

# 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:

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

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  $\xi_{ij}$  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  $\xi_{ij}$ .

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

$$\mathbf{X} = \begin{array}{c} \begin{array}{c} \text{votações} \\ \downarrow \end{array} \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ N \end{array} \left[ \begin{array}{ccccc} & \xrightarrow{\text{membros}} & & & \\ & 1 & j & & M \\ \begin{array}{c} x_{11} \quad \dots \quad x_{1j} \quad \dots \quad x_{1M} \\ \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ x_{i1} \quad \dots \quad x_{ij} \quad \dots \quad x_{iM} \\ \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ x_{N1} \quad \dots \quad x_{NM} \quad \dots \quad x_{NM} \end{array} \end{array} \right] \end{array}$$

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:

$$x_{ij}^* = x_{ij} - \langle x_{ij} \rangle_j \quad (1)$$

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:

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

de forma que:

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

A variância (amostral)  $\text{var}(i)$  de cada votação, ou dimensão é:

$$\begin{aligned}\text{var}(i) &= \frac{\sum_{j=1}^M \left(x_{ij} - \langle x_{ij} \rangle_j\right)^2}{M-1} = \frac{M}{M-1} \left(\langle x_{ij}^2 \rangle_j - \langle x_{ij} \rangle_j^2\right) \\ \text{var}(i) &= \frac{M}{M-1} \langle x_{ij}^{*2} \rangle_j\end{aligned}\quad (2)$$

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  $\mathbf{R}_{(d)}^t \cdot \mathbf{\Gamma}_{(d)}$  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<sup>1</sup>.

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  $\mathbf{X}^* \approx \mathbf{R}_{(2)}^t \cdot \mathbf{\Gamma}_{(2)}$  foi reescrita abaixo:

$$\begin{array}{c} \text{membros} \\ \left[ \begin{array}{ccc} x_{11}^* & \cdots & x_{1M}^* \\ \vdots & \ddots & \vdots \\ x_{N1}^* & \cdots & x_{NM}^* \end{array} \right] \\ \text{vot.} \end{array} \approx \begin{array}{c} \text{C.P.} \\ \left[ \begin{array}{cc} R_{11} & R_{21} \\ \vdots & \vdots \\ R_{1N} & R_{2N} \end{array} \right] \\ \text{vot.} \end{array} \cdot \begin{array}{c} \text{membros} \\ \text{C.P.} \left[ \begin{array}{ccc} \gamma_{11} & \cdots & \gamma_{1M} \\ \gamma_{21} & \cdots & \gamma_{2M} \end{array} \right] \end{array}$$

<sup>1</sup>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  $\hat{\mathbf{X}}$ :

$$\hat{\mathbf{X}} = \mathbf{R}_{(d)}^t \cdot \mathbf{\Gamma}_{(d)} + \mathbf{C} \quad (3)$$

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

Para  $\hat{x}_{ij} > 0$  o modelo prevê que o parlamentar  $j$  vota SIM na votação  $i$ ; para  $\hat{x}_{ij} < 0$  o modelo prevê voto NÃO; e para  $\hat{x}_{ij} = 0$  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:

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

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  $\nu_d$  mais preciso será o modelo. Uma prática comum é adotar  $d$  tal que se fosse adotado  $d+1$  o ganho em  $\nu_d$  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  $\nu_d$  correspondente, afim de que se possa ter uma idéia de quanta informação está sendo desconsiderada.

## 3 Revisão de outros métodos

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.

Primórdios... discrete choice model...

Os trabalhos mais proeminentes publicados na área são os de Poole... optimal classification, nominate...

Problemas do nominate...

Izumi [10] utiliza o OC de Poole para analisar o senado brasileiro, e encontra que... justifica porque o NOMINATE não seria adequado para analisar o congresso brasileiro, devido às fortes premissas do modelo.

**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?* ◀

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 [11], 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.

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 prever o resultado de votações futuras, ainda mais quando tais modelos tentam explicar o modelo de escolha de opção do legislador com base em atributos das opções de escolha e na distribuição de preferências dos legisladores 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.

Controvérsias sobre quantidade de dimensões

## 4 Peculiaridades de nossa abordagem considerando o contexto brasileiro

### 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:

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

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 estadunidenses, e as assumem aleatórias em relação a resultados das votações e a preferências dos parlamentares [11, 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 por cento ausentes em uma dada votação.

## 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**<sup>2</sup> Como referência

---

<sup>2</sup>O código dos scripts utilizado está disponível em [https://github.com/leonardofl/radar\\_parlamentar](https://github.com/leonardofl/radar_parlamentar) sob licença *AGPL v3*

para *benchmarking* foi adotado o algoritmo WNOMINATE, através do pacote *wnominate* para **R** <sup>3</sup>.

## 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, 00 2002.
- [6] M. Kantardzic, “Section 3.4 Principal Component Analysis,” in *Data Mining, Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, INC., 2003.

---

<sup>3</sup>O pacote *wnominate* pode ser encontrado na *Comprehensive R Archive Network*, no endereço <http://CRAN.R-project.org/package=wnominate>

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