Métodos Matriciais e Análise de Clusters

Redução de dimensionalidade

Laura de Oliveira F. Moraes

Recomendação de filmes

	Quem vai ficar com Mary?
Comédia	4
Drama	0

	Comédia	Drama
João	5	1

Você recomendaria o filme **Quem vai ficar com Mary?** para **João** sabendo os dados acima?

Recomendação de filmes

	Comédia	Drama
João	5	1

	Quem vai ficar com Mary?
Comédia	5 x 4 + 1 x 0 = 20

	Quem vai ficar com Mary?
Comédia	4
Drama	0

Você recomendaria o filme **Quem vai ficar com Mary?** para **João** sabendo os dados acima?

$$nota\ para\ filme = \sum_{todo\ genero} peso\ do\ genero\ no\ filme\ *\ gosto\ do\ usuario\ pelo\ genero$$

Recomendação de filmes

	Comédia	Drama
João	5	1

	Quem vai ficar com Mary?	**Pequena Miss Sunshine**
João	20	13

	Quem vai ficar com Mary?	**Pequena Miss Sunshine**
Comédia	4	2
Drama	0	3

Qual dos dois filmes você **recomendaria** para o João?

Recomendação de filmes

	Comédia	Drama
João	5	1

	Quem vai ficar com Mary?	**Pequena Miss Sunshine**
João	20	13

	Quem vai ficar com Mary?	**Pequena Miss Sunshine**
Comédia	4	2
Drama	0	3

Qual dos dois filmes você **recomendaria** para o João?

$$nota\ para\ filme = \sum_{todo\ genero} peso\ do\ genero\ no\ filme\ *\ gosto\ do\ usuario\ pelo\ genero$$

Recomendação de filmes

	Quem vai ficar com Mary?	Pequena Miss Sunshine
Comédia	4	2
Drama	0	3

Característica intrínseca do dado!
Nem sempre é fácil de quantificar!

	Comédia	Drama
João	5	1

Tem que pedir para o usuário. Será que ele vai responder?

Recomendação de filmes

	Quem vai ficar com Mary?	Pequena Miss Sunshine
João	20	13
Maria	13	18
Camila	5	12

Pode ser montado **automaticamente** baseado nas transações do usuário com o sistema. Quantas vezes viu, por quanto tempo, etc.

Como obter o **modelo de gostos por gênero do usuário**?

Fatoração de matrizes

$$NOTAS = filmes\ \ por\ \ genero \quad * \quad genero\ \ por\ \ usuario$$

Fatoração de matrizes

$$NOTAS \approx filmes\ \ por\ \ genero \quad * \quad genero\ \ por\ \ usuario$$

Fatoração de matrizes

$$NOTAS \approx filmes\ \ por\ \ genero \quad * \quad genero\ \ por\ \ usuario$$

$$V \approx WH$$

Fatoração de matrizes

$$NOTAS \approx filmes\ \ por\ \ genero \quad * \quad genero\ \ por\ \ usuario$$

$$V \approx WH$$

- **NMF** *(Non-negative Matrix Factorization)*: requer que os números em W e H sejam sempre **positivos**

Fatoração de matrizes

$$NOTAS \approx filmes\ \ por\ \ genero \quad * \quad genero\ \ por\ \ usuario$$

$$V \approx WH$$

- **NMF** *(Non-negative Matrix Factorization)***:** requer que os números em W e H sejam sempre **positivos**

- Outras fatorações:
  - **SVD** *(Singular Value Decomposition)* / **PCA** *(Principal Component Analysis)*: buscam a direção de **maior variância** dos dados = mais **próximos** dos dados originais
  - **MDS** (*Multidimensional Scaling)*: utiliza a **distância** entre os dados e não sua posição no espaço

Redução de dimensionalidade

- Encontrar as características **intrínsecas** ou **latentes** (escondidas) dos dados (como o gosto por gênero dos usuários e dos filmes ou assuntos em um conjunto de notícias)

- **Eliminar** um pouco do **ruído**, encontrando as dimensões que mais dão **informação**

- **Compressão** de dados

- **Visualização** de dados de alta dimensão

SVD/PCA

Objetivo: Ao diminuirmos dimensões, estamos comprimindo informação. Então é preciso encontrar no espaço reduzido, a posição dos pontos mais se **parece** com a original.

SVD/PCA

Objetivo: Ao diminuirmos dimensões, estamos comprimindo informação. Então é preciso encontrar no espaço reduzido, a posição dos pontos mais se **parece** com a original.

- Como definir o que mais se **parece**?
  - Possui o **menor erro** (distância euclidiana do ponto até o hiperplano)

SVD/PCA

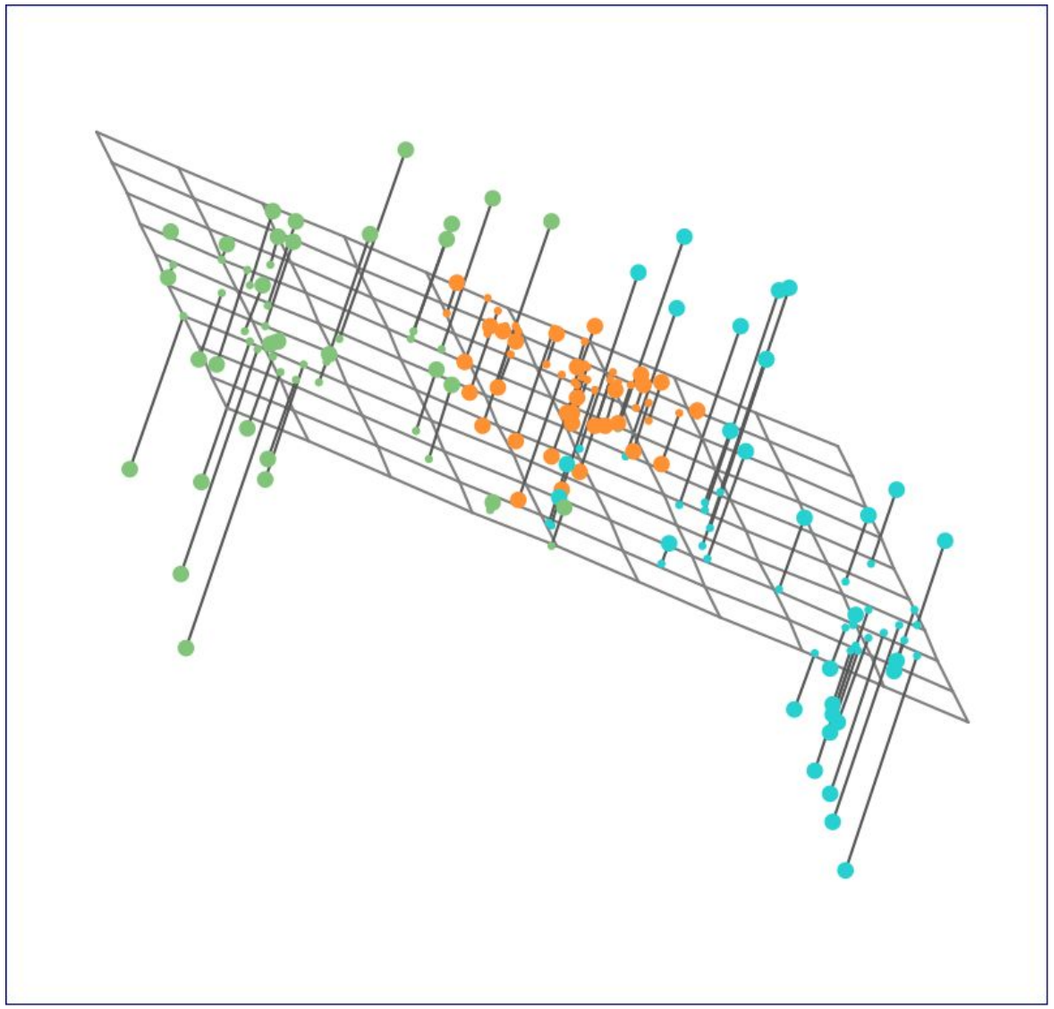
Objetivo: Ao diminuirmos dimensões, estamos comprimindo informação. Então é preciso encontrar no espaço reduzido, a posição dos pontos mais se **parece** com a original.

- Como definir o que mais se **parece**?
  - Possui o **menor erro** (distância euclidiana do ponto até o hiperplano)

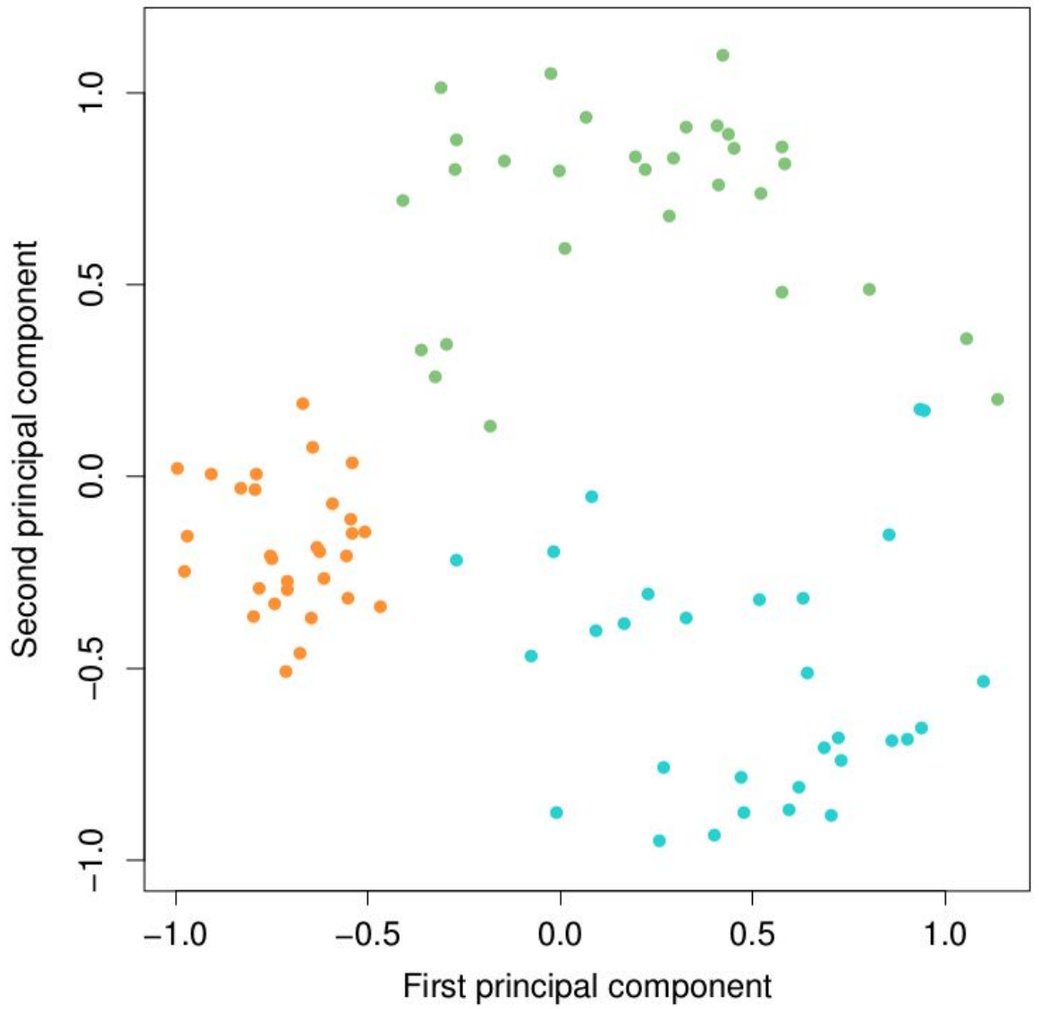
=

- Combinação das dimensões que produz a **maior variância**

SVD/PCA



Espaço original em R^3. Marcado o plano em que as distâncias euclidianas dos pontos ao plano somadas são as menores.



Espaço reduzido para R^2. Vista somente do plano.

SVD/PCA

- Algoritmo **iterativo**
- Descobre as dimensões (componentes principais) e as posições em cada dimensão **progressivamente:**
  - 1º descobre a combinação linear das dimensões originais de menor erro (maior variância) e calcula os pontos sobre ela

SVD/PCA

- Algoritmo **iterativo**
- Descobre as dimensões (componentes principais) e as posições em cada dimensão **progressivamente:**
  - 1º descobre a combinação linear das dimensões originais de menor erro (maior variância) e calcula os pontos sobre ela
  - Descobre a 2ª combinação linear das dimensões originais de menor erro (maior variância) e calcula os pontos

SVD/PCA

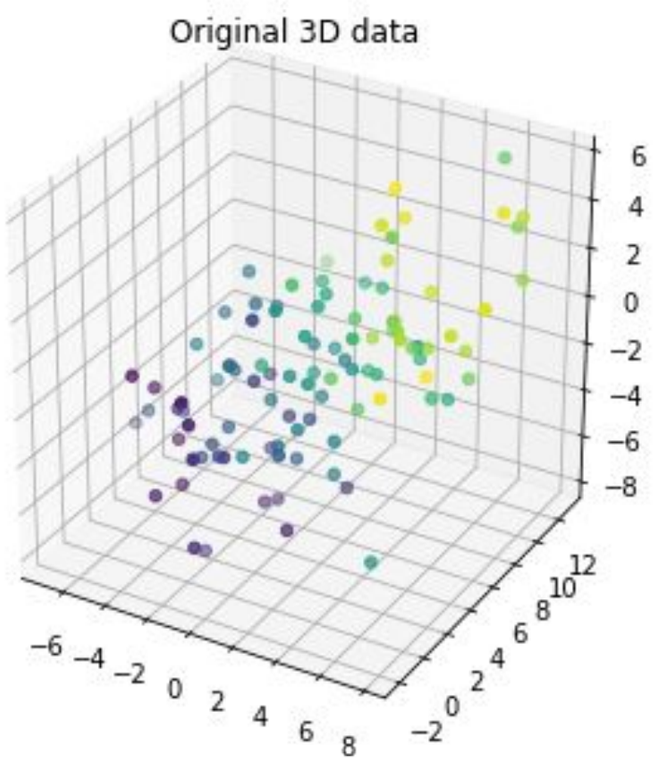
- Algoritmo **iterativo**
- Descobre as dimensões (componentes principais) e as posições em cada dimensão **progressivamente:**
  - 1º descobre a combinação linear das dimensões originais de menor erro (maior variância) e calcula os pontos sobre ela
  - Descobre a 2ª combinação linear das dimensões originais de menor erro (maior variância) e calcula os pontos
  - E assim sucessivamente até o número de dimensões **que eu queira**

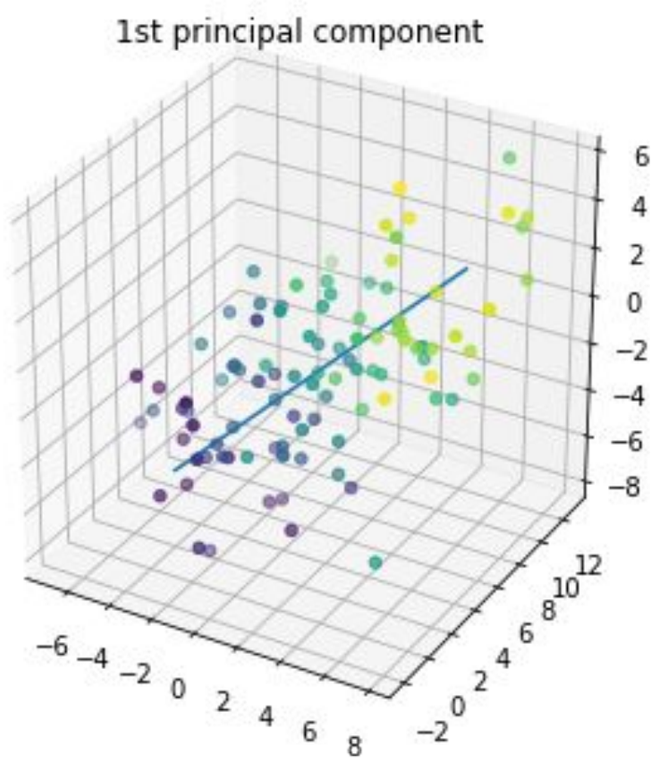
SVD/PCA

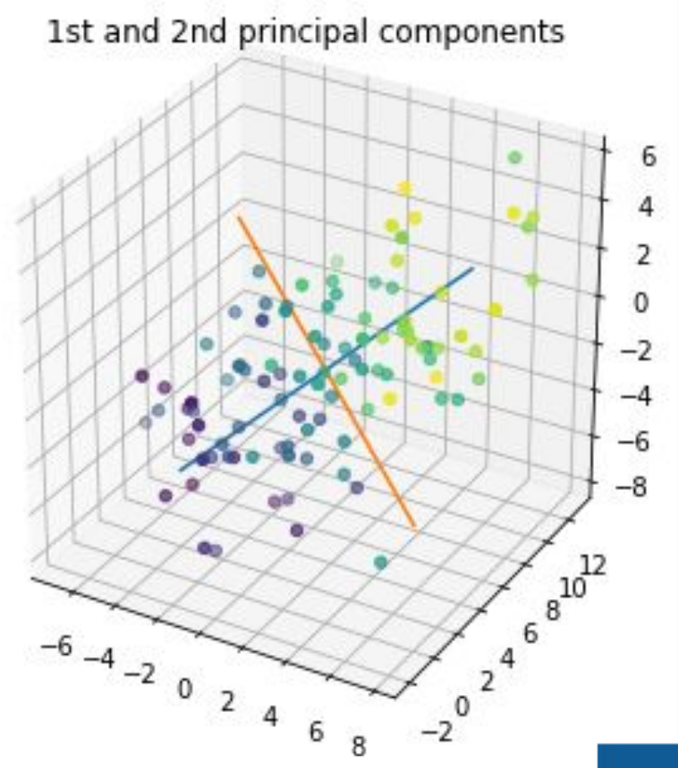
- Algoritmo **iterativo**
- Descobre as dimensões (componentes principais) e as posições em cada dimensão **progressivamente:**
  - 1º descobre a combinação linear das dimensões originais de menor erro (maior variância) e calcula os pontos sobre ela
  - Descobre a 2ª combinação linear das dimensões originais de menor erro (maior variância) e calcula os pontos
  - E assim sucessivamente até o número de dimensões **que eu queira**

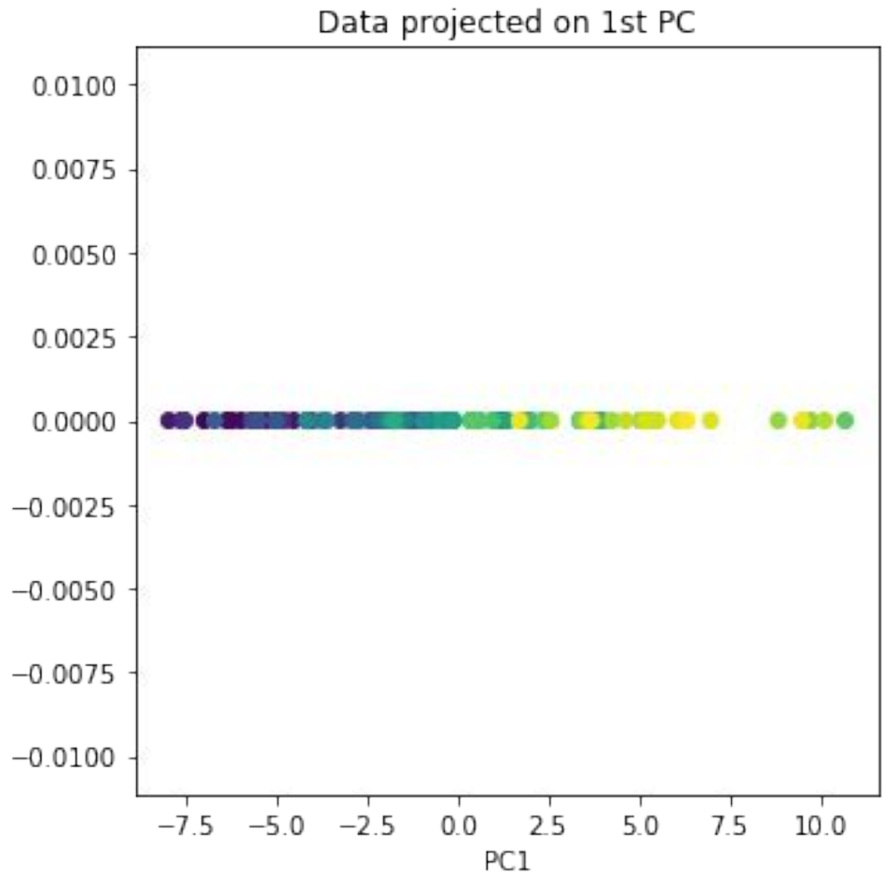
- SVD e PCA possuem diversos nomes na literatura: EOF (meteorologia), transformada de Karhunen-Loève (processamento de sinais), transformada de Hotelling (processamento de imagens)

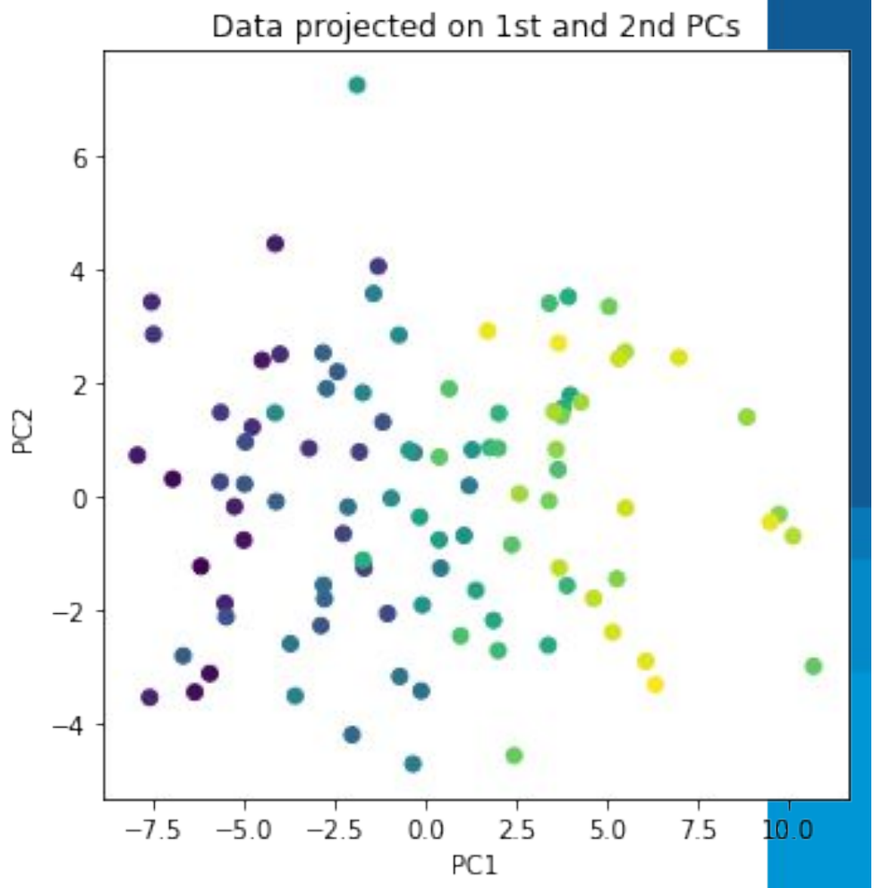
SVD/PCA


Original 3D data


1st principal component

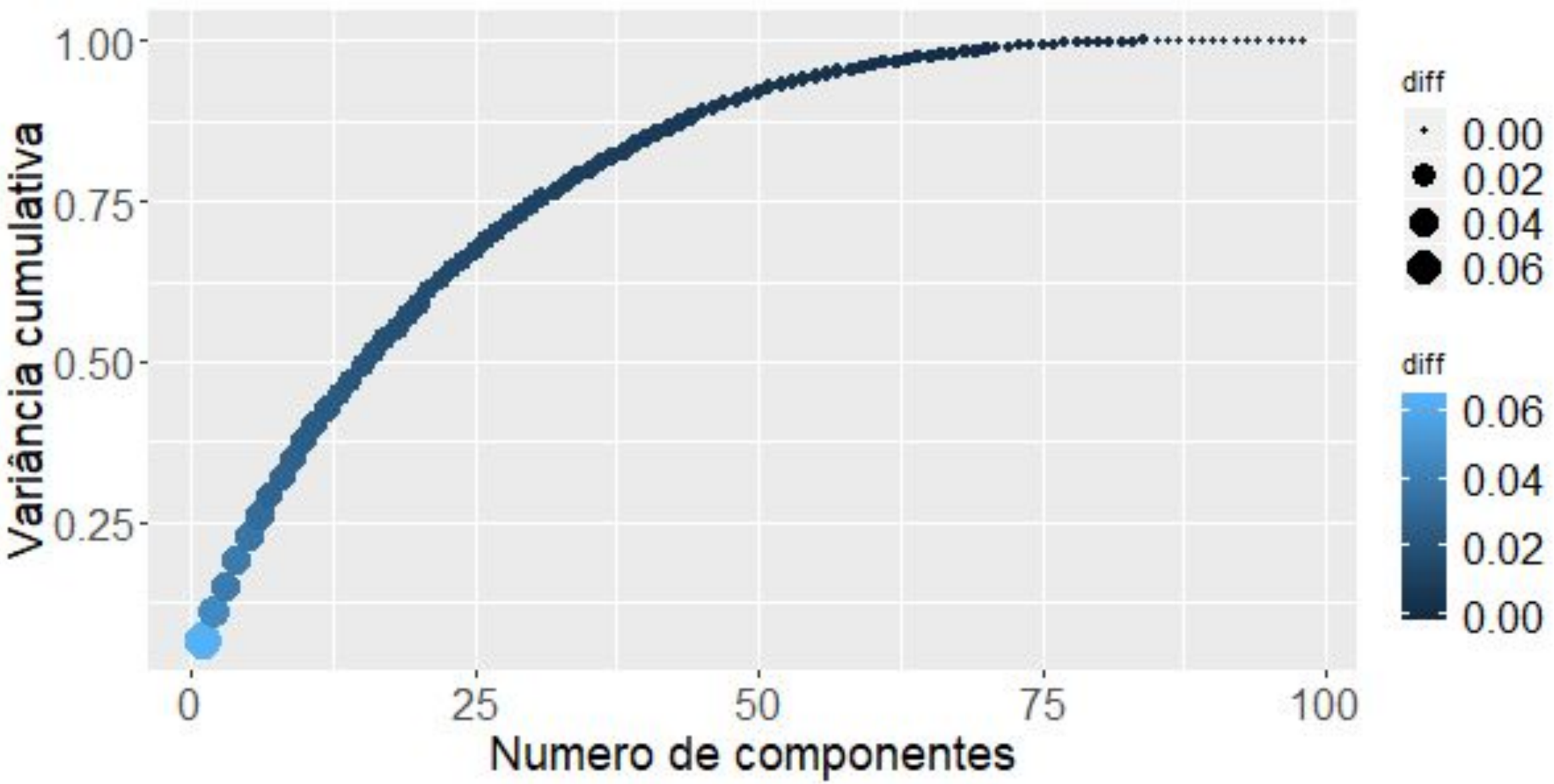

1st and 2nd principal components


Data projected on 1st PC
PC1

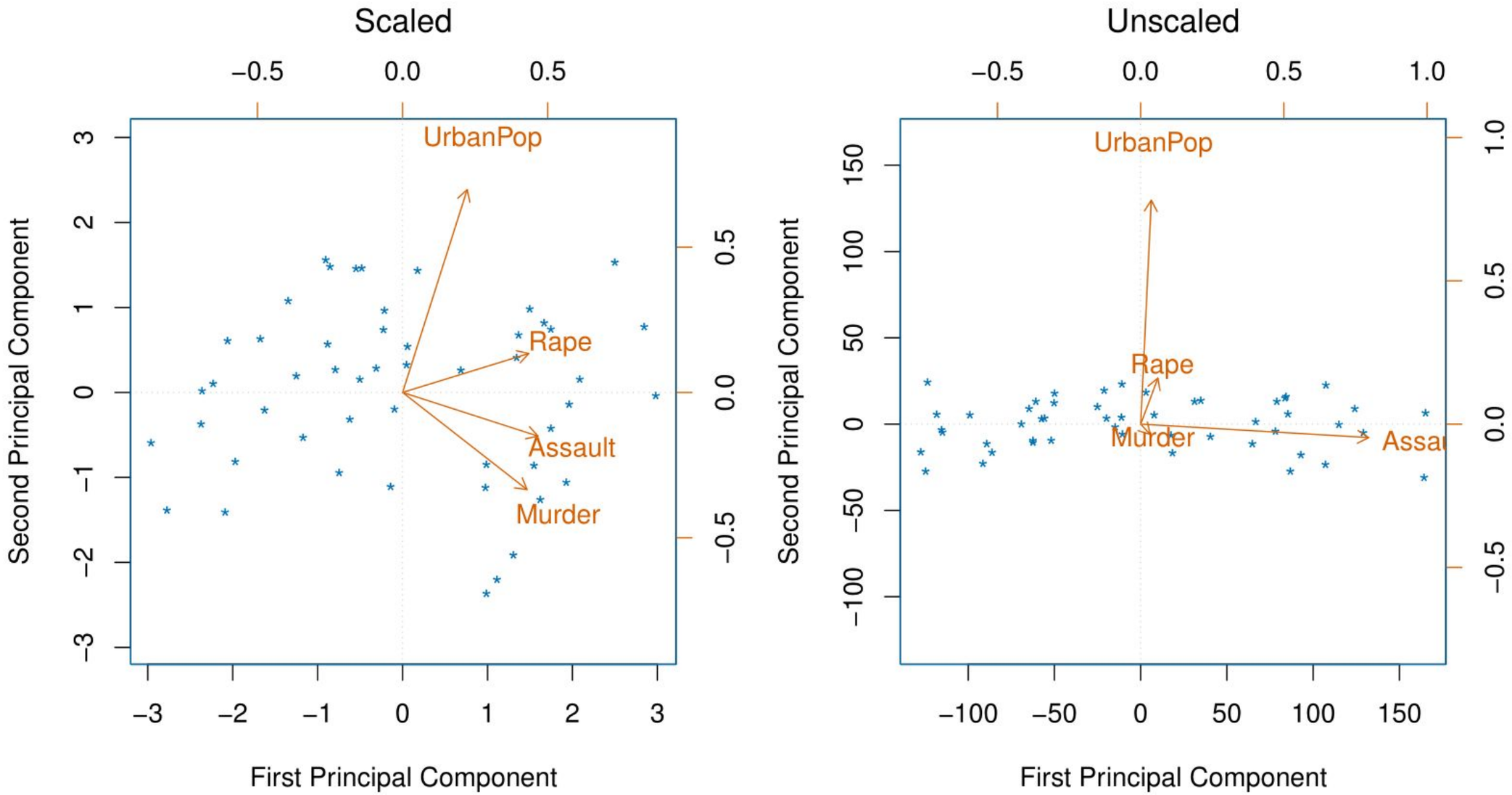

Data projected on 1st and 2nd PCs
PC1
PC2

Como se define um bom número de dimensões?

- Visualização requer que sejam 2 ou 3 dimensões
- Outras aplicações:

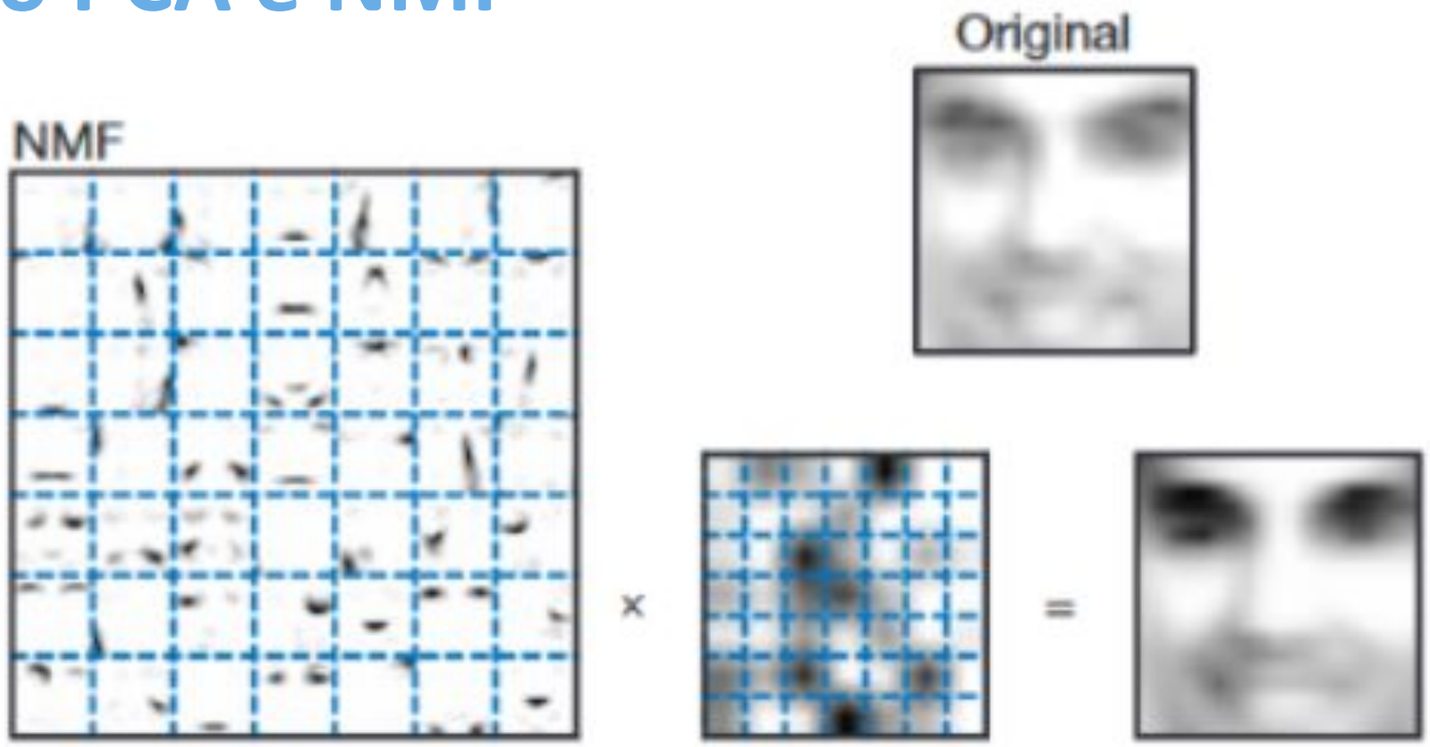


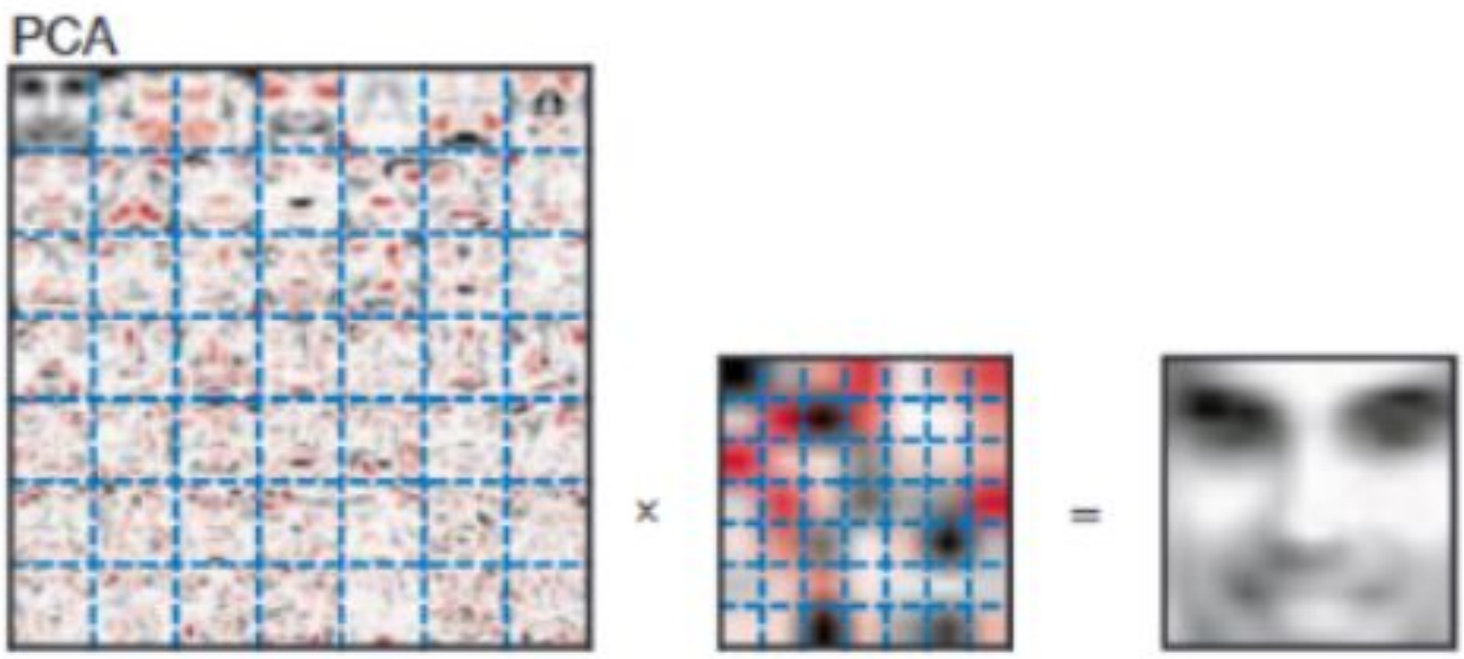
Não se esqueça de normalizar os dados!



Nos dados originais, Assault tem a maior variância. No gráfico da esquerda, vemos o resultado do PCA feito sobre os dados normalizados para média zero e variância unitária. No da direita o PCA é feito sobre os dados originais. A componente de maior variância acaba dominando a dimensão principal.

Comparação PCA e NMF



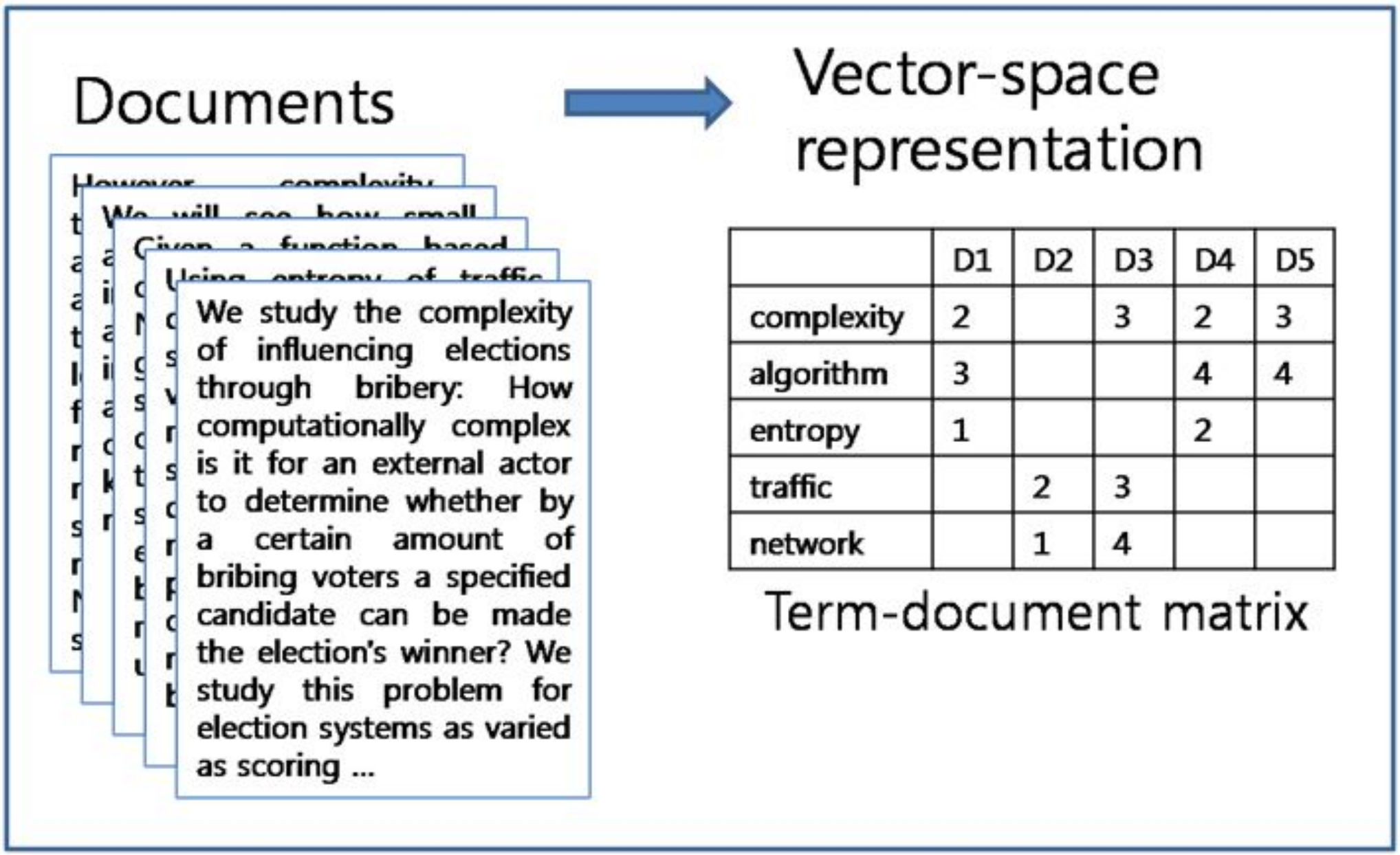


Netflix Prize

$$R = \begin{pmatrix} 1 & ? & 2 & ? & ? \\ ? & ? & ? & ? & 4 \\ 2 & ? & 4 & 5 & ? \\ ? & ? & 3 & ? & ? \\ ? & 1 & ? & 3 & ? \\ 5 & ? & ? & ? & 2 \end{pmatrix} \begin{matrix} \text{Alice} \\ \text{Bob} \\ \text{Charlie} \\ \text{Daniel} \\ \text{Eric} \\ \text{Frank} \end{matrix}$$

- SVD não fatora matriz com dados faltantes.
- Pode ser resolvido com:
  - Imputação dos dados e depois usar o SVD
  - SVD modificado: otimização com derivada e gradiente descendente

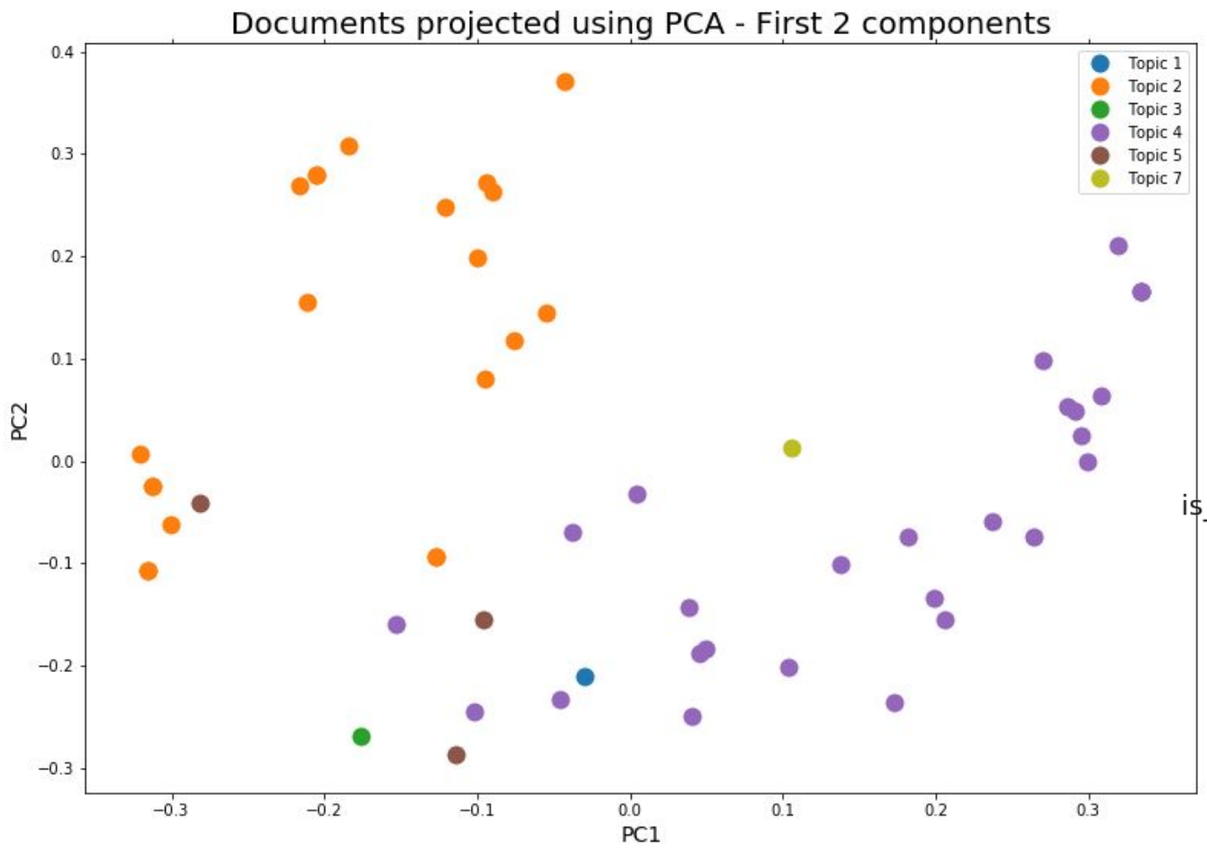
Tópicos

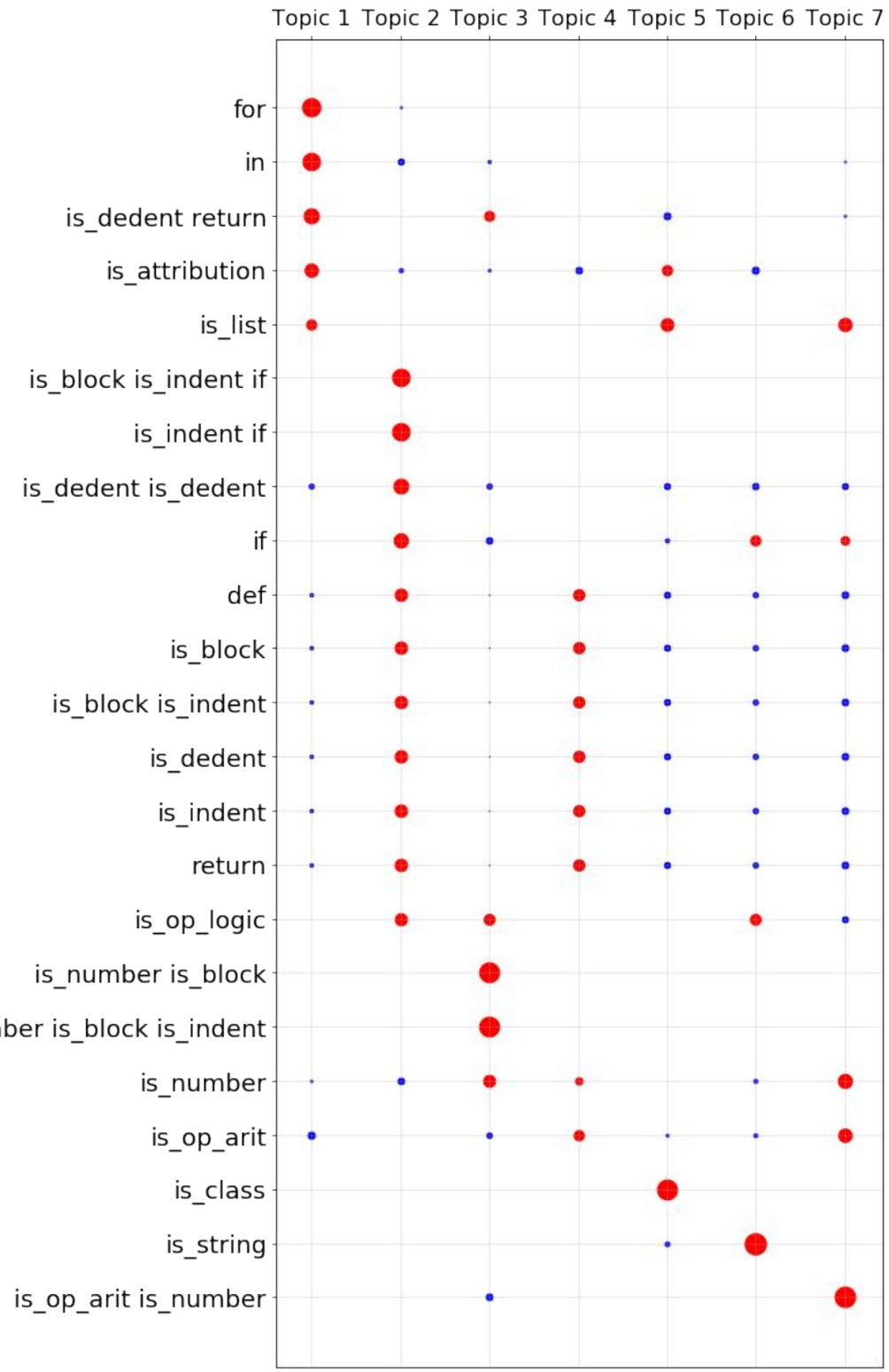


	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Tópicos

$$\frac{palavras}{documento} = \frac{assuntos}{documento} * \frac{palavras}{assunto}$$





Tópicos

FGV | EDUCAÇÃO EXECUTIVA

Sort by: Number of news ▾

- Most recent
- Oldest
- Number of news
- Relevance

Show: Todos ▾

- Gaining repercussion
- Decreasing repercussion
- Todos

rihanna, cantora, instagram, hotel, riri, fasano, sábado
Last news on: 05/10/2015 19:05:01

88 notícias

brt, terminal, alvorada, ônibus, embarque, consórcio, transporte
Last news on: 30/09/2015 08:53:02

70 notícias

Confira as linhas de ônibus regulares que passam mais próximas à Cidade do Rock

Após longas filas, sistema de ônibus BRT no Rock in Rio deve ter mudanças

Saiba como chegar à Cidade do Rock

Clique no tópico e descubra mais sobre o assunto

Adicionais

- Matriz de escalonamento e rotação
- Fatoramento PCA e SVD
- Teorema espectral
- Autovetores e autovalores