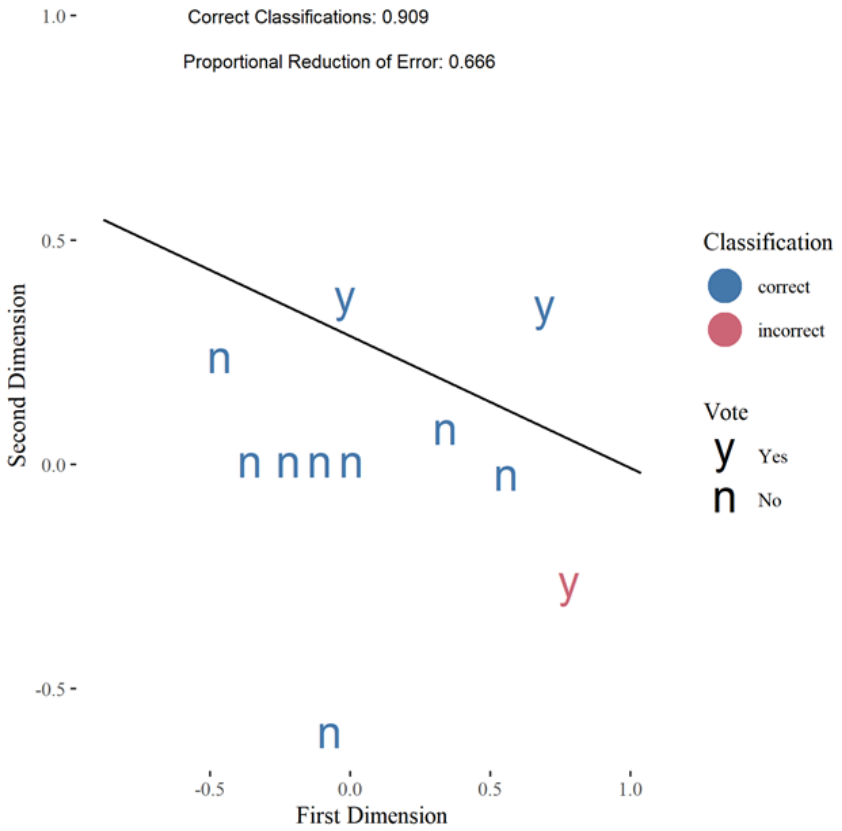


Figure 4: Example of evaluation of performance measures

Voters represented by the letter "y" voted "yes" and voters represented by the letter "n" voted "no". The only voter represented in red was the voter who was wrongly classified by the model, given that they voted "yes" when all the others represented in the same side of the vote voted "no". They should have been on the other side of the line, along with the others who voted "yes" and were correctly classified. Thus, the percentage of correct classification of this vote is 0.90922[1]. The proportional reduction in error (PRE) is 0.66[2]. This measure is more adequate and stricter in evaluating the performance of the models. Using the example of Armstrong et al. (2014), if at any vote there is a 65-35 split in favor of a proposal, we could correctly classify 65% of the votes, just by "naively" saying that all voters will support it. The proportional reduction in error measures how much the model improves the classification of votes in relation to the minority. Going back to our hypothetical example in figure 4, the PRE is 0.66 because it correctly classifies two out of three minority voters. The APRE is a measure that aggregates the PRE of each vote.

It is important to highlight that this method uses split votes, that is, votes that were not unanimous. This criterion is important, given that once one is looking to identify how divergence happens among voters, the issue at hand must necessarily produce dissent among them. Unanimous votes cannot give information on groupings and divergences and are thus disregarded. This characteristic may affect studies on institutions with a high rate of consensus. For the Federal Supreme Court, for example, where most decisions are monocratic or unanimous, the use of this method would lead to loss of explanatory capacity of the court as a whole. However, by using it only on majoritarian decisions, we would observe cases that deal with controversial issues, subject to dispute of distinct interpretations and polit-

1 number of correct classifications / number of voters

2 (votes in the minority – incorrect classifications) / votes in the minority

ical conflict. It is more likely that important cases, that have larger repercussion, fit into this scenario.

Care is needed when analyzing the dispersion of voters in space, because the model does not have the capacity to infer what determines that dispersion, that is, what is the factor that structures voting in spatial dimensions. This type of statistical estimation has been used in Brazilian Political Science, for example, when analyzing the Chamber of Deputies (Leoni, 2002; Zucco, 2009; Zucco and Lauderdale, 2011), Senate (Izumi, 2016), and Federal Supreme Court (Desposato et al., 2015; Ferreira and Mueller, 2014; Mariano Silva, 2018). For the cases of the Chamber of Deputies and the Senate, it is not ideological division that ends up being highlighted with ideal point estimation, but the division between opposition and government.

The first statistical ideal point estimation model elaborated in Political Science directly derived from the spatial theory of voting was NOMINATE (Poole and Rosenthal, 1985, 1991, 1997). This is a parametric model applied to binary votes and widely used in the area of legislative studies. There are several subsequent developments implemented in R, the most popular of which being W-NOMINATE (Poole et al., 2011), while D-NOMINATE and DW-NOMINATE are used to estimate dynamic ideal points, to observe the oscillation of voters' behaviors over time.

With the increase in computational performance, new implementations of ideal point estimations were elaborated with the application of Bayesian statistics. Derived from item-response theory, well developed in the field of psychometrics, and developed from the works of Clinton et al. (2004) and Martin and Quinn (2002), the implementation in R of IDEAL in the **pscl** package is one of the most popular, along with the MCMCpack package, which relies on an implementation of the Bayesian dynamic model, among others. Item-response theory is used mainly in educational studies, where they seek to measure the knowledge of individuals through tests, replaced in Political Science by measuring ideology from votes.

With distinct implementations, the NOMINATE family and Bayesian models have slightly different assumptions. While both assume that errors are independent, identically distributed, with normal distribution, and lower probability of occurring as the voter's ideal point is distant the vote line, the way each voter treats the available options when voting *yes* or *no* differs. In NOMINATE, if both alternatives are distant from the voter's ideal point, they are indifferent to the options offered, given that the preference curve takes on a Gaussian shape. In Bayesian models, the voter's preference curve takes on a quadratic shape, allowing them to maintain sensibility when facing options distant from their ideal point. In practice, despite this assumption not having much effect on the models' results (Carroll et al., 2009), to deal with this issue α-NOMINATE was created (Carroll et al., 2013). This Bayesian implementation of NOMINATE was made to replace previous implementations and it estimates if the best functional form of the voters' preferences is Gaussian, quadratic, or an intermediary mix of both.

In this paper, given the various statistical implementations of the spatial model to conduct ideal point estimation, I will demonstrate how to use *Optimal Classification* (OC). This is a non-parametric ideal point estimation technique, unlike the previous models. Its algorithm consists of the following steps (Poole, 2005, p. 82):

1. To generate initial values for the voters' ideal points from the decomposition into eigenvalues and eigenvectors from the concordance matrix between voter pairs.

2. Given the voters' values, to find the estimations of the votes' vectors that maximize the number of correct classifications.

3. Given the values of the votes' vectors, to find new values of the voters' ideal points that maximize the number of correct classifications.

4. Return to step 2.

This process is repeated until the number of correct classifications is stabilized, maximizing their number. According to Poole, this technique is consistent with data with few voters and few votes and one of its advantages is the fact that it has less assumptions when compared to parametric models of ideal point estimation, making it less susceptible to statistical estimation problems.

I chose this technique because, being non-parametric, it requires that less statistical assumptions on the voters' behaviors be satisfied. Rosenthal and Voeten (2004) argue that the strategic behavior that legislatures may present due to party discipline would be less problematic for *Optimal Classification* when compares to other parametric models[3].

In the following section, I will begin the explanation of how to implement an ideal point estimation, presenting the code in R for the required commands and functions to use this method.

3. First step: create a rollcall object

The packages that implement *Optimal Classification*, *W-NOMINATE*, α*-NOMINATE,* and *IDEAL* require that a *rollcall* type object be created in R, so that we can use its respective functions. To do so, it is necessary to specify each of the following elements in the rollcall() function:

1. the data where each line is a voter, and each column is a roll call vote

2. the values corresponding to a "yes" vote

3. the values corresponding to a "no" vote

4. the values corresponding to *missing*, when a voter did not vote for some reason

5. the values corresponding to when a voter was not part of the legislature

6. the values corresponding to the legislators' names, in the same order that they appear in the lines of the database

7. the values corresponding to the roll call votes' names, in the same order that they appear in the columns of the database

8. the values corresponding to additional information on the voters, such as their parties, in the same order that they appear in the lines of the database

9. the values corresponding to additional information on the roll call votes, such as their dates, in the same order that they appear in the columns of the database

The first three types of information are mandatory, while the others are optional. The code below illustrates the shape of a hypothetical database and how to create a *rollcall* object from it:

3 This problem is not fully solved by this technique (Spirling and McLean, 2007) since the monotonicity assumption still applies. See Duck-Mayr and Montgomery ([S.d.]) for more on this issue.

```
# Observe the database
hypotethical_data[1:5,1:13]
##          name      state      party V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
## 1 Legislator1     Alaska   Democrat  1  1  6  6  1  1  6  1  1   1
## 2 Legislator2      Texas   Democrat  1  1  6  1  1  6  6  6  1   1
## 3 Legislator3       Ohio Republican  1  1  6  6  1  6  6  6  6   6
## 4 Legislator4      Texas   Democrat  1  1  1  1  1  1  1  1  1   6
## 5 Legislator5 California   Democrat  6  6  6  1  6  6  6  6  6   1
# Create a rollcall object from the hypothetical_database object
hypotethical_rollcall<- rollcall(hypotethical_data[,-c(1:3)],
                                 yea = 1,
                                 nay = 6,
                                 legis.names = hypotethical_data$name,
                                 legis.data = hypotethical_data[,2:3])

hypotethical_rollcall
## Number of Legislators:    20
## Number of Votes:  100
##
## Using the following codes to represent roll call votes:
## Yea:      1
## Nay:      6
## Abstentions:  NA
## Not In Legislature:   9
##
## Legislator-specific variables:
## [1] "state" "party"
## Detailed information is available via the summary function.
```

First, we observe the database's first five lines and 13 columns to get an idea of its structure and organization. The first column represents the name of each legislator, the second their state of origin, and the third, their party. Following these, the columns represent votes, where 1 means a favorable vote and 6 means a contrary vote on the issue. To create a *rollcall* object, the first argument specified is the database, delimiting the voting columns only, and therefore excluding columns 1 through 3. Next the *yea* and *nay* arguments are specified, respectively 1 and 6. Then we specify the two optional arguments, *legis.names*, indicating the name column in the *hypotethical_data* object as the legislators' names, and columns 2 and 3 in the *legis.data* argument, given that these columns have additional information on the legislators. If we wanted to name the votes and include additional information on each one of them, we would specify the arguments *vote.names* and *vote.data* with their values. When observing the object created, *hypotethical_rollcall*, we can verify the structure of the data, indicating the number of legislators and votes, the votes' codifications as specified, and the additional variables on the legislators.

To analyze the Chamber of Deputies' behavior during the first year of the Bolsonaro government, we obtained voting data using the R package *congressbr* (McDonnell et al., 2019), which allows searches in the Chamber of Deputies and Senate's API[4] to collect various information on the National Congress's legislative process. When we request data from a given year, the object returned is in tidy format[5] where the unit of analysis is a parliamentarian's vote and a given roll call vote. The code below illustrates how to download the data to put them in the proper format for ideal point estimation and convert them to a *rollcall* object:

4 The API acronym means *Application Programming Interface*, interfaces that simplify several processes; among them, collection of data in a more accessible manner.

5 For an explanation on the principles of this data organization, see (Wickham, 2014)

```
# Install the tidyverse package, required for data manipulation
install.packages("tidyverse")

# Load the tidyverse package
library(tidyverse)

# Install the congressbr package, required to access the data from the #Chamber of
Deputies
install.packages("congressbr")

# Load the congressbr package
library(congressbr)

# Install the oc package
install.packages("oc")

# Load the oc package
library(oc)

# Download the Chamber of Deputies' votes from 2019
chamber_votes <- cham_votes_year(2019)

# Transform chamber_votes to wide format
chamber_wide <-
chamber_votes %>%
  transmute(legislator_name = str_to_upper(legislator_name),
            legislator_party = str_to_upper(legislator_party),
            rollcall_id, legislator_vote) %>%
  mutate(legislator_vote = recode(legislator_vote,
                                  "Sim" = 1,
                                  "Obstrucao" = 4,
                                  "Art. 17" = 5,
                                  "Nao" = 6,
                                  "Abstencao" = 7)) %>%
  mutate(legislator_id = paste0(legislator_name,
                                " - ",
                                 legislator_party)) %>%
  pivot_wider(names_from = rollcall_id,
              values_from = legislator_vote)

# Create a rollcall object
chamber_rollcall <- rollcall(chamber_wide[,-c(1:3)],
                             yea = 1,
                             nay = 4:7,
                             missing = NA,
                             legis.names = chamber_wide$legislator_id,
                             vote.names = colnames(chamber_wide)[-c(1:3)],
                             legis.data = chamber_wide[,1:3])
```

First, we need to install and load the packages required to manipulate and download the data. *Tidyverse* is a set of R packages with wide usage in data science[6]. Next, we download and install *congressbr* itself. The cham_votes_year() function downloads all the plenary votes in a given year, 2019 in this case.

We need to put the *chamber_votes* object in the adequate format to use the *rollcall* function. To do so, we need the database in *wide* format, where every line is a deputy, and every column is a roll call vote or information on the deputies. First, we select only the columns we want to keep, regarding parliamentarians' name, party, and vote and the column that identifies the roll call vote. Then, we recode the values in the votes column to numerical values. An important step is to create a new column, which I called *legislador_id*, comprised of the deputy's name and party. Then, if a parliamentarian changes party, they will be identified as another legislator. This is important as previous studies show (Izumi, 2016; McCarty et al., 2001) that party change and the entry or exit of parliamentarians from the governing coalition causes change in their behavior. If we do not take this step, we assume that a legislator's change in party has no impact in their parliamentary behavior. By making each point a legislator in a certain party, we can verify how much their behavior changes with the move. Lastly, we convert the data to the *wide* format, specifying that new columns will have the name of each roll call votes' ID, and its values will be the legislators' votes.

Finally, we can create the object in *rollcall* format. First, we indicate the object that has the data, selecting only the columns that correspond to roll call votes. Then we establish which values correspond to *yes*, *no*[7], and *missing*. The last arguments in the function refer to the column of legislators' names (which, in this case, is the combination of name and party), the name of the votes, and the additional information on the deputies.

4. Running Optimal Classification

With the data obtained and organized in the adequate manner, we case use them to estimate the ideal points. *Optimal Classification* is implemented by the *oc* package. To run the oc()[8] function it is necessary to specify the *rollcall* object that has the data to be used, the number of dimensions to be estimated, the minimum number of votes a legislator needs to be kept in the analysis, the minimum proportion (0 to 0.5) that the vote must have as minority to be kept in the analysis, and the polarity, which corresponds to the legislator that must have positive value in each of the estimated dimensions. Before estimating the model, some explanations about these elements are required.

The number of estimated dimensions, as mentioned previously, is a problem every researcher must keep in mind when using any ideal point estimation technique. Therefore, it is up to the researcher to elaborate a theoretical or statistical justification for their choice. One of the forms proposed by Poole (2005) of deciding the number of dimensions, which will be used in this paper, is by observing the *scree plot*. Widely used in psychometric studies to evaluate dimensionality, the idea behind this graph is to obtain some adjustment measure of the data by varying the number of dimensions estimated and observing the moment in which an "elbow" is formed, making the line more horizontal and less vertical, indicating that from a given dimension there is no great improvement in adding another dimension. This graph

6 For more details, see (Wickham et al., 2019)

7 To make it simple, we assigned all votes different from "yes" as contrary votes.

8 *W-NOMINATE*'s and α-NOMINATE's R implementations also require these same elements in their functions.

is generated by the NOMINATE packages with values calculated from the eigenvalues of the concordance matrix of votes among legislators. With the example used ahead, this will become clearer. However, this way of eliminating the number of dimensions should not be applied without reflecting on the substantive meaning of each of them, because it is frequent that an improvement in the measure of the models' fit has no meaning at all, being the result of a mere incorporation of noise to the data (Armstrong et al., 2014).

The minimum number of votes that legislators must have to remain in the estimation, as the minimum proportion that the minority must have so that a given vote remains in the estimation, are important because they can influence, albeit slightly at times, the final result. Votes with a very small minority are of little use to inform the division among its voters, and legislators with few votes bring low confidence to the consistency of their preferences. These votes and voters may end up influencing the estimations due to their outlier characteristics when compared to the data set. Therefore, it is common to exclude them, but it is up to the researcher which criteria will be established for that. The standard for *Optimum Classification*, if these values are not specified by researchers, is that voters with a minimum of 20 votes and votes with minimum minority of 2.5% be kept.

The polarity of the vote concerns specifying a voter who has a positive value in each dimension. This may seem like a crucial choice for the researcher but, in reality, it does not affect the final result in a significant way. For example, if we believed that the main dimension represents an ideological distribution of parliamentarians, and the higher the value the more conservative is the politician, then we should choose a conservative voter as polarity reference. If we specified a progressive, the result would only mirror our expectation, progressives with positive values and conservatives with negative values. In this way, the relative position among the points is not altered, but only the negative or positive sign of their values. If we are not certain or confident in who to choose as polarity, we can estimate the *Optimal Classification* with any legislator and verify the result a posteriori.

The code below shows how I ran the *Optimal Classification* for the Chamber of Deputies in 2019, from the data obtained previously:

```
# Running Optimal Classification's algorithm
chamber_oc <- oc(chamber_rollcall,
                 dims = 2,
                 polarity = rep("EDUARDO BOLSONARO - PSL", 2))
```

I specified the *rollcall* object built, asked to estimate two dimensions, and informed that in both dimensions deputy Eduardo Bolsonaro must have positive value. This process took roughly 20 seconds to run[9].

Unveiling dimensionality

We can have an idea on about the data by observing the object created:

9 This is an advantage when compared to the time for estimation of Bayesian models, which can take considerably longer.

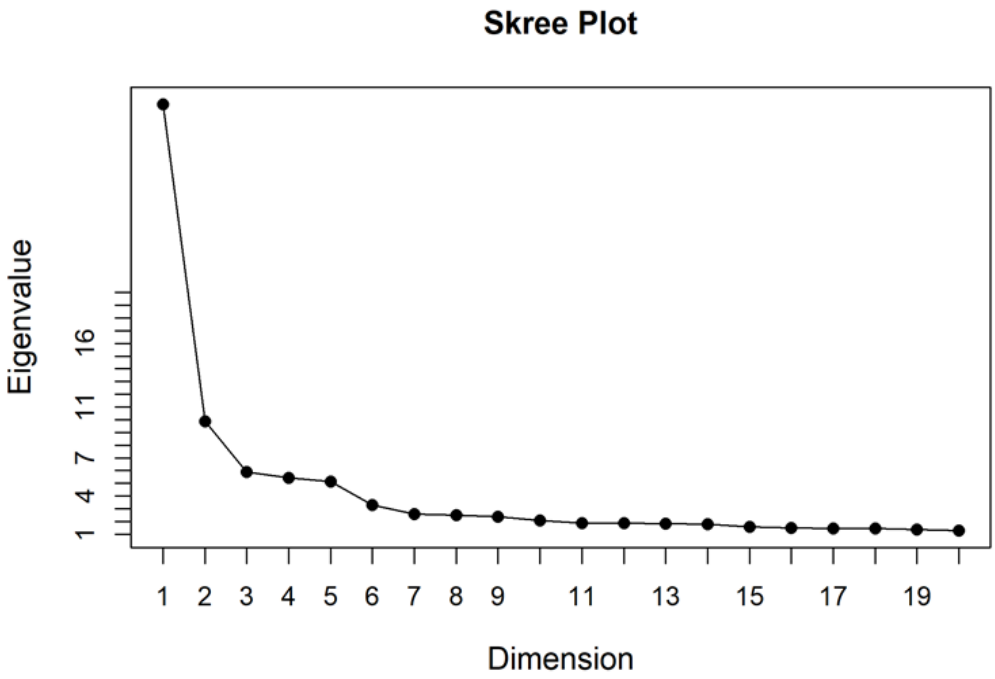
```
# Summary of the object generated by Optimal Classification
summary(chamber_oc)
##
##
## SUMMARY OF OPTIMAL CLASSIFICATION OBJECT
## --------------------------
##
## Number of Legislators:    616 (113 legislators deleted)
## Number of Votes:  259 (73 votes deleted)
## Number of Dimensions:     2
## Predicted Yeas:       47153 of 50123 (94.1%) predictions correct
## Predicted Nays:       46840 of 49446 (94.7%) predictions correct
##
## The first 10 legislator estimates are:
##                                coord1D coord2D
## EDIO LOPES - PL                  0.113  -0.360
## HAROLDO CATHEDRAL - PSD          0.143  -0.091
## HIRAN GONCALVES - PP             0.079  -0.380
## JHONATAN DE JESUS - PRB          0.249  -0.075
## JOENIA WAPICHANA - REDE         -0.529   0.703
## NICOLETTI - PSL                  0.283   0.009
## OTACI NASCIMENTO - SOLIDARIED    0.075  -0.126
## SHERIDAN - PSDB                  0.219  -0.025
## ACACIO FAVACHO - PROS            0.136  -0.265
## ALINE GURGEL - PRB              -0.067   0.085
```

Some useful information is displayed when we request a summary of the object created by *oc*. The number of legislators and roll call votes kept and excluded, how many dimensions were estimated, the number and percentage of correct classifications of favorable and contrary votes on the issues proposed, and the values estimated for the first 10 legislators in the database. It is possible to obtain the overall percentage of correct classifications of the votes and the APRE with the following command:

```
# Obtain performance measures
chamber_oc$fits
## [1] 0.9439986 0.7446653
```

The first value observed is the percentage of correct classifications, followed by the APRE. The values obtained are relatively high when compared to the values presented by other studies (Leoni, 2002). It is worth noting that when we estimate only one dimension, the percentage of correct classifications falls to 92.8% and the APRE to 0.67. The difference is small, giving us few elements to believe that there is a relevant second dimension. Another way to check the relevance of adding more dimensions is to observe the *scree plot*, with the eigenvalues generated by *Optimal Classification*. The logic behind this type of graph, as mentioned earlier, is to identify at which dimension an "elbow" is formed and to consider only the number of dimensions prior to that point. We can see the graph with the following code:

```
# Make scree plot
plot.OCskree(chamber_oc)
```

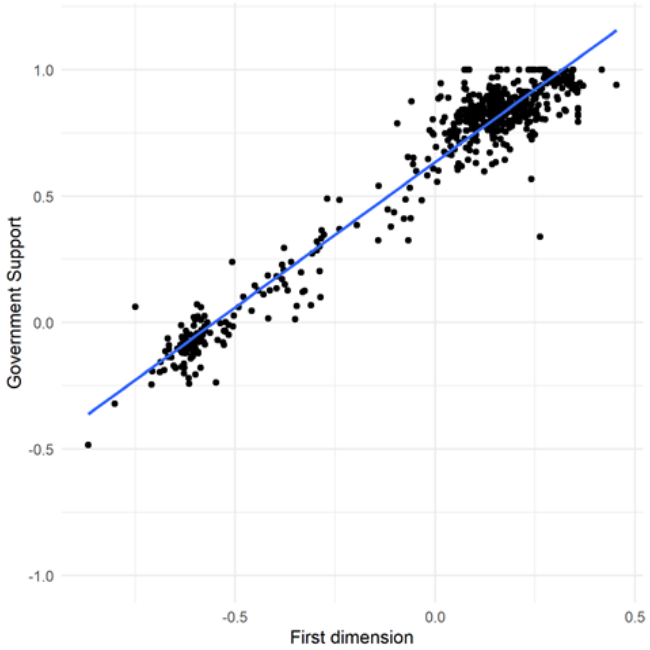


Graph 1: Screeplot for dimensionality evaluation

We can see that a single dimension is predominant in the data structure, with a second one being very discreet. This result follows previous studies that identified a single dimension as relevant in the structuring of votes in the Chamber of Deputies (Leoni, 2002) and Senate (Izumi, 2016). These studies also state that this dimension concerns the division between government and opposition. As mentioned previously, the interpreting the meaning of the estimated dimensions must be done by the researcher, as these are not given or informed by the ideal point estimation techniques.

We can check if, in 2019, in the first year of a government that did not establish a coalition government similar to ones prior, this dimension remains characterized by the government and opposition conflict. The graph below shows the correlation between estimated points for the first dimension and correlation of deputies' votes with the position the leader of the government laid out[10].

10 I excluded the votes for which the leader of the government allowed the deputies to vote freely.



Graph 2: Correlation between ideal points and government support

The graph shows a high correlation between the points estimated in the first dimension and the deputies' vote in the same direction stated by the leader of the government. This association is of 0.97. That is, deputies who usually vote in the same way as the leader of the government are further to the right of the graph, with higher values at their ideal points. Therefore, even though there is no coalition presidentialism in the same manner of prior governments, the dimension of political conflict in the Chamber of Deputies occurs in a similar way.

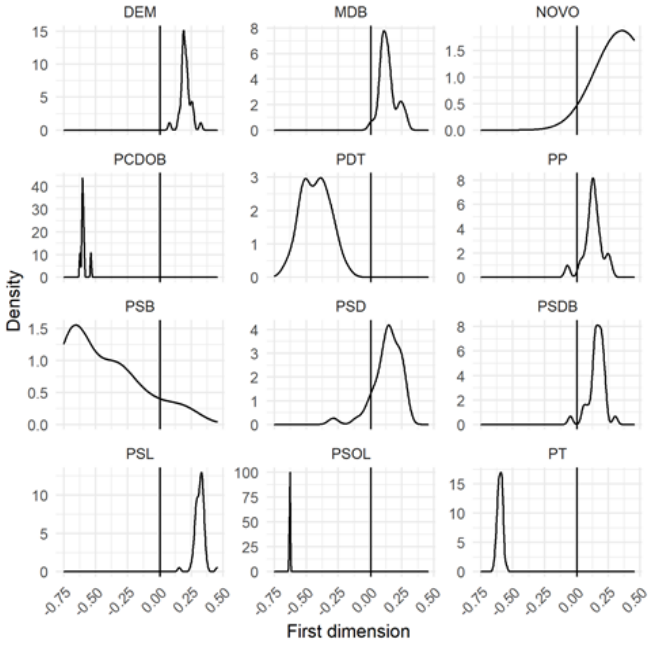
5. Distribution of ideal points

The values of each legislators' ideal points can be verified in the object created by *Optimal Classification*, in a list named *legislators*. In it can be found, for each voter, the amount of votes correctly classified, the estimated coordinates, and the variables on the legislators that were specified in the *legis.data* argument in the rollcall() function

```
glimpse(chamber_oc$legislators)
## Rows: 729
## Columns: 11
## $ legislator_name  <chr> "EDIO LOPES", "HAROLDO CATHEDRAL", "HIRAN GONCALVE...
## $ legislator_party <chr> "PL", "PSD", "PP", "PRB", "REDE", "PSL", "SOLIDARI...
## $ legislator_id    <chr> "EDIO LOPES - PL", "HAROLDO CATHEDRAL - PSD", "HIR...
## $ rank             <dbl> 282.0, 333.0, 220.0, 511.0, 95.0, 548.0, 212.0, 46...
## $ correctYea       <dbl> 69, 106, 79, 44, 67, 104, 100, 105, 95, 33, 97, 80...
## $ wrongYea         <dbl> 1, 8, 4, 4, 6, 0, 5, 2, 2, 5, 0, 1, 4, 0, 5, 0, 0,...
## $ wrongNay         <dbl> 1, 3, 0, 1, 20, 1, 4, 5, 4, 10, 1, 4, 4, 4, 6, 0, ...
## $ correctNay       <dbl> 47, 87, 71, 34, 91, 90, 82, 77, 83, 44, 83, 125, 6...
## $ volume           <dbl> 0.311, 0.019, 0.068, 0.006, 0.003, 0.070, 0.020, 0...
## $ coord1D          <dbl> 0.11288447, 0.14326323, 0.07852262, 0.24926999, -0...
## $ coord2D          <dbl> -3.600102e-01, -9.135407e-02, -3.802511e-01, -7.45...
```

With this information, we can elaborate a graph to show the parliamentarians' ideal points. The following code shows how to make a graph using the party variable to separately analyze the behavior of parties in the first dimension, using the *coord1D* variable:

```
chamber_oc$legislators %>%
  filter(legislator_party %in% c("PSL", "NOVO", "PT", "PSOL",
                                 "PCDOB", "PSB", "PDT", "PSD",
                                 "PSDB", "MDB", "PP", "DEM")) %>%
ggplot(aes(coord1D)) +
  geom_density() +
  facet_wrap(~legislator_party, drop = T, ncol = 3, scales = "free_y") +
  geom_vline(aes(xintercept = mean(chamber_oc$legislators$coord1D,
                                   na.rm = T))) +
  theme_minimal() +
  labs(x = "First dimension", y = "Density") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

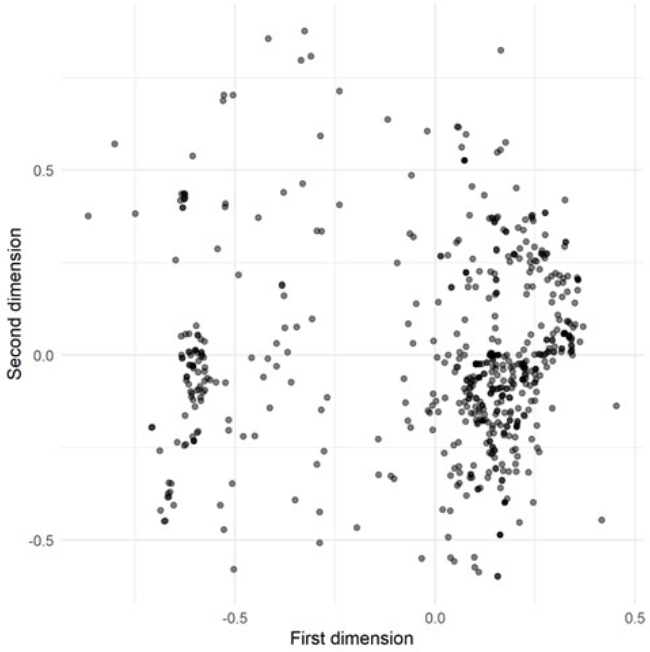


Graph 3: Party distribution in the first dimension

I opted to keep in the graph only 12 parties, given that if all 33 different parties and those without parties were added, the graph would be of little use to distinguish party behavior. I have also represented with a vertical line the mean value of points obtained of all parliamentarians. That way, it is possible to verify parties with different behaviors according to the dispersion of their ideal points over the first dimension. PT, PSOL, and PC do B appear concentrated in the left with little dispersion. PSB and PDT appear concentrated to the left of the mean, but very dispersed. Other parties, PSL, NOVO, DEM, MDB, PP, PSD, and PSDB are concentrated on the right.

In addition to the parliamentarians' dispersion in a single dimension, we can visualize the dispersion of their points on the two estimated dimensions:

```
ggplot(chamber_oc$legislators, aes(coord1D, coord2D)) +
  geom_point(alpha = 0.5) +
  theme_minimal() +
  labs(x = "First dimension", y = "Second dimension")
```



Graph 4: Distribution of ideal points on both dimensions

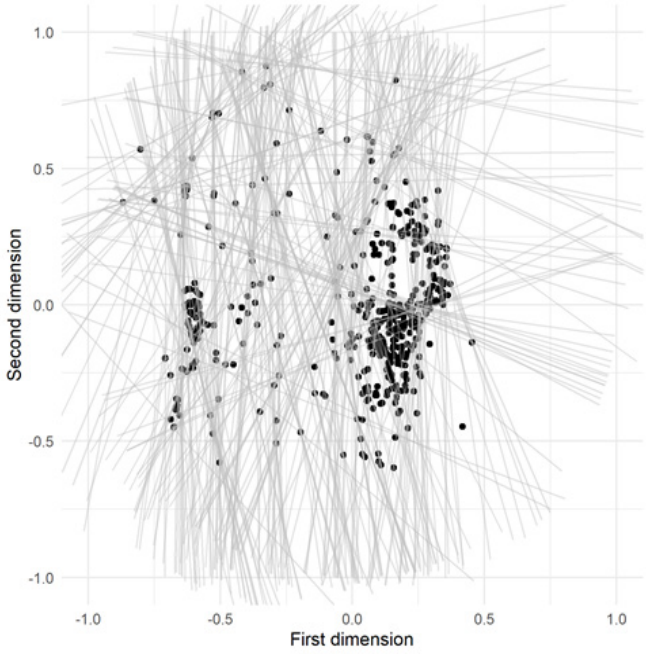
6. Distribution of the vote lines

I mentioned previously that the estimation of ideal points estimates both legislators' points and the vote lines that divide them. To do so, I used the information present in the *rollcalls* list that make up the object created by *Optimal Classification*. We can check, for every roll call vote, how many votes were correctly and incorrectly classified, the proportional reduction in error, and the necessary information to determine the lines that divide the parliamentarians.

```
head(chamber_oc$rollcalls)
##      correctYea wrongYea wrongNay correctNay       PRE normVector1D
## [1,]         NA       NA       NA         NA        NA           NA
## [2,]         NA       NA       NA         NA        NA           NA
## [3,]        297       12        1         72 0.8452381    0.9985689
## [4,]         94        4       36        216 0.6923077    0.9254343
## [5,]        121       31       13        244 0.6716418    0.2525467
## [6,]        212       26       20        112 0.6666667    0.2703514
##      normVector2D   midpoints
## [1,]           NA          NA
## [2,]           NA          NA
## [3,]  -0.05347965  0.23146515
## [4,]  -0.37890820 -0.06372496
## [5,]  -0.96758471  0.08072122
## [6,]  -0.96276172  0.08239360
```

The qualitative analysis of specific votes is also an important way of getting evidence on the substantive meaning of the dimensions estimated. We can not only identify the most frequent coalitions through this analysis, but we can also know the angle of these lines and determine which ones predominantly belong to the first dimension, which ones are influenced by a second dimension, and which are not explained well by either dimension. To add vote lines to the previous graph, the following code is necessary:

```
ggplot() +
  geom_point(data = chamber_oc$legislators,
             aes(coord1D, coord2D)) +
  geom_segment(data = data.frame(chamber_oc$rollcall),
               aes(x = (midpoints*normVector1D)+normVector2D,
                   y = (midpoints*normVector2D)-normVector1D,
                   xend = (midpoints*normVector1D)-normVector2D,
                   yend = (midpoints*normVector2D)+normVector1D),
               color = "grey", alpha = 0.4) +
  coord_cartesian(xlim = c(-1,1), ylim = c(-1,1)) +
  theme_minimal() +
  labs(x = "First dimension", y = "Second dimension")
```

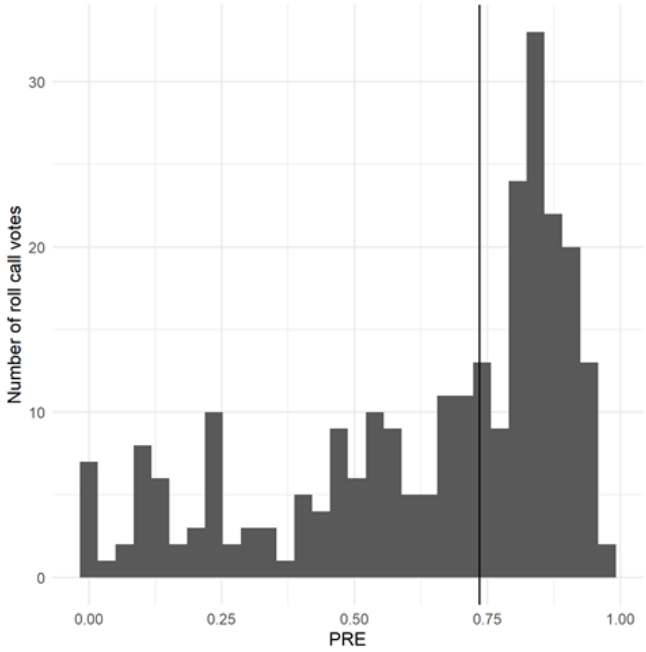


Grpah 5: Distribution of ideal points on both dimensions and vote lines

Note that in the geom_point() function, I specified the values of the ideal points, which are in the *legislators* list in the object created by *Optimal Classification*. In the geom_segment() function, I specified the values needed to calculate the initial and final points of the lines on the *x* and *y* coordinates, using the information from the roll call votes in the *rollcall* list of the same object created by *Optimal Classification*.

The graph gives us the impression that most lines have a more vertical inclination, crossing its *x* axis, indicating a dominance by the first dimension in the structuring of the voters' behaviors. However, we can analyze in more depth specific votes, to evaluate which are explained well by the estimation of ideal points. The first way of selecting votes is to observe which ones had a better proportional reduction in error. That information is available in the *legislators* list in object created by *Optimal Classification*. With the graph below we can observe the distribution of that value on all votes:

```
ggplot(data.frame(chamber_oc$rollcalls), aes(PRE)) +
  geom_histogram() +
  geom_vline(aes(xintercept = median(PRE, na.rm = T))) +
  theme_minimal() +
  labs(x = "PRE", y = "Number of roll call votes")
```



Graph 6: Distribution of the proportional reduction in error

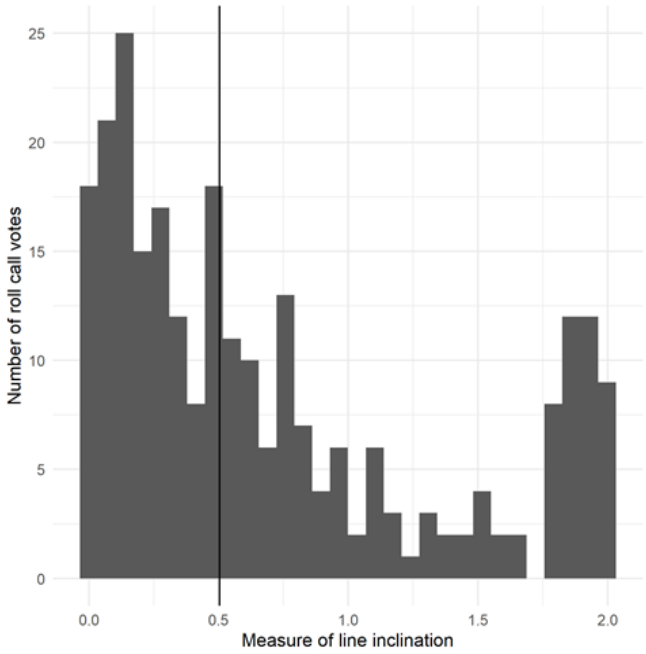
The median value of the proportional reduction in error is close to 0.75. Therefore, a first criterion to verify the substantive meaning of the first dimension should be to note the roll call votes with PRE higher than 0.75. That information alone does not help us in knowing which votes concern the first or second dimension, or both, but which votes are explained well by the estimation overall. One way of accessing which ones correspond to the first dimension is to use the same variables that were used to estimate the lines that divide the parliamentarians in roll call votes. We can do that in the following way:

```
x1 <- (chamber_oc$rollcalls[,"midpoints"]*chamber_oc$rollcalls[,"normVector1D"])+
  chamber_oc$rollcalls[,"normVector2D"]

x2 <- (chamber_oc$rollcalls[,"midpoints"]*chamber_oc$rollcalls[,"normVector1D"])-
  chamber_oc$rollcalls[,"normVector2D"]

x_diff <- abs(x2 - x1)

ggplot(NULL, aes(x_diff)) +
  geom_histogram() +
  geom_vline(aes(xintercept = median(x_diff, na.rm = T))) +
  theme_minimal() +
  labs(y = "Number of roll call votes", x = "Measure of line inclination")
```

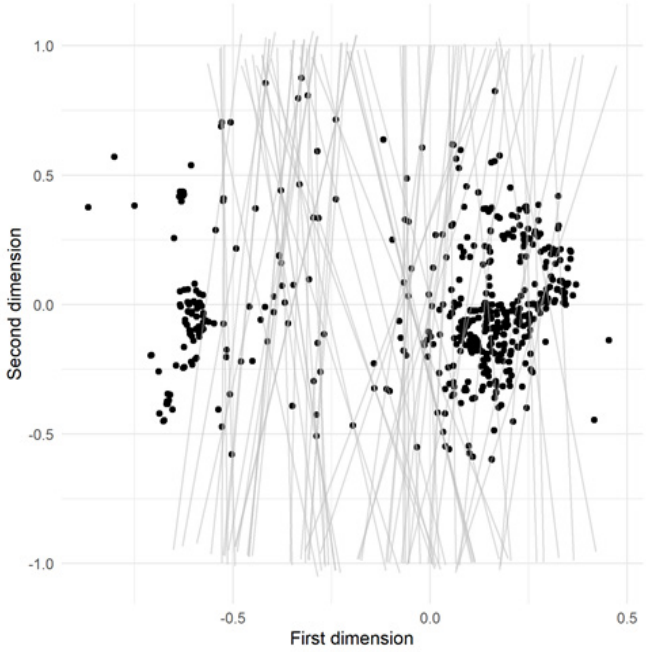


Graph 7: Measure of line inclination

The idea behind what I have done is: 1) calculate the value on the x axis, which corresponds to the first dimension, from the initial point in the line that divides the parliamentarians; 2) do the same for the final point; 3) know the absolute value of the difference between the two values. If the difference is low, that means that the line tends to have a more vertical position, being better explained by the first dimension. With the graph, we can verify that the median of that difference if very close to 0.5. therefore, values inferior to this are a second indicator of the first dimension's predominance. Lastly, I joined both criteria to select the votes:

```
selected_rollcalls <- which(chamber_oc$rollcalls[,"PRE"] > 0.75 &
                              x_diff < 0.5)

colnames(chamber_rollcall$votes)[selected_rollcalls]
##  [1] "PL-1292-1995-3"  "PL-1292-1995-7"  "PL-4742-2001-1"  "PLP-441-2017-7"
##  [5] "PLP-441-2017-9"  "PLP-441-2017-11" "MPV-852-2018-1"  "MPV-866-2018-1"
##  [9] "MPV-867-2018-3"  "MPV-867-2018-19" "MPV-871-2019-3"  "MPV-871-2019-4"
## [13] "MPV-871-2019-5"  "PEC-6-2019-2"    "PEC-6-2019-3"    "PEC-6-2019-4"
## [17] "PEC-6-2019-5"    "PEC-6-2019-6"    "PEC-6-2019-7"    "PEC-6-2019-8"
## [21] "PEC-6-2019-9"    "PEC-6-2019-13"   "PEC-6-2019-14"   "PEC-6-2019-18"
## [25] "PEC-6-2019-22"   "PEC-6-2019-24"   "PEC-6-2019-31"   "PEC-6-2019-32"
## [29] "PEC-6-2019-33"   "PEC-6-2019-35"   "PEC-6-2019-37"   "PEC-6-2019-38"
## [33] "PEC-6-2019-39"   "PEC-6-2019-40"   "PEC-6-2019-41"   "PEC-6-2019-42"
## [37] "PEC-34-2019-4"   "MPV-881-2019-1"  "MPV-881-2019-5"  "MPV-881-2019-6"
## [41] "MPV-881-2019-7"  "MPV-881-2019-8"  "MPV-881-2019-9"  "MPV-881-2019-11"
## [45] "PL-2787-2019-1"  "PL-2788-2019-2"  "REQ-1464-2019-1" "PL-2999-2019-6"
## [49] "PL-3261-2019-5"  "PL-3261-2019-6"  "PL-3261-2019-7"  "PL-3261-2019-9"
## [53] "PL-3261-2019-10" "PL-3261-2019-12" "PL-3261-2019-13" "MPV-886-2019-1"
## [57] "MPV-886-2019-2"  "PL-3715-2019-2"  "PL-3715-2019-9"  "MPV-890-2019-3"
## [61] "MPV-890-2019-5"  "MPV-890-2019-6"  "MPV-890-2019-7"  "MPV-890-2019-9"
## [65] "PL-4162-2019-1"  "PL-4162-2019-2"  "REQ-2110-2019-1" "REQ-2157-2019-1"
## [69] "REQ-2234-2019-1" "PDL-523-2019-1"  "PDL-523-2019-4"  "PL-5815-2019-1"
## [73] "PL-5815-2019-2"  "PL-5815-2019-3"
```

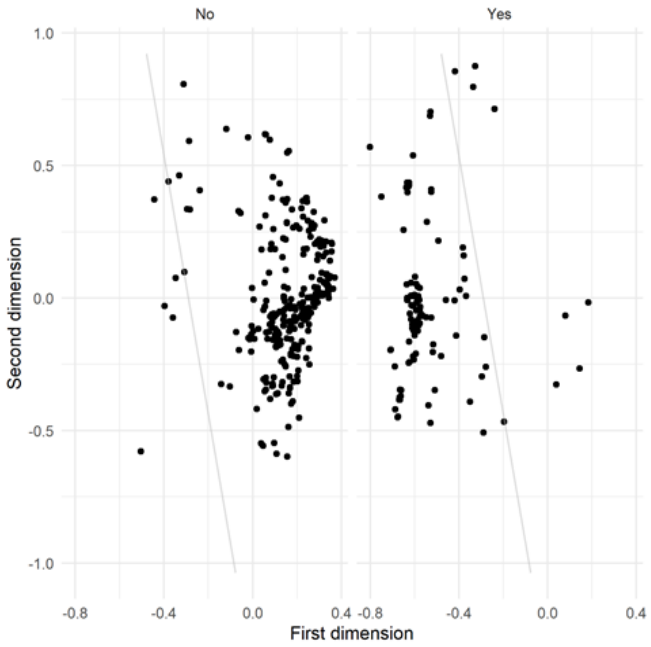


Graph 8: Ideal points and vote lines corresponding to the first dimension

Note that the difference is in the specification of the geom_segment() function, where I indicated in the *chamber_oc$rollcall* object that I only wanted the lines that correspond to the votes selected. We succeeded in our goal of selecting only the roll call votes that were explained well by the technique and concern the first dimension. We can observe that there are votes that end up splitting both the left-wing and the right-wing party blocs.

With the information gathered on each voter's ideal points and the lines for each roll call vote, it is possible to analyze specific decisions qualitatively. Thus, it is possible to better understand how the structure of preferences and political conflict is formed among deputies, contributing even more towards an interpretation of the substantive meaning of the dimensions estimated and the results overall. Below I show an example, with of one of the votes for the social security reform PEC.

```
chamber_oc$legislators %>%
  mutate(Vote = chamber_rollcall$votes[,147]) %>%
  mutate(Vote = case_when(Vote == 1 ~ "Yes",
                          Vote != 1 ~ "No")) %>%
  filter(!is.na(Vote)) %>%
ggplot() +
  geom_point(aes(coord1D, coord2D)) +
  geom_segment(data = data.frame(chamber_oc$rollcall) %>%
                 slice(147),
               aes(x = (midpoints*normVector1D)+normVector2D,
                   y = (midpoints*normVector2D)-normVector1D,
                   xend = (midpoints*normVector1D)-normVector2D,
                   yend = (midpoints*normVector2D)+normVector1D),
               color = "grey", alpha = 0.5) +
  theme_minimal() +
  facet_wrap(vars(Vote), nrow = 1) +
  labs(x = "First dimension", y = "Second dimension")
```

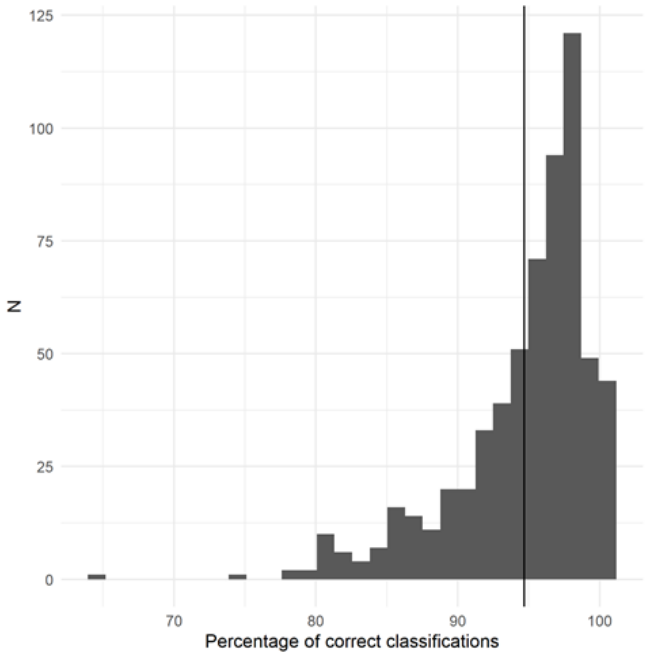


Graph 9: Ideal points and vote line for one of the votes concerning PEC 6/2019

In the code above, I used the object created by *Optimal Classification* and added a column for the votes from column 147, which refers to the selected roll call vote on social security reform. Next, I recoded the values to indicate if the vote was in favor or against the issue, removing those who did not vote, and generated the graph. I specified that two separate panels be made, according to the vote, and kept only the line for the selected roll call vote. We can see on the graph that government supporters who, in this case, voted against the proposal, are concentrated on the right side, with higher ideal points values in the first dimension, while the opposition voted for it. We can observe that the estimation gets wrong five deputies who voted *no* and eight deputies who voted *yes*, given that they are located at the opposite side the line than it would be expected. That roll call vote has 97.1% of the votes correctly classified and a PRE of 0.88. This graph enables a qualitative analysis of this specific vote and specific voters. One example is the behavior of deputy Tabata Amaral. Within the context of the social security reform, there was a lot of criticism, both on social media and internally in the PDT, about some of the votes that the deputy had in favor of the approved text. Her ideal point is the one that voted *no*, located higher on the graph. As we have noted, her vote is correctly classified by the estimation, placing her on the right side of the line that splits the vote. Thus, her vote is coherent with her behavior on other votes, placing her point in that space. In this case, even if located more to the left, her vote with those on the right would not be atypical, given how political conflict is structured in the Chamber of Deputies when we spatially analyze all the deputies and roll call votes.

Another possibility for analysis of voter behavior is to observe which ones are more "unpredictable", with a higher number of incorrect classifications because their behavior does not follow the logic of the dimensions found, but others not estimated. There are specific models to handle these types of voters more appropriately (Lauderdale, 2010). In the object created by *Optimal Classification*, columns are presented with the result of how many *yes* and *no* were correctly and incorrectly classified. Thus, it is possible to know the percentage of votes for each voter that the estimation of ideal points was able to adequately allocate in the space. We can get the results and elaborate a graph in the following way:

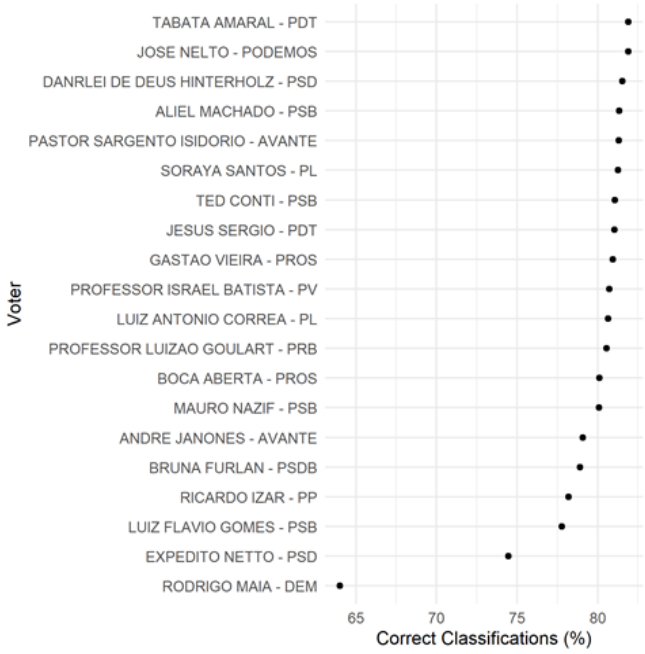
```
chamber_oc$legislators %>%
  mutate(correct_classifications = 100 * (correctYea + correctNay) /
           (correctYea + wrongYea + wrongNay + correctNay)) %>%
  ggplot(aes(correct_classifications)) +
  geom_histogram() +
  geom_vline(aes(xintercept = mean(correct_classifications, na.rm = T))) +
  theme_minimal() +
  labs(x = "Percentage of correct classifications", y = "N")
```



Graph 10: Distribution of the percentage of the voters' correct classifications

With the histogram, we can observe that the average of correct classifications of voters is 94.6% and the overwhelming majority of deputies is above 90%. To observe who are the voters whose behaviors are less explained by the two dimensions estimated, I have used the following code:

```
chamber_oc$legislators %>%
  mutate(correct_classifications = 100 * (correctYea + correctNay) /
           (correctYea + wrongYea + wrongNay + correctNay)) %>%
  arrange(correct_classifications) %>%
  slice(1:20) %>%
  mutate(legislator_id = fct_reorder(legislator_id, correct_classifications)) %>%
  ggplot(aes(correct_classifications, legislator_id)) +
  geom_point() +
  theme_minimal() +
  labs(x = "Correct Classifications (%)", y = "Voter")
```



Graph 11: Voters with smaller percentage of correct classifications

I kept in the graph only the 20 voters with the smallest percentage of correctly classified votes. We can observe that Rodrigo Maia is the one who the smallest percentage, with less than 65%, followed by Expedito Netto, with a little under 75%. It is curious that deputy Tabata Amaral, mentioned previously due to the critiques on her behavior, is among those whose behavior is not well explained by the two dimensions estimated.

7. Discussion

This paper showed how to estimate ideal points and some ways to explore the results from one of the technique's several applications. Many other R packages have been implemented to explain the behavior of various political actors with varied types of data.

For studies that work with text, instead of votes on bills, the choice of actors is in the use of words. When dealing with a certain topic, using certain words and omitting others would reveal an ideological preference on how to approach an issue. Moreover, the relevance of a topic itself, to the detriment of others, in a political speech, would reveal ideological content. Implementations such as *Wordfish* (Slapin and Proksch, 2008), *Wordscores* (Laver et al., 2003), and *Wordshoal* (Lauderdale and Clark, 2014) seek to deal with this type of issue, analyzing party manifestos, parliamentary speeches, among others.

The relative ease in obtaining Twitter data has also enable the use of this type of data for ideal point estimation. Not only the text of the tweets made by users can be analyzed with the implementations just mentioned, but also the way users interact and connect to one another. The act of choosing to follow certain accounts to the detriment of others would reveal an ideological preference of users of that social network in getting certain types of content (Barberá, 2015).

Several R packages that surface daily, more than expanding what type of data can be used for ideal point estimation, also look to expand the possibilities for better statistical inference from data already in use for a long time. We cannot estimate ideal points, for example, of the 2019 Chamber of Deputies and directly compare the results with an estimation from the previous year.

For that, theoretical and statistical assumptions are required that are implemented by other algorithms (Bailey et al., 2017; Martin e Quinn, 2002; Poole and Rosenthal, 2001). In addition to intertemporal comparison, it is possible compare ideal points between different institutions and actors, with the proper data and methods (Bailey, 2007; Bonica, 2014; Tausanovitch and Warshaw, 2013).

Regardless of future developments, it is necessary to understand the possibilities that these techniques offer in the present, with the limitations and virtues. I hope to have contributed so that new studies can adequately implement these techniques.

BIBLIOGRAPHICAL REFERENCES

ARMSTRONG, David A et al. (2014), *Analyzing spatial models of choice and judgment with R*. [S.l.] CRC Press.

BAILEY, Michael A. (2007), "Comparable preference estimates across time and institutions for the court, congress, and presidency". *American Journal of Political Science*, vol. 51,no. 3: 433–448.

BAILEY, Michael A & STREZHNEV, Anton & VOETEN, Erik. (2017), "Estimating dynamic state preferences from United Nations voting data". *Journal of Conflict Resolution*, vol. 61,no. 2: 430–456.

BARBERÁ, Pablo. (2015), "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". *Political analysis*, vol. 23,no. 1: 76–91.

BONICA, Adam. (2014), "Mapping the ideological marketplace". *American Journal of Political Science*, vol. 58,no. 2: 367–386.

CARROLL, Royce et al. (2009), "Comparing NOMINATE and IDEAL: Points of difference and Monte Carlo tests". *Legislative Studies Quarterly*, vol. 34,no. 4: 555–591.

CARROLL, Royce et al. (2013), "The structure of utility in spatial models of voting". *American Journal of Political Science*, vol. 57,no. 4: 1008–1028.

CLINTON, Joshua & JACKMAN, Simon & RIVERS, Douglas. (2004), "The statistical analysis of roll call data". *American Political Science Review*, vol. 98,no. 2: 355–370.

CONVERSE, Philip E. (1964), "The nature of belief systems in mass publics. Ideology and discontent". *Ideology and Discontent*, 206–261.

DESPOSATO, Scott W. & INGRAM, Matthew C. & LANNES JR., Osmar P. (2015), "Power, Composition, and Decision Making: The Behavioral Consequences of Institutional Reform on Brazil's Supremo Tribunal Federal". *Journal of Law, Economics, and Organization*, vol. 31,no. 3: 534–567.

DOWNS, Anthony. (1957), *An economic theory of democracy*. New York, Harper; Row.

DUCK-MAYR, JBrandon & MONTGOMERY, Jacob M. *Ends Against the Middle: Scaling Votes When Ideological Opposites*. [S.d.].

FERREIRA, Pedro Fernando Almeida Nery & MUELLER, Bernardo. (2014), "How judges think in the Brazilian Supreme Court: Estimating ideal points and identifying dimensions". *Economia*, vol. 15,no. 3: 275–293.

IZUMI, Mauricio. (2016), "Governo e Oposição no Senado Brasileiro (1989-2010)". *Dados*, vol. 59,no. 1: 91–138.

LAUDERDALE, Benjamin. (2010), "Unpredictable voters in ideal point estimation". *Political Analysis*, no. 18: 151–171.

LAUDERDALE, Benjamin E. & CLARK, Tom S. (2014), "Scaling politically meaningful dimensions using texts and votes". *American Journal of Political Science*, vol. 58,no. 3: 754–771.

LAVER, Michael & BENOIT, Kenneth & GARRY, John. (2003), "Extracting policy positions from political texts using words as data". *American Political Science Review*, vol. 97,no. 02: 311–331.

LEONI, Eduardo. (2002), "Ideologia, Democracia e Comportamento Parlamentar: A Camara dos Deputados (1991-1998)". *Dados*, vol. 45,no. 3: 361–386.

LIMONGI, Fernando. (2019), "Presidencialismo do Desleixo". *Piauí*, no. 158: Disponível em: <https://piaui.folha.uol.com.br/materia/presidencialismo-do-desleixo/>.

MARIANO SILVA, Jeferson. (2018), "Mapeando o Supremo. As posições dos ministros do STF na jurisdição constitucional (2012-2017)". *Novos Estudos - CEBRAP*, vol. 37,no. 1: 35–54.

MARTIN, Andrew D. & QUINN, Kevin M. (2002), "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis*, vol. 10,no. 2: 134–153.

MCCARTY, Nolan & POOLE, Keith T. & ROSENTHAL, Howard. (2001), "The Hunt for Party Discipline in Congress". *American Political Science Review*, vol. 95,no. 3: 673–87.

MCDONNELL, Robert Myles & DUARTE, Guilherme Jardim & FREIRE, Danilo. (2019), "congressbr: An R Package for Analyzing Data from Brazil's Chamber of Deputies and Federal Senate". *Latin American Research Review*, vol. 54,no. 4:

POOLE, Keith et al. (2011), "Scaling Roll Call Votes with wnominate in R". *Journal of Statistical Software*, vol. 42,no. 14: 1–21.

POOLE, Keith T. (2005), *Spatial models of parliamentary voting*. [S.l.] Cambridge University Press.

POOLE, Keith T & ROSENTHAL, Howard. (1985), "A spatial model for legislative roll call analysis". *American Journal of Political Science*, 357–384.

POOLE, Keith T & ROSENTHAL, Howard. (1997), *Congress: A political-economic history of roll call voting*. [S.l.] Oxford University Press on Demand.

POOLE, Keith T & ROSENTHAL, Howard. (2001), "D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting". *Legislative Studies Quarterly*, 5–29.

POOLE, Keith T & ROSENTHAL, Howard. (1991), "Patterns of congressional voting". *American Journal of Political Science*, 228–278.

ROSENTHAL, Howard & VOETEN, Erik. (2004), "Analyzing roll calls with perfect spatial voting: France 1946–1958". *American Journal of Political Science*, vol. 48,no. 3: 620–632.

SLAPIN, Jonathan B & PROKSCH, Sven-Oliver. (2008), "A scaling model for estimating time-series party

positions from texts". *American Journal of Political Science*, vol. 52,no. 3: 705–722.

SPIRLING, Arthur & MCLEAN, Iain. (2007), "UK OC OK? Interpreting optimal classification scores for the UK House of Commons". *Political Analysis*, 85–96.

TAUSANOVITCH, Chris & WARSHAW, Christopher. (2013), "Measuring constituent policy preferences in congress, state legislatures, and cities". *The Journal of Politics*, vol. 75,no. 2: 330–342.

WICKHAM, Hadley. (2014), "Tidy data". *Journal of Statistical Software*, vol. 59,no. 10: 1–23.

WICKHAM, Hadley et al. (2019), "Welcome to the Tidyverse". *Journal of Open Source Software*, vol. 4,no. 43: 1686.

ZUCCO, Cesar. (2009), "Ideology or What? Legislative Behavior in Multiparty Presidential Settings". *The Journal of Politics*, vol. 71,no. 3: 1076–1092.

ZUCCO, Cesar & LAUDERDALE, Benjamin E. (2011), "Distinguishing Between Influences on Brazilian Legislative Behavior". *Legislative Studies Quarterly*, vol. 3,no. 36: 363–396.