Análise de Votações Legislativas Utilizando Componentes Principais

Resumo

A literatura sobre análise quantitativa de votações nominais em casas legislativas é bem extensa e relativamente antiga. No entanto, poucos são os trabalhos que se debruçam sobre a aplicação de tais métodos nas casas legislativas brasileiras. Este artigo trás uma revisão de métodos de análise de votações legislativas nominais e discute a utilização da análise de componentes principais (ACP) como um método simples e eficaz para analisar votações nominais de casas legislativas. Apresentaremos também nossa abordagem para o tratamento de abstenções e sobre a realização de análises agregadas por partido, temas pouco explorados na literatura, que normalmente foca em análises de casas legislativas dos Estados Unidos.

São apresentados gráficos bidimensionais de algumas legislaturas do Congresso Americano e da Câmara dos Deputados e Senado brasileiros, por parlamentar e agrupados por partido, e comparados os resultados do modelo ACP com os do conhecido modelo WNOMINATE. O modelo ACP além de ser mais simples, computacionalmente mais rápido e de mais fácil interpretação, apresentou métricas de adequação (fitness) melhores para as casas legislativas brasileiras. Saulo ►*este último, a confirmar!*◄

1 Introdução

Leo ►*deixar pra ajeitar a coesão da introdução por último*◄

Modelos espaciais para análise de votações no âmbito legislativo existem pelo menos desde 1957, com Downs [1], e se tornaram mais numerosos e mais utilizados a partir da década de 1980, com o aumento da disponibilidade e redução de custo de processamento computacional e com a proposição em 1985 do famoso algoritmo NOMINATE por Poole e Rosenthal [2], até hoje o mais conhecido e utilizado. O objetivo destes modelos de escalonamento dimensional é representar os parlamentares ou partidos em um espaço

geométrico com algumas poucas dimensões (frequentemente uma ou duas) de tal forma que o comportamento de cada um nas votações seja em grande parte explicado por sua posição (coordenadas) neste espaço, sendo que esta posição, também chamada "ponto ideal" do legislador ou partido, é estimada a partir dos votos observados nas votações.

Existe uma literatura relativamente ampla sobre o assunto focando no Congresso Americano ou outras entidades estadounidenses como o senado e suprema corte, inclusive comparando a performance de diferentes modelos [3, 4], mas são poucos os estudos de votações em entidades brasileiras. Um exemplo é Leoni, que analisou as votações da Câmara dos Deputados entre os anos 1991 e 1998 utilizando W-NOMINATE [5].

A ACP é o método estatístico mais popular para redução dimensional de grandes conjuntos de dados [6], e algoritmos de determinação de componentes principais através de decomposição em valores singulares (SVD) são amplamente disponíveis em softwares e bibliotecas de matemática e estatística.

Leo ►*Falar que implementamos o ACP no Radar Parlamentar.*◄

Os objetivos deste trabalho são:

- Contextualizar o uso da ACP no histórico da literatura de análise quantitativa de votações nominais em casa legislativas.
- Apresentar nossa abordagem do uso da ACP no contexto brasileiro, com o devido preenchimento de algumas lacunas da literatura.
- Mostrar os resultados da aplicação de nossa abordagem para a ACP nas casas legislativas federais do Brasil.

Este artigo está organizado da seguinte forma: ...

2 O modelo ACP para análise de votações nominais

Considera-se uma casa legislativa com M membros (parlamentares) e N votações nominais de interesse. O voto x_{ij} de um parlamentar j em uma votação i será modelado por um valor numérico como segue:

$$x_{ij} = \begin{cases} 1 & \text{, se parlamentar votou } sim \\ -1 & \text{, se parlamentar votou } não \\ 0 & \text{, em qualquer outro caso} \end{cases}$$

Os outros casos além do sim e do não podem consistir em abstenção, obstrução ou ausência do parlamentar, ou situação em que este não esteja exercendo o mandato na data em que a votação ocorreu. Todos esses casos representam uma impossibilidade de verificar a opinião do parlamentar sobre a votação, e por isso são modelados por um valor euclidianamente equidistante das duas opções.

Normalmente a análise de componentes principais não é adequada para variáveis categóricas, porém neste caso as categorias podem ser claramente representadas em um eixo cartesiano com dois extremos: SIM e NÃO. O valor de x_{ij} pode ser interpretado como um estimador para um ponto de utilidade máxima ξ_{ij} do legislador j face à decisão i situado em uma escala contínua de valores deste eixo, tal que quando $\xi_{ij} > 0$ o legislador tende a preferir o SIM, e com mais convicção ou maior importância dada à questão quanto mais distante do zero, e analogamente para $\xi_{ij} < 0$ e a opção NÃO. Ora, o comportamento observado que é o voto, por sua natureza categórica, não permite dizer o grau de importância dada ou a convicção com que o parlamentar decidiu por uma ou outra opção, mas é razoável supor que os x_{ij} tal como definidos acima forneçam um estimador para os ξ_{ij}.

Fica definida a matriz de votações $\mathbf{X}$:

$$\mathbf{X} = \begin{array}{c} \\ \text{votações} \downarrow \end{array} \begin{array}{c} \overset{\text{membros}}{\longrightarrow} \\ \begin{array}{cccccc} & 1 & & j & & M \\ 1 & x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ & \vdots & \ddots & \vdots & \ddots & \vdots \\ i & x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ & \vdots & \ddots & \vdots & \ddots & \vdots \\ N & x_{N1} & \dots & x_{NM} & \dots & x_{NM} \end{array} \end{array}$$

Por definição esta matriz contém apenas os valores -1, 0 e 1. Para realizar a análise de componentes principais, define-se a matriz centralizada $\mathbf{X}^*$, subtraindo de cada entrada a média da linha:

$$x^*_{ij} = x_{ij} - \langle x_{ij} \rangle_j \tag{1}$$

onde $\langle \cdot \rangle_j = \frac{1}{M} \sum_{j=1}^{M} \cdot$ denota a média nos j.

Define-se a matriz de centralização $\mathbf{C}$ por:

$$c_{ij} = \langle x_{ij} \rangle_j \qquad i = 1..N;\ j = 1..M$$

de forma que:

$$\mathbf{X}^* = \mathbf{X} - \mathbf{C}$$

A variância (amostral) var(i) de cada votação, ou dimensão é:

$$\mathrm{var}(i) = \frac{\sum_{j=1}^{M} \left(x_{ij} - \langle x_{ij} \rangle_j \right)^2}{M-1} = \frac{M}{M-1} \left(\langle {x_{ij}}^2 \rangle_j - \langle x_{ij} \rangle_j^2 \right)$$

$$\mathrm{var}(i) = \frac{M}{M-1} \langle {x^*_{ij}}^2 \rangle_j \tag{2}$$

A análise de componentes principais consiste em uma rotação de base $\mathbf{R}$ deste espaço vetorial tal que os dados (centralizados) transformados $\mathbf{\Gamma} = \mathbf{R} \cdot \mathbf{X}^*$ concentram a máxima variância possível na primeira dimensão, a segunda dimensão possui a máxima variância possível sob a restrição de ser ortogonal à primeira, e assim sucessivamente. A cada vetor da nova base é dado o nome de *componente principal*, os valores de $\mathbf{R}$ são chamados *pesos* (ou *loadings*) e as coordenadas obtidas em $\mathbf{\Gamma}$ são chamadas de *scores*.

//////////////// A execução é tipicamente muito rápida, com complexidade $O(mn^2)$ onde $m > n$ são as dimensões da matriz de dados, utilizando a notação *big-Oh* [7]. Neste trabalho foi utilizada a função *prcomp* do R e a execução para $m \approx n \approx 400$ leva menos de 2 segundos em um computador pessoal com processador de 2,4 GHz.

Como a matriz de rotação $\mathbf{R}$ é ortonormal, sua inversa é igual à transposta $\mathbf{R}^t$, e tem-se $\mathbf{X}^* = \mathbf{R}^t \cdot \mathbf{\Gamma}$.

Se forem mantidos apenas os $d \leq N$ primeiros componentes principais, a parte relevante da matriz de rotação, que chamaremos de $\mathbf{R}_{(d)}$, e da matriz de scores, $\mathbf{\Gamma}_{(d)}$, terão apenas d linhas, e $\mathbf{R}^t_{(d)} \cdot \mathbf{\Gamma}_{(d)}$ será a melhor aproximação de $\mathbf{X}^*$ que pode ser obtida com um modelo linear deste tipo com d dimensões, onde "a melhor aproximação" se refere à minimização da soma dos quadrados das diferenças das entradas[1].

Utilizando uma nomenclatura usual em análise de votações legislativas, as coordenadas de cada parlamentar j retidas em $\mathbf{\Gamma_{(d)}}$ podem ser entendidas como o *ponto ideal* do parlamentar no espaço d-dimensional de preferências políticas.

Exemplificando para o caso comum em que $d = 2$, a equação $\mathbf{X}^* \approx \mathbf{R}^t_{(2)} \cdot \mathbf{\Gamma}_{(2)}$ foi reescrita abaixo:

$$\underset{\text{vot.}}{\begin{bmatrix} x^*_{11} & \cdots & x^*_{1M} \\ \vdots & \ddots & \vdots \\ x^*_{N1} & \cdots & x^*_{NM} \end{bmatrix}}_{\text{membros}} \approx \underset{\text{vot.}}{\begin{bmatrix} R_{11} & R_{21} \\ \vdots & \vdots \\ R_{1N} & R_{2N} \end{bmatrix}}_{\text{C.P.}} \cdot \underset{\text{C.P.}}{\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1M} \\ \gamma_{21} & \cdots & \gamma_{2M} \end{bmatrix}}_{\text{membros}}$$

[1] Em outras palavras, o modelo minimiza a norma de Frobenius da matriz de votações.

Centralização e Normalização

Em diversos contextos em que se aplica a ACP é comum realizar a *centralização* (subtraindo de cada entrada o valor médio da linha) e a *normalização* (multiplicando cada entrada por um fator de escala igual ao inverso da variância da linha, de forma a obter variância unitária para todas as direções da base original) de $\mathbf{X}$ antes de proceder à análise.

O algoritmo de determinação das componentes por SVD não é baseado na variância em si, e sim na soma dos quadrados. Para variáveis centralizadas as duas quantidades são proporcionais (vide equação 2), por isso a centralização é recomendável para variáveis que não possam ser supostas de média zero. No caso de votações legislativas a centralização introduz N parâmetros ao modelo (através dos valores L.I. da matriz $\mathbf{C}$), que podem ser interpretados como sendo relacionados aos tamanhos da maioria e minoria de cada votação.

Já a normalização é em geral recomendável quando as componentes originais possuem unidades de medida distintas, para evitar que dimensões com variâncias numericamente grandes predominem artificialmente. Como todas as votações possuem a mesma "escala", não se faz necessária a normalização. De fato, para o caso de uma votação quase unânime o fator de escala (1/variância) seria muito alto, pois a variância de uma votação quase unânime é baixa, e esta votação receberia um peso maior na composição das componentes principais apenas por ter sido menos acirrada.

Estas considerações sugerem a adoção da centralização, mas não da normalização, na análise de votações utilizando ACP.

Preditor

Para o modelo de classificação, define-se a matriz $\widehat{\mathbf{X}}$:

$$\widehat{\mathbf{X}} = \mathbf{R}^t_{(d)} \cdot \mathbf{\Gamma}_{(d)} + \mathbf{C} \tag{3}$$

$\widehat{\mathbf{X}}$ possui valores em $\mathbb{R}$ que se aproximam dos valores discretos da matriz de votos original $\mathbf{X}$.

Para $\widehat{x}_{ij} > 0$ o modelo prevê que o parlamentar j vota SIM na votação i; para $\widehat{x}_{ij} < 0$ o modelo prevê voto NÃO; e para $\widehat{x}_{ij} = 0$ o modelo prevê um voto arbitrário (para facilitar a reprodutibilidade dos resultados foi adotado SIM nestes casos).

Este modelo prevê apenas votos SIM ou NÃO, ou seja, não prevê a possibilidade de abstenções, obstruções ou ausências.

Escolha do Número de Dimensões d

O modelo será tanto mais preciso na classificação correta das votações quanto maior for o número de dimensões retidas $d \leq N$. Porém está claro que um modelo simples é mais útil: analisar cada uma do total de N dimensões seria tão trabalhoso quanto analisar individualmente cada uma das N votações (e tão completo quanto). O objetivo é simplificar, retendo o essencial da informação.

Uma forma de quantificar a informação retida (ou perdida) ao considerar apenas d dimensões é observar qual é a fração $\nu_d \leq 1$ da variância total explicada:

$$\nu_d = \frac{\sum_{i=1}^{d} \frac{M}{M-1} \langle {\gamma_{ij}}^2 \rangle_j}{\sum_{i=1}^{N} \mathrm{var}(i)}$$

onde o numerador é a soma da variância das d primeiras componentes principais, e o denominador é a variância total da matriz de votações.

Quanto maior for ν_d mais preciso será o modelo. Uma prática comum é adotar d tal que se fosse adotado $d+1$ o ganho em ν_d seria pequeno. Dito isso, o critério é arbitrário, e deve depender do objetivo da análise. Para uma visualização do aspecto geral de distribuição dos parlamentares é prático utilizar $d = 2$, já que assim a visualização no plano é muito mais simples. Seja qual for, a escolha deve vir acompanhada do valor de ν_d correspondente, afim de que se possa ter uma idéia de quanta informação está sendo desconsiderada.

3 Revisão da literatura

Leo ►*Um objetivo desta revisão é ser o mais didática possível, inclusive para não fluentes em matemática. Isso por si só pode ser uma boa contribuição para auxiliar novos estudantes de ciências políticas, inclusive da graduação, a entrarem nessa área. Esse objetivo deve ser dito aqui e também na introdução*◄ .

Leo ►*Talvez dizer que se preferiu uma análise mais aprofundada sobre os principais trabalhos da literatura, com espaço para trabalhos sobre o legislativo brasileiro, do que uma análise extensiva sobre tudo o que há na literatura. Essa abordagem corrobora o objetivo descrito acima*◄ .

%%%%

Primórdios... discrete choice model...

Outros métodos similares de escalonamento multidimensional foram propostos desde então, construindo sobre bases teóricas de análise de componentes principais que remontam ao início do século 20 com Pearson [8] e

Hotelling [9], e notando a equivalência destes com métodos utilizados em outras disciplinas, como em psicologia e educação na padronização de provas e análise estatística de resultados em testes de múltipla escolha.

3.1 Keith Poole: optimal classification e NOMINATE

Os trabalhos mais proeminentes publicados sobre a construção de mapas espaciais de votações são os de Keith Poole. Apresentaremos na sequência conceitos básicos utilizados por Poole na construção dos mapas especiais e os procedimentos utilizados para a construção dos mapas. Todos esses conceitos e procedimentos estão descritos detalhadamente no livro de Poole sobre modelos espaciais de votações [10].

O primeiro conceito é do que seria uma *votação perfeita*. Considere um mapa espacial de votações onde cada ponto representa um parlamentar e cada linha representa uma votação, de forma que pontos de um lado da linha representam parlamentares que votaram SIM, enquanto que os pontos do outro lado da linha representam parlamentares que votaram NÃO. Para uma pequena quantidade de votações pode ser possível uma construção perfeita de tal mapa, como, por exemplo, podemos observar na Figura 1.

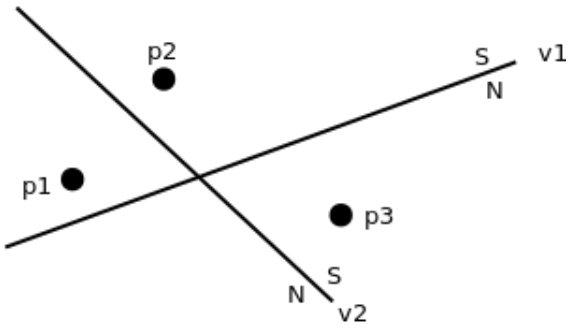

Figura 1: Mapa espacial de votação com classificação perfeita.

No exemplo fornecido, podemos ver pelo mapa que o parlamentar $p1$ votou SIM na votação $v1$ e NÃO para a votação $v2$. Já $p2$ votou SIM para $v1$ e SIM para $v2$. Por fim, $p3$ votou NÃO para $v1$ e SIM para $v2$. Como a partir do mapa descrevemos perfeitamente o comportamento dos parlamentares, dizemos que trata-se de uma votação perfeita.

Mas conforme o número de votações e de parlamentares cresce, percebe-se que é impossível posicionar perfeitamente todos os pontos em relação a todas as retas. Por isso é importante entender que o mapa de votações não é construído de forma a descrever perfeitamente o comportamento dos parla-

mentares. Em vez disso, o mapa de votações tenta maximizar a quantidade de classificações corretas.

Na concepção dos trabalhos de Poole, considera-se que em um mapa de votações cada parlamentar possui seu *ponto ideal* no espaço. Nesse mesmo espaço, uma votação também possui pontos associados às suas possíveis opções (SIM e NÃO). A reta que representa a votação na Figura 1 é construída em função desses pontos que representam a votação. Dessa forma, o comportamento de um parlamentar numa dada votação é função de relações entre seu ponto ideal e a representação espacial da votação.

Dados os pontos ideias associados a parlamentares e votações, uma primeira abordagem simplista para determinar o voto do parlamentar em uma votação seria dizer que o parlamentar vota deterministicamente na opção mais próxima de seu ponto ideal. Mas em vez disso, Poole utiliza o conceito de *função utilidade*, que atribui a cada ponto no espaço um valor. Quanto mais alto esse valor, maior é a *probabilidade* de que o parlamentar vote na opção associada a esse ponto do espaço.

Duas premissas importantes são usualmente aplicadas às funções utilidade: 1) as funções são de pico único (i.e., possuem apenas um ponto de valor máximo), sendo esse pico localizado no ponto ideal do parlamentar; 2) a função é simétrica, ou seja, o parlamentar é indiferente a duas opções igualmente distantes de seu ponto ideal.

ToDo ►*Fazer figura ilustrativa?*◄

A função utilidade possui uma parcela determinística e também uma parcela estocástica, que possibilita a modelagem de *erros de votação*. Um erro de votação seria a ideia de que o parlamentar não votou de acordo com suas preferências políticas. Um erro pode ter acontecido no sentido de que o parlamentar pode ter avaliado erroneamente a localização espacial das opções de uma dada votação. Mas o erro pode refletir também o fato de que fatores subjacentes não captados pelo modelo foram decisivos na determinação da opção escolhida.

Dados os conceitos básicos apresentados (pontos ideias, funções utilidade etc.) vamos descrever agora os principais métodos de construção de mapas espaciais de votações elaborados por Poole. São eles o *Optimal Classification* e o NOMINATE.

O *Optimal Classification* (OC) consiste em um processo iterativo[2] que procura maximizar a proporção de classificações corretas em um mapa espacial de votações. Dada uma configuração inicial de um mapa de votações,

[2]Um processo iterativo é aquele no qual um algoritmo é repetido várias vezes, sendo que depois de uma certa quantidade de repetições há uma *convergência*, ou seja, o resultado não mais se altera com mais repetições do processo.

primeiramente aplica-se um algoritmo que maximiza a classificação correta fixando os pontos e movendo os linhas. Em um segundo passo, fixa-se as linhas e move-se os pontos para maximizar a classificação correta. Esses dois passos são repetidos várias vezes até que o erro (proporção de classificações incorretas) estabilize. Os algoritmos empregados garantem que a cada passo o erro nunca aumenta.

O OC não define uma posição exata dos parlamentares no mapa de votações, assim como não define uma distância exata entre dois dados parlamentares. O que o OC fornece são regiões do espaço nas quais os parlamentares podem ser posicionados. Essas regiões são denominadas de politopos e representam padrões de opções escolhidas nas votações. Voltemos à Figura 1: note que se alterarmos ligeiramente a posição de um parlamentar, digamos $p1$, seu padrão de opções escolhidas não se altera. Esse padrão se mantêm enquanto o ponto ideal não atravessar uma das retas que representam as votações. Dessa forma, dizemos que essa região do espaço delimitada pelas retas que mantêm o padrão de escolhas de $p1$ é o seu politopo.

Após a construção do mapa podemos observar que normalmente alguns parlamentares caem do lado errado de algumas retas. Isso significa que se o leitor do mapa fosse reconstituir os votos dados pelos parlamentares em cada votação, ele se enganaria em alguns casos. Esses erros representam uma imperfeição do mapa construído. A literatura costuma apresentar esses erros como uma incapacidade do modelo obtido em *predizer* corretamente os resultados de algumas votações. É preciso ficar atento com o uso do termo *predição*, pois os mapas de votações ou funções utilidades obtidos não serão utilizados na tentativa de predizer o resultado de votações futuras, ou mesmo votações passadas que não foram utilizadas para a produção do mapa de votações.

Para a utilização do OC, não é preciso premissas sobre a distribuição da função utilidade. Apenas considerar que ela é simétrica e de pico-único. Embora o conceito de função utilidade não apareça diretamente na aplicação do algoritmo, ele é importante para explicar os erros de classificação, no sentido de que há uma certa probabilidade de que o legislador vote na opção contrária do que o mapa de votações indica.

Como exemplo de aplicação, Poole mostra os resultados do OC quando aplicado à votação de revogação das leis do milho na Casa dos Comuns do parlamento inglês em 1846. Nessa situação, o algoritmo apresentou uma taxa de classificação correta de 95,2% para os 430 parlamentares que votaram nessa matéria. Leo ►*hum, que estranho o exemplo dado de apenas uma votação... acho q OC deve ser interessante quando aplicado a várias votações de uma vez.*◄

O outro método de construção de mapas espaciais de votações consagrado por Poole é o NOMINATE, que constitui na verdade uma família de

algoritmos com ligeiras variações entre si. No NOMINATE temos algumas premissas a mais sobre as funções utilidade. A mais importante é de que a função utilidade é exponencial. Outra opção utilizada em outros trabalhos [4] são funções quadráticas. Essa diferença diz respeito a como o parlamentar vai se comportar em relação a opções cada vez mais longes de seu ponto ideal. Com funções exponencias, as opção cada vez mais longe se tornam cada vez mais indistinguíveis, enquanto que na exponencial elas se tornam cada vez pior avaliadas.

ToDo ►*Fazer figura ilustrativa de função exponencial vs função quadrática.*◄

Agora vamos detalhar a função utilidade utilizada no NOMINATE [2]. Considere a localização x de um parlamentar e a localização o de uma opção de uma votação. Essas localizações são pontos em um espaço multidimensional, onde cada dimensão representa preferências sobre um determinado tema político. Para uma votação temos o_y, a localização da opção SIM, e o_n, a localização da opção NÃO.

A função utilidade utilizada pelo NOMINATE [2] é

$$U(x, o) = \beta * e^{\frac{-w*d^2}{2}} + \varepsilon,$$

onde d representa a distância entre os pontos x e o, dada por $|x - o|$. Quanto mais perto z está de o, maior o valor de U, sendo U simétrica e de pico único. β e w são parâmetros da função utilidade, sendo β o fator de ruído, que determina o peso da parcela determinística da função utilidade. ε é a parcela estocástica, que representa erros distribuídos independentemente de conforme a distribuição logística. Poole afirma que teoricamente a distribuição normal para ε seria mais adequada. No entanto, devido às limitações computacionais da época, Poole optou pela distribuição logística, que é mais simples do ponto de vista computacional e razoavelmente similar à função normal.

No NOMINATE temos então três grupos de parâmetros: pontos ideias dos parlamentares (um x para cada parlamentar), pontos ideais das opções das votações envolvidas (um o_n e um o_y para cada votação) e os parâmetros da função utilidade (β e w). O objetivo do NOMINATE é encontrar os valores de todos esses parâmetros para que se possa desenhar o mapa espacial de votações.

Dada uma configuração inicial de pontos ideias de legisladores e opções de votações, o NOMINATE aplica sucessivamente os três seguintes passos: 1) estima-se os parâmetros da função utilidade com base nos parâmetros dos legisladores e das votações; 2) estima-se os parâmetros dos legisladores com base nos parâmetros das votações e da função utilidade; 3) estima-se os parâmetros das votações com base nos parâmetros dos legisladores e da

função utilidade. Cada um desses passos possui seu próprio algoritmo com suas complexidades. Os três passos são repetidos até a convergência, que é quando o refinamento para de ter efeito e os mesmos valores são produzidos.

Tanto o OC quanto o NOMINATE partem de uma *configuração inicial* do mapa de votações a ser refinada. Poole define essa configuração inicial pelo seguinte processo: 1) constrói-se uma matriz de concordância entre parlamentares; 2) transforma-se os valores de concordância em distâncias quadráticas; 3) centraliza-se duplamente a matriz de distâncias quadráticas; 4) realiza-se sobre essa matriz uma decomposição de autovetores-autovalores. Leo ►*O que comentar sobre esse processo maluco? Alguma relação com o PCA?*◄

Leo ►*OC produz apenas politopos, mas NOMINATE produz distâncias... comentar sobre isso?*◄

ToDo ►*Resultados obtidos por Poole com o NOMINATE. Acho que não achei isso no livro.*◄

Para a valiar os resultados de um modelo como o NOMINATE podemos usar a taxa de classificação correta, que nos informa a porcentagem de acertos e erros do modelo. Mas os cientistas políticos costumam usar uma outra medida, a PRE (redução proporcional do erro). Isso porque modelos extremamente ingênuos poderiam obter boas taxas de classificação correta. Exemplo: o modelo pode prever que todos os parlamentares votam na opção vencedora. Dessa forma, numa votação em que 90% dos legisladores votaram com a maioria, o modelo ingênuo teria apenas 10% de erro. Já a PRE mede em quantos porcentos o erro foi reduzido do modelo ingênuo para o modelo avaliado. Ou seja,

$$PRE = \frac{\text{votos da minoria} * \text{erros do modelo}}{\text{votos da minoria}}.$$

A fórmula da PRE considera apenas uma votação. Para se avaliar um conjunto de votações, se utiliza a APRE (redução proporcional do erro agregado), onde a quantidade de votos na minoria e a quantidade de erros são consideradas para todas as votações.

3.2 Aplicações dos trabalhos de Poole ao congresso brasileiro

Aplicando o W-NOMINATE, uma das variações do NOMINATE, à Câmara dos Deputados, Leoni [5] encontra taxas de classificação correta de 86,4% e 90,4% para as 49ª e 50ª legislaturas respectivamente. Já em termos de APRE, as taxas encontradas são de 52,3% e 64,8% para as mesmas legislaturas.

Izumi [11] elabora mapas espaciais de votações para o Senado brasileiro. No entanto, ele argumenta que o NOMINATE pode não ser adequado por

causa dos pressupostos envolvendo a função utilidade. O primeiro pressuposto questionado é a simetria da função utilidade, uma vez que, por exemplo, a redução em 5% dos impostos pode ser algo muito mais importante para um parlamentar do que um aumento da mesma magnitude.

O segundo pressuposto questionado da função utilidade é o de que os erros são independentes e identicamente distribuídos entre os legisladores e as votações. Argumentos contrários a esse pressuposto: 1) existem partidos mais coesos que outros (o PT, por exemplo, costuma ser bem mais coeso que outros partidos); 2) migrações de parlamentares entre partidos e migrações de partidos para dentro ou fora da base governista podem alterar a variação de erros ao longo do tempo; 3) em determinados contextos o voto estratégico pode prevalecer sobre o voto sincero. Interessante notar aqui como os argumentos 1) e 2) são características que diferenciam o estudo do legislativo brasileiro do legislativo norte-americano.

Embasado por esses questionamentos, Izumi prefere adotar o Optimal Classification, pois este não se sustenta sobre pressupostos da função utilidade ou da distribuição de erros. Dessa forma, aplicando o OC a seis legislaturas do Senado (48ª a 53ª), Izumi encontra taxas de classificação correta entre 90,7% e 98,5%. Já em termos de APRE, as taxas encontradas são entre 50% e 94%. Embora o trabalho de Izumi se aplique ao Senado, cabe notar aqui que todos seus argumentos também se aplicariam à Câmara dos Deputados.

Cabe notar que parte das críticas de Izumi ao NOMINATE se aplicam também ao OC. Como disse Poole [12], as únicas premissas para a aplicação do OC são 1) o espaço de escolha é euclideano; e 2) preferências são simétricas e de pico-único. Como o próprio Izumi disse, a simetria é uma premissa que pode ser questionada na prática, e essa não somente para o caso brasileiro.

Leo ►*Com as considerações do Izumi, talvez não seja mais tão interessante comparar o PCA com o nominate... será q faz sentido comparar com o OC?*◄

3.3 Modelos lineares de Heckman e Snyder

Muitos trabalhos consideram modelos lineares para a análise de votações nominais menos adequados do que os modelos não lineares, como o nominate de Poole. Heckman e Snyder [13], porém, demonstram rigorosamente a equivalência de modelos lineares com os resultados obtidos por Poole, tendo como principal vantagem a simplicidade e eficiência computacional dos métodos lineares.

Para Heckman e Snyder, a decisão de se votar SIM ou NÃO em uma votação é modelada como o resultado de um processo de escolha racional no qual os legisladores usam suas preferências para ponderar sobre as carac-

terísticas da votação. Assim, dizemos que uma opção é localizada em um espaço no qual cada dimensão seria uma característica da opção. Nesse caso, a função de utilidade recebe um vetor no espaço de características e retorna um número real. Cada legislador teria sua própria função utilidade e escolheria a opção que resultasse no maior valor da função utilidade. Diferentemente do NOMINATE, temos aqui uma função utilidade quadrática.

A função utilidade de Heckman e Snyder, assim como a do NOMINATE, é incrementada com uma parcela de erros aleatórios. Considera-se que uma das fontes de erro seja a dificuldade para o parlamentar estimar o valor de cada característica.

Os autores utilizam diferentes métodos algébricos para estimar as preferências dos legisladores. Um dos métodos mais simples utilizados é justamente a Análise de Componentes Principais (ACP). Leo ►*Elaborar relação entre o PCA e a função utilidade*◄

3.4 Dimensionalidade dos mapas de votações

Um debate existente na literatura é sobre a quantidade de dimensões necessárias para representar um mapa espacial de votações. De acordo com McCarty [14], essa é uma polêmica sem conclusão definitiva.

Nos estudos sobre o NOMINATE, Poole afirma que para o congresso americano, na maior parte do tempo Leo ►*quando?*◄ uma dimensão explica suficientemente bem Leo ►*quantificar*◄ o comportamento dos legisladores. Essa dimensão estaria ligada ao espectro político-ideológico presente nos EUA, possuindo uma escala que vai do extremo liberal ao extremo conservador. No entanto, em determinado período da história dos EUA Leo ►*quando?*◄ é preciso a utilização de duas dimensões. A segunda dimensão estaria ligada principalmente a questões associadas a direitos civis, tendo uma correlação com a região dos parlamentares (norte ou sul dos EUA).

Já para Heckman e Snyder, existem pelo menos cinco dimensões significativas. Essas dimensões a mais podem não fazer tanta diferença na taxa global de sucesso de classificação, mas são decisivas em algumas votações específicas, por representarem substantivamente assuntos específicos. Exemplos de tais assuntos: direitos civis e eleitorais, agricultura, ajuda internacional, gasto militar, teto da dívida, água, aborto e reforma do congresso.

Utilizando o W-NOMINATE, Leoni [5] chegou à conclusão de que uma dimensão explica a maior parte das votações na Câmara dos Deputados, pois dimensões adicionais não melhoram significativamente a capacidade explicativa do modelo.

3.5 Discussão

Mesmo Heckman se apoia em uma teoria profunda sobre discrete choice model, porém acreditamos que independente de um modelo que explique o processo de tomada de decisão de parlamentares, a ACP é útil enquanto análise de conjuntura de uma casa legislativa que de fato aconteceu. Assim, tem-se uma figura simplificada para facilitar a análise de tal conjuntura.

Além disso, tanto Poole quanto Heckman adotam uma postura de análise preditiva. Embora seja possível avaliar a qualidade do modelo com base na capacidade de reconstrução dos dados originais a partir do mapa de votações, os autores usam um tom que pode confundir um leitor desavisado, dando a impressão de que se trata da tentativa de predizer o resultado de votações futuras, ainda mais quando tais modelos tentam explicar o modelo de escolha de opção do legislador com base em atributos da opções de escolha e na distribuição de preferências do legislados no espaço de possibilidade.... Nesse sentido, de fato, se se soubesse os atributos de cada votação a priori, e as funções de utilidade de cada parlamentar, poderíamos prever o resultado das votações. Mas na prática, temos que em função dos resultados observáveis é que podemos estimar coisas como funções de utilidade dos parlamentares ou os atributos das votações. Dessa forma, queremos apenas destacar que ao descartar a ambição de se ter um modelo preditivo, mais ainda a ACP se mostra conceitualmente útil em prover uma ferramenta que auxilie o cientista político na análise de conjuntura de uma casa legislativa por meio de um mapa espacial de votações. De outra forma... os trabalhos revisados se apoiam em uma teoria de escolha racional do voto. Os pressupostos da teoria podem ser fortes. Nosso tipo de análise dispensa a teoria. Trata-se de uma fotografia que resume o que aconteceu.

4 Decisões de nossa abordagem

Dizer que utilizamos o PCA e como modelamos nossa matriz de entrada conta apenas parte da história. Ao utilizar o PCA diversas outras pequenas decisões devem ser feitas. Nesta seção tratamos dessas decisões, assim como algumas variações do tipo de análise. Algumas dessas decisões são feitas em função das características do sistema político brasileiro, em contraste com o sistema político do EUA, usalmente abordado na literatura.

Análise por partido

No modelo apresentado, nada impede que os valores de $\mathbf{X}$ possuam valores reais, situados por exemplo no intervalo [-1;1], em vez de apenas os valores

discretos {-1;0;1}. Esta observação permite uma extensão direta do modelo para analisar os parlamentares agregados por partido em vez de considerá-los individualmente, bastando considerar o voto médio do partido em cada votação antes de iniciar a análise.

O voto médio do partido k na votação i é definido por:

$$x_{ik} = \frac{1}{|k|} \sum_{j \in k} x_{ij} \tag{4}$$

onde $j \in k$ denota que o parlamentar j pertence ao partido k, e $|k|$ é o número de parlamentares do partido k considerados.

Esta análise é útil para analisar afinidades partidárias e coalizões em ambientes com vários partidos, como é tipicamente o caso das casas legislativas no Brasil.

No Radar Parlamentar, porém, não aplicamos essa técnica, pois para esse software optamos pela coexistência de partidos e parlamentares individuais no mesmo mapa especial. O método proposto acima é válido, mas seu resultado não possui relação direta com o resultado da análise por parlamentar. Assim sendo, no Radar Parlamentar os partidos são posicionados no centroide das posições ocupadas por seus parlamentares. Segue fórmula da hora...

Tratamento de valores faltantes

Todos os métodos de análise de votações legislativas encontrados na literatura revisada descartam ausências e abstenções antes de iniciar a análise, considerando explicita- ou implicitamente que tais atitudes não trazem informação acerca das preferências políticas do legislador, e notando que tais situações representam a minoria dos casos. Por exemplo, Heckman e Snyder notam que as abstenções representam menos de 1% dos votos para câmara e senado estadounidenses, e as assumem aleatórias em relação a resultados das votações e a preferências dos parlamentares [13, p.40].

Já 53a legislatura da Câmara dos Deputados brasileira (período de 2003 a 2006) soma para votações nominais abertas cerca de 6% entre abstenções e obstruções, e as ausências são próximas de 49%. Propõe-se que o comportamento observado de ausentar-se ou abster-se de uma votação traz sim informação acerca das preferências do parlamentar que se deseja estimar, e por isso essa informação não deve ser descartada na análise.

No modelo proposto, ausência, abstenção e obstrução são modeladas através do valor 0. Em relação à alternativa de descartar estas situações, que serão referidas genericamente como votos "nulos", esta modelagem introduz um viés no sentido oposto ao voto médio dos que realmente votaram.

Ou seja, supondo sem perda de generalidade que o voto da maioria é sempre SIM, o voto médio dos que votaram será sempre maior que zero, e se o parlamentar faz voto nulo sua preferência nesta votação será modelada como sendo ligeiramente oposta ao SIM (pois seu voto é numericamente menor do que a média), mas não tão oposta quanto se o parlamentar tivesse efetivamente votado NÃO.

Este viés pode parecer arbitrário, porém esta abordagem é consistente tanto com a idéia de que um voto nulo representaria uma indiferença do parlamentar quanto aos resultados SIM e NÃO (o voto nulo é euclidianamente equidistante das duas alternativas) quanto da idéia de que ao não votar o parlamentar pode ter uma preferência contrária àquela que se imagina que será aprovada na votação, como em um "boicote" pessoal (ou em grupo) à votação. Em outras palavras, um parlamentar teria maior tendência de comparecer e não se abster nem obstruir a votação em propostas nas quais ele esteja inclinado a votar com a maioria. Se estas hipóteses são arbitrárias, pode-se dizer que são pelo menos tão arbitrárias quanto a alternativa de considerar que um voto nulo equivale a um parlamentar com preferência igual à preferência média da casa. Nossos resultados sugerem que de fato a forma proposta de modelagem melhora os índices de classificação correta.

No caso da análise por partidos, ao excluir votos nulos do cálculo da média na equação 4 estaria-se buscando considerar que a opinião "do partido" é composta apenas pela opinião daqueles que votaram ou SIM ou NÃO. Outra opção é excluir apenas as ausências, se os dados permitirem discriminar esta opção. Os resultados aqui apresentados não excluem estes votos, para que a análise reflita o fato de que uma abstenção ou mesmo uma ausência não são equivalentes a concordar com a opinião geral do partido. Além disso a análise fica mais simples, já que não há necessidade de tratamento especial de partidos que tenham estado por exemplo cem porcento ausentes em uma dada votação.

Tratamento de votações unânimes

Lidando com migração partidária

Citar Izumi

Análise temporal

O objetivo de nossa análise temporal não é que se possa comparar as posições de um parlamentar/partido ao longo do tempo, mas que se possa comparar as distâncias relativas desses elementos entre si ao longo do tempo.

Efeito de extremistas na visualização temporal: um extremista de um período afeta a visualização de outros períodos (deixando parlamentares muito concentrados). Procede ???

Leo ►*Comparar em linhas gerais nosso método de análise temporal com o D-NOMINATE (dynamic nominate).*◄

Heckman encontra uma alta correção entre as coordenadas de um legislador ao longo do tempo (e não são apenas para as duas primeiras coordenadas... chega-se a até 6 ou 8 fatores com correlação ao longo do tempo). Mas duvido que no caso brasileiro essas correlações seriam altas.

5 Medidas de Adequação (Fitness)

Aqui, explicar os indicadores e tal.

6 Resultados

Mostrar resultados de indicadores e gráficos.

As análises foram feitas no software de estatística **R** [3] Como referência para *benchmarking* foi adotado o algoritmo WNOMINATE, através do pacote *wnominate* para **R** [4].

...

Heckman e Snyder [13] encontram padrões políticos na distribuição do mapa espacial que são compatíveis com análises já feitas por cientistas políticos utilizando outros métodos. These findindgs illustrate the "reasonableness" of estimates obtained from our model. Vamos na mesma linha... mostrar que nossos mapas obtidos fazem sentido do ponto de vista político.

...

Problema da taxa de classificação correta no PCA: para modelar as votações, mapeando opções em valores numéricos, tudo bem. Mas o cálculo da taxa de classificação correta depende do mapeamento inverso, dado um valor entre -1 e 1 definir uma das opções (SIM, NÃO ou ABSTENÇÃO). Definir os valores limiares desse mapeamento inverso é arbitrário, sendo que essa escolha arbitrária pode alterar drasticamente os valores obtidos.

Em vez de se valer da taxa de correta classificação, uma medida melhor para avaliar o resultado do mapa espacial gerada pela ACP é o porcentual

[3] O código dos scripts utilizado está disponível em `https://github.com/leonardofl/radar_parlamentar` sob licensa *AGPL v3*

[4] O pacote wnominate pode ser encontrado na *Comprehensive R Archive Network*, no endereço `http://CRAN.R-project.org/package=wnominate`

da variância dos dados explicados pelas dimensões utilizadas no mapa. No entanto, essa medida tem a desvantagem de não ser diretamente comparável com outros métodos.

7 Discussão

Comparar pros e contras dos algoritmos, e contextualizar (ferramenta web, características brasil, vantagens políticas de análise mais transparente e simples etc.)

8 Conclusões

resultados coisas pra melhorar / investigar ex: modelagem de partidos q não votam em alguma votação; suplentes; troca de parlamentares análise de sensibilidade

Referências

[1] A. Downs, "An economic theory of political action in a democracy," *The Journal of Political Economy*, vol. 65, no. 2, pp. 135–150, 1957.

[2] K. T. Poole and H. Rosenthal, "A spatial model for legislative roll call analysis," *American Journal of Political Science*, vol. 29(2), pp. 357–384, 1985.

[3] K. T. Poole and H. Rosenthal, *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press, USA, Nov. 2000.

[4] J. Clinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," *American Political Science Review*, vol. 98, no. 02, pp. 355–370, 2004.

[5] E. Leoni, "Ideologia, democracia e comportamento parlamentar: a Câmara dos Deputados (1991-1998)," *Dados*, vol. 45, pp. 361 – 386, 2002.

[6] M. Kantardzic, "Section 3.4 Principal Componente Analysis," in *Data Mining, Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, INC., 2003.

[7] G. H. Golub and C. F. van Loan, *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. The Johns Hopkins University Press, 3rd ed., Oct. 1996.

[8] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine (6)*, vol. 23, pp. 559–572, 1901.

[9] H. Hotelling, "Analysis of a complex of statistical variables into principal components," 1933.

[10] K. T. Poole, *Spatial models of parliamentary voting*. Cambridge University Press, 2005.

[11] M. Y. Izumi, "Governo e oposição no senado brasileiro," *DADOS – Revista de Ciências Sociais*, vol. 59, pp. 91–138, 2016.

[12] K. T. Poole, "Nonparametric unfolding of binary choice data," *Political Analysis*, vol. 8, no. 3, pp. 211–237, 2000.

[13] J. J. Heckman and J. M. Snyder, "Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators," *The RAND Journal of Economics*, vol. 28, pp. S142–S189, 1997.

[14] N. McCarty, "Measuring legislative preferences," in *The Oxford handbook of the American Congress*, pp. 66–94, Oxford University Press Oxford, 2011.