Parlamentares no Radar Parlamentar

Maratona Hacker – Câmara dos Deputados – 2013

Saulo Trento, Leonardo Leite

1 Introdução

O *Radar Parlamentar* é uma aplicação que permite a visualização gráfica das afinidades entre partidos políticos em função de uma análise matemática sobre como partidos votam em uma dada casa legislativa (ex: Câmara dos Deputados). É exibido um gráfico em que cada partido é representado por uma circunferência, sendo que quanto mais afinidade dois partidos possuem, mais próximos eles estarão no gráfico.

A versão anterior do Radar Parlamentar (utilizada como base para as novas funcionalidades introduzidas pelo projeto "Parlamentares no Radar Parlamentar") se encontra disponível em: `http://radarparlamentar.polignu.org/`. Já versão apresentada na *Maratona Hacker 2013 da Câmara dos Deputados* está em `http://hackathona.radarparlamentar.polignu.org`.

2 "Parlamentares no Radar Parlamentar"

O objetivo do projeto "Parlamentares no Radar Parlamentar" inscrito na *Maratona Hacker 2013 da Câmara dos Deputados* é fazer com que o *Radar Parlamentar* exiba também um gráfico de afinidade entre parlamentares. Dessa forma o cidadão observará não somente um gráfico com partidos, mas também com parlamentares.

3 Contribuições

O *Radar Parlamentar* fornece uma interessante ferramenta de análise de conjuntura das casas legislativas e suporte à decisão de voto do cidadão nas eleições legislativas. A nova funcionalidade implementada no *Radar* pelo nosso projeto acrescenta mais uma importante dimensão: o acompanhamento do mandato de um parlamentar. Assim um eleitor irá facilmente identificar seu parlamentar em uma conjuntura macro e poderá responder a perguntas como "meu parlamentar costuma ser fiel ao seu partido?", "se não, a que outros partidos ele se aproxima mais?". Com futuras funcionalidades do *Radar* poderemos também verificar a posição do parlamentar em temáticas mais específicas, como "meio ambiente", por exemplo.

Outra melhoria implementada durante a Hackaton da Câmara dos Deputados foi a apresentação de bolhas com tamanhos diferentes, em função do tamanho do partido, com a área da bolha proporcional ao tamanho da bancada durante o período sob análise.

Por fim, a documentação do *Radar Parlamentar* foi incrementada, notadamente com a criação de um documento que explica detalhadamente o algoritmo matemático desenvolvido pela equipe para a análise dos dados (vide anexo). Tal documentação é fundamental para que a validade do método possa ser julgada por quaisquer interessados, como estatísticos, cientistas políticos, jornalistas, etc.

Sobre o suporte à decisão do voto do cidadão nas eleições legislativas, o *Radar* já fornece um subsídio ao eleitor para verificar o "perfil" do partido no qual ele pretende votar. Com a nova funcionalidade, esse subsídio será mais refinado, uma vez que poder-se-á verificar diretamente o comportamento do parlamentar, possibilitando uma diferenciação entre candidatos de um mesmo partido.

4 Implementação e Interação com o Usuário

Um primeiro requisito importante é que continue sendo possível a visualização por partidos. Considerando isso e a grande quantidade de parlamentares na Câmara dos Deputados, a visualização de todos os elementos ao mesmo tempo poderia ser confusa.

Então definimos o seguinte modelo de interação do usuário com a aplicação: o gráfico é inicialmente carregado com os partidos, assim como na versão antiga. O usuário pode então clicar sobre um partido para "explodi-lo", o que fará com que a circunferência do partido desapareça dando lugar às circunferências dos parlamentares daquele partido. Quando o usuário clicar sobre a circunferência de um parlamentar, o partido será "implodido", e a circunferência do partido voltará a aparecer.

Esse modelo de interação permite um controle maior do usuário sobre a quantidade de elementos no gráfico, de forma a torná-lo menos poluído e mais inteligível.

O anexo "O Modelo ACP para Análise de Votações" traz informaçãoes detalhadas sobre como as posições de parlamentares e partidos no gráfico são calculadas.

5 Dados Utilizados

O *Radar Parlamentar* possui módulos dedicados a realizar a importação de dados abertos de votações que estejam disponíveis na Internet. Atualmente o *Radar* possui importadores de votações da Câmara dos Deputados, do Senado Federal, e da Câmara Municipal de São Paulo.

Como a estrutura dos dados disponibilizados por cada casa pode ser ligeiramente diferente, é tarefa do importador não apenas automatizar as solicitações de dados ao webservice como também adequar os dados ao modelo de dados do *Radar*. Por esse motivo, um importador dedicado deve ser desenvolvido para cada casa, porém uma vez que isso tenha sido feito torna-se automaticamente possível fazer todos os tipos de análise já implementadas. Nesse sentido gostamos de dizer que o *Radar Parlamentar* é facilmente extensível para analisar dados abertos de votações parlamentares de outras casas.

Nosso importador de dados da Câmara dos Deputados automatiza o processo de montar nossa base de dados acessando os seguintes web services disponibilizados pela casa: Deputados e Proposicoes. A informação mais importante obtida pelos web services é o voto dado por cada deputado em cada uma das votações realizadas na casa. No intuito de informar quais foram os dados utilizados para gerar o gráfico visualizado no *Radar Parlamentar*, nós disponibilizamos links para os dados em formato SQLite.

6 Tecnologias

Front-end A principal tecnologia de front-end envolvida neste projeto foi a utilização da biblioteca javascript d3js para a geração de um gráfico animado em SVG.

Back-end O projeto é codificado em Python, utilizando o framework web Django e banco de dados MySQL. Uma das principais ferramentas para a análise matemática é a biblioteca numpy.

ANEXO: O Modelo ACP para Análise de Votações

O objetivo deste anexo é dar detalhes sobre como são feitas as análises matemáticas de votações legislativas.

A análise de votações do *Radar Parlamentar* é baseada na análise de componentes principais (ACP), que é um método cujos resultados são de interpretação relativamente simples, e que é muito conhecido e utilizado em análise exploratória de dados. Esta seção introduz a ACP e como ela se aplica no contexto do *Radar*.

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 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 $\mathbf{X}$:

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

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:

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

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

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*.

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 $\mathbf{\Gamma}_{(\mathbf{d})}$ 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:

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

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 utilizado para determinação das componentes (que utiliza a teoria de decomposição de valores singulares) 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.

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

Precisão do Modelo em Duas Dimensões

Ao reter apenas $d = 2$ dimensões, parte da informação sobre o comportamento dos parlamentares nas votações está sendo ignorada, afim de possibilitar uma visualização simples no plano bidimensional.

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

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

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

Ao calcular para $d = 2$, o valor ν_2 dá uma idéia de quanta informação está sendo desconsiderada (quanto mais próximo de 1, menos informação foi desconsiderada).

Análise por Partido

Na primeira versão do *Radar Parlamentar* não era feita a análise dos deputados tal como apresentado até aqui, a qual foi implementada durante a *Maratona Hacker 2013 da Câmara dos Deputados*. Até então era feita apenas uma análise parecida, porém mais simples, considerando os votos agregados por partidos. Dessa forma era possível ver partidos no gráfico de bolhas, mas não era possível ver parlamentares individualmente.

Para entender a análise de votos agregados por partidos tal como era feita, deve-se notar que no modelo apresentado nada impede que os valores de $\mathbf{X}$ sejam números reais, situados por exemplo no intervalo [-1;1], em vez de apenas os valores discretos {-1;0;1}. Utiliza-se para tais valores justamente o "voto médio" do partido em cada votação. O voto médio do partido k na votação i é definido por:

$$x_{ik} = \frac{1}{|k|} \sum_{j \in k} x_{ij} \tag{3}$$

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.

Uma vez que a análise "desagregada", por parlamentares, tenha sido feita, uma alternativa para apresentar os partidos no gráfico é calcular a posição de um partido como sendo a posição média dos parlamentares que o compõem. Esta solução foi adotada no *Radar Parlamentar* para permitir que partidos e parlamentares possam ser apresentados em um mesmo gráfico interativo, permitindo "explodir" e "implodir" a visualização de um partido para ver a posição de seus parlamentares.

Transição entre Períodos

Para dois períodos consecutivos o *Radar Parlamentar* faz duas análises separadas, e uma animação mostra a transição entre um período e outro.

Nada garante que os dois resultados colocarão um mesmo partido em posições próximas no plano para os dois períodos, mas espera-se que as posições relativas entre os partidos mude pouco. Para melhor visualizar as mudanças nestas posições relativas pode ser interessante rotacionar o resultado de uma das análises em torno da origem, de tal forma que seja minimizada alguma medida das "distâncias percorridas" pelos partidos entre os períodos. A figura 1 ilustra esta idéia, em que o resultado da PCA de 2011 foi rotacionado de $\theta = 160°$.

Está claro que temos em mãos um problema de otimização, em que buscamos o ângulo θ que minimiza uma função objetivo $E(\theta)$ que dê uma medida das distâncias percorridas pelos partidos no plano.

Função Objetivo

É razoável esperar que partidos com mais parlamentares apresentem uma maior consistência de opiniões entre períodos, já que para que um partido com apenas um parlamentar altere radicalmente suas opiniões basta que o seu representante único o faça: assim associamos um peso M_k a cada partido k igual ao número de parlamentares que tem na casa.

Também é razoável penalizar mais as distâncias muito grandes, de forma que a função objetivo que usaremos será proporcional à soma das distâncias quadráticas, D_k^2, percorridas por cada partido k, ponderados pelos M_k.Temos portanto:

$$E(\theta) = \sum_{k \in \{partidos\}} M_k \cdot D_k^2 \tag{4}$$

Ou seja:

$$E(\theta) = \sum_{k \in \{partidos\}} M_k \cdot \left[(x_{k1} - x_{k2}(\theta))^2 + (y_{k1} - y_{k2}(\theta))^2 \right] \tag{5}$$

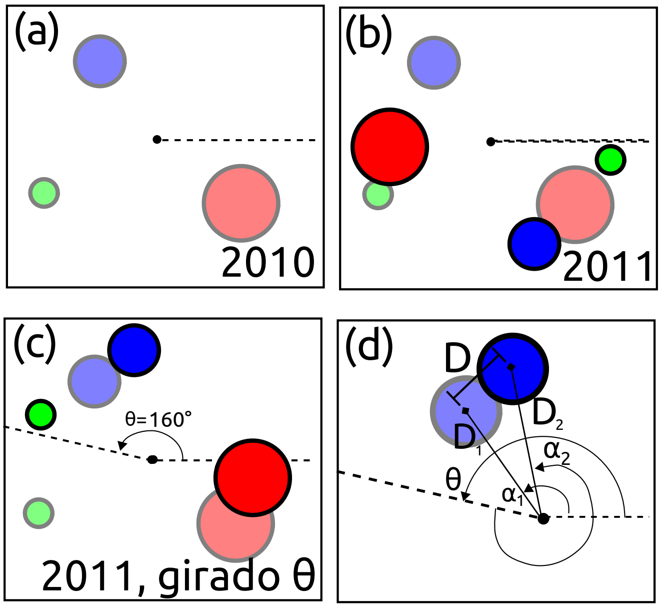

Figura 1: (a) Um resultado hipotético da análise de componentes principais para o ano de 2010 envolvendo três partidos. (b) Superposto ao primeiro, um resultado para o ano seguinte. (c) O mesmo resultado anterior, porém rotacionado de 160° para minimizar a movimentação dos partidos na animação entre um ano e outro. (d) Detalhe para um dos partidos.

Onde os índices 1 e 2 referem-se aos dois períodos consecutivos, e o período 2 é o que está sendo rotacionado de θ. x e y referem-se às coordenadas cartesianas.

Note-se a analogia desta função objetivo com a definição de energia: para um conjunto de bolas k massa M_k que iriam de (x_{k1}, y_{k1}) a (x_{k2}, y_{k2}) em um tempo Δt (o tempo da animação), estamos buscando o ângulo θ que minimiza a energia envolvida neste movimento.

Resolução Analítica

Representando as coordenadas na forma polar, conforme a notação da figura 1(d), definimos D_{k1}, D_{k2}, α_{k1} e α_{k2} como segue:

$$\begin{aligned} x_{k1} &= D_{k1} \cdot \cos(\alpha_{k1}) \\ y_{k1} &= D_{k1} \cdot \sin(\alpha_{k1}) \\ x_{k2} &= D_{k2} \cdot \cos(\theta + \alpha_{k2}) \\ y_{k2} &= D_{k2} \cdot \sin(\theta + \alpha_{k2}) \end{aligned} \tag{6}$$

Pela lei dos cossenos temos que a distância quadrática D_k^2 percorrida pelo partido k é dada por

$$D_k^2 = D_{k1}^2 + D_{k2}^2 - 2D_{k1}D_{k2}\cos(\alpha_{k1} - \alpha_{k2} - \theta)$$

Fazendo $\alpha_k = \alpha_{k1} - \alpha_{k2}$:

$$D_k^2 = D_{k1}^2 + D_{k2}^2 - 2D_{k1}D_{k2}\cos(\alpha_k - \theta)$$

E a função objetivo $E(\theta)$ pode ser reescrita na forma:

$$E(\theta) = \sum_{k \in \{partidos\}} M_k \left[D_{k1}^2 + D_{k2}^2 - 2D_{k1}D_{k2}\cos(\alpha_k - \theta)\right] \tag{7}$$

Usando a fórmula do arco duplo, temos $\cos(\alpha_k - \theta) = \cos\alpha_k \cos\theta + \sin\alpha_k \sin\theta$. Estamos procurando um mínimo, então vamos derivar $E(\theta)$, para depois igualar a derivada a zero:

$$\frac{\mathrm{d}E(\theta)}{\mathrm{d}\theta} = \sum_k \left[(2M_k D_{k1} D_{k2} \cos\alpha_k) \sin\theta + (2M_k D_{k1} D_{k2} \sin\alpha_k) \cos\theta\right]$$

O seno e cosseno de θ não dependem de k, logo podem ser colocados em evidência antes da somatória:

$$\frac{\mathrm{d}E(\theta)}{\mathrm{d}\theta} = \sin\theta \sum_k 2M_k D_{k1} D_{k2} \cos\alpha_k + \cos\theta \sum_k 2M_k D_{k1} D_{k2} \sin\alpha_k$$

Igualando a derivada a zero, dividindo tudo por $\cos\theta$ (desde que $\theta \neq \pm\frac{\pi}{2}$), e isolando a incógnita:

$$\tan\theta = \frac{\sum_k M_k D_{k1} D_{k2} \sin\alpha_k}{\sum_k M_k D_{k1} D_{k2} \cos\alpha_k}$$

$$\boxed{\theta = \arctan\frac{\sum_k M_k D_{k1} D_{k2} \sin\alpha_k}{\sum_k M_k D_{k1} D_{k2} \cos\alpha_k} + C\pi} \tag{8}$$

onde $C = 0\,\mathrm{ou}\,1$, um dos casos correspondendo ao mínimo e o outro ao máximo.

Para encontrar a solução basta calcular $E(\theta)$ para os θ que satisfazem (8) e verificar qual delas corresponde ao mínimo. Se durante o cálculo o denominador do argumento do arco-tangente for nulo, então o mínimo está em $\frac{\pi}{2}$ ou $\frac{3\pi}{2}$.

O mesmo procedimento deve ser repetido espelhando-se um dos eixos (por exemplo multiplicando todas as coordenadas x por -1), já que o sentido dos eixos que resulta da PCA é arbitrário. Assim o problema se resume à avaliação de $E(\theta)$ em 4 casos (espelhado e não-espelhado, cada um com dois valores de θ), e escolha do caso que resultar no mínimo.

Desvantagens de Rotacionar

Ao rotacionar os resultados o significado dos eixos horizontal e vertical torna-se menos claro. Uma alternativa para manter a interpretação do eixo horizontal como sendo a primeira compontente principal e o eixo vertical como sendo a segunda componente principal, sem contudo permitir que partidos e parlamentares “pulem” de um lado para o outro no gráfico ao avançar ou retroceder no tempo, é considerar apenas a possibilidade de espelhar os eixos. Para isso, deve-se avaliar $E(\theta = 0)$ em 4 casos (primeiro e segundo eixos espelhados ou não) e escolher o caso que resultar no mínimo.