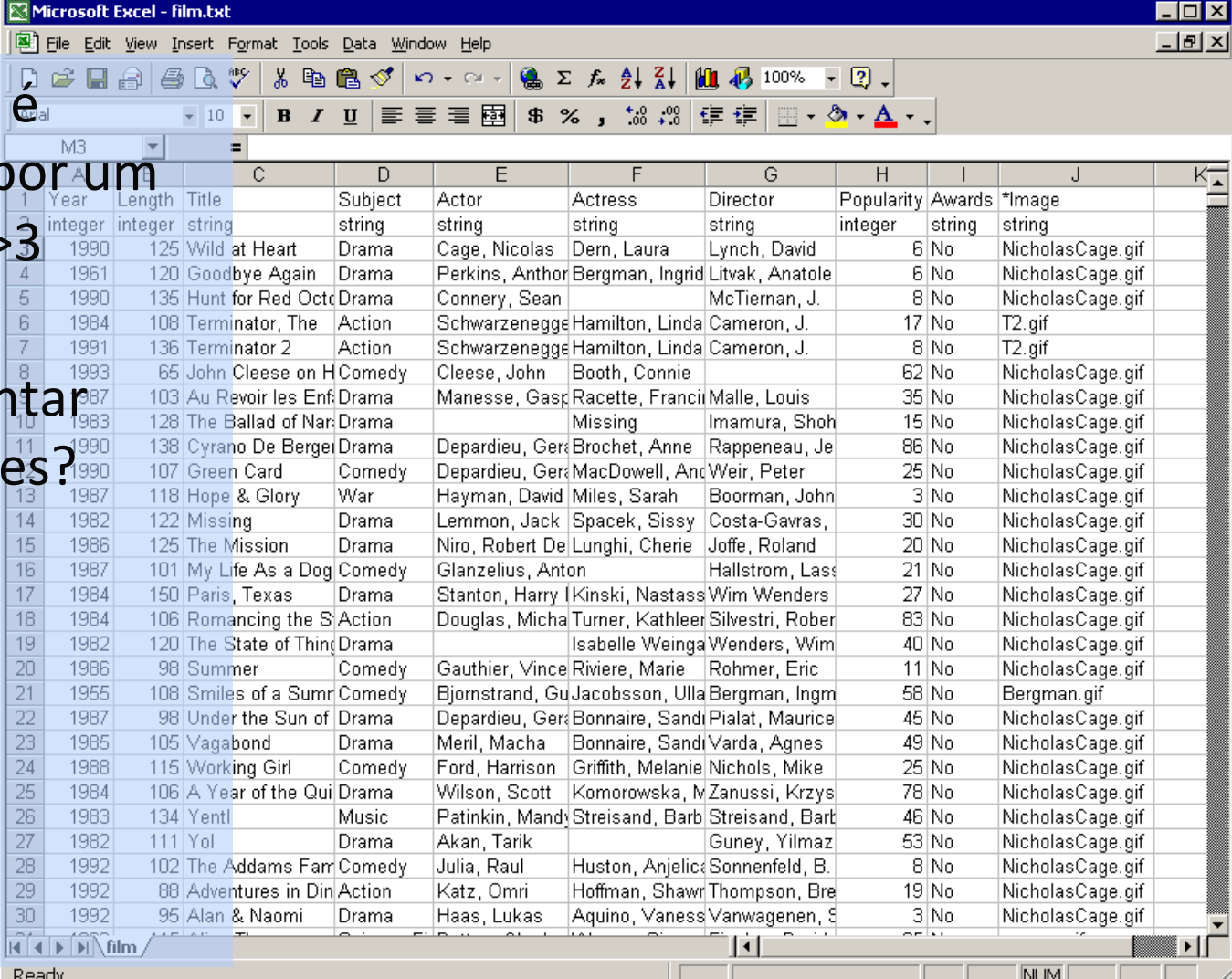


Visualização de Dados Multidimensionais

Parte 2

Relembrando: dados multidimensionais

- Cada entidade é caracterizada por um conjunto de $n > 3$ atributos
- Como representar essas dimensões?



The screenshot shows a Microsoft Excel window titled "Microsoft Excel - film.txt". The spreadsheet contains a table of movie data. The columns are labeled as follows: Year (integer), Length (integer), Title (string), Subject (string), Actor (string), Actress (string), Director (string), Popularity (integer), Awards (string), and *Image (string). The data rows are numbered 1 through 30. The status bar at the bottom indicates "Ready" and "NUM".

	Year	Length	Title	Subject	Actor	Actress	Director	Popularity	Awards	*Image
1	1990	125	Wild at Heart	Drama	Cage, Nicolas	Dern, Laura	Lynch, David	6	No	NicholasCage.gif
2	1961	120	Goodbye Again	Drama	Perkins, Anthor	Bergman, Ingrid	Litvak, Anatole	6	No	NicholasCage.gif
3	1990	135	Hunt for Red Oct	Drama	Connery, Sean		McTiernan, J.	8	No	NicholasCage.gif
4	1984	108	Terminator, The	Action	Schwarzenegge	Hamilton, Linda	Cameron, J.	17	No	T2.gif
5	1991	136	Terminator 2	Action	Schwarzenegge	Hamilton, Linda	Cameron, J.	8	No	T2.gif
6	1993	65	John Cleese on H	Comedy	Cleese, John	Booth, Connie		62	No	NicholasCage.gif
7	1987	103	Au Revoir les Enf	Drama	Manesse, Gas	Racette, Franci	Malle, Louis	35	No	NicholasCage.gif
8	1983	128	The Ballad of Nar	Drama		Missing	Imamura, Shoh	15	No	NicholasCage.gif
9	1990	138	Cyrano De Berge	Drama	Depardieu, Ger	Brochet, Anne	Rappeneau, Je	86	No	NicholasCage.gif
10	1990	107	Green Card	Comedy	Depardieu, Ger	MacDowell, And	Weir, Peter	25	No	NicholasCage.gif
11	1987	118	Hope & Glory	War	Hayman, David	Miles, Sarah	Boorman, John	3	No	NicholasCage.gif
12	1982	122	Missing	Drama	Lemmon, Jack	Spacek, Sissy	Costa-Gavras,	30	No	NicholasCage.gif
13	1986	125	The Mission	Drama	Niro, Robert De	Lunghi, Cherie	Joffe, Roland	20	No	NicholasCage.gif
14	1987	101	My Life As a Dog	Comedy	Glanzelius, Anton		Hallstrom, Lass	21	No	NicholasCage.gif
15	1984	150	Paris, Texas	Drama	Stanton, Harry	Kinski, Nastass	Wim Wenders	27	No	NicholasCage.gif
16	1984	106	Romancing the S	Action	Douglas, Micha	Turner, Kathleer	Silvestri, Rober	83	No	NicholasCage.gif
17	1982	120	The State of Thing	Drama		Isabelle Weinga	Wenders, Wim	40	No	NicholasCage.gif
18	1986	98	Summer	Comedy	Gauthier, Vince	Riviere, Marie	Rohmer, Eric	11	No	NicholasCage.gif
19	1955	108	Smiles of a Sumr	Comedy	Bjornstrand, Gu	Jacobsson, Ulla	Bergman, Ingm	58	No	Bergman.gif
20	1987	98	Under the Sun of	Drama	Depardieu, Ger	Bonnaire, Sandi	Pialat, Maurice	45	No	NicholasCage.gif
21	1985	105	Vagabond	Drama	Meril, Macha	Bonnaire, Sandi	Varda, Agnes	49	No	NicholasCage.gif
22	1988	115	Working Girl	Comedy	Ford, Harrison	Griffith, Melanie	Nichols, Mike	25	No	NicholasCage.gif
23	1984	106	A Year of the Qui	Drama	Wilson, Scott	Komorowska, M	Zanussi, Krzys	78	No	NicholasCage.gif
24	1983	134	Yentl	Music	Patinkin, Mand	Streisand, Barb	Streisand, Barb	46	No	NicholasCage.gif
25	1982	111	Yol	Drama	Akan, Tarik		Guney, Yilmaz	53	No	NicholasCage.gif
26	1992	102	The Addams Fam	Comedy	Julia, Raul	Huston, Anjelica	Sonnenfeld, B.	8	No	NicholasCage.gif
27	1992	88	Adventures in Din	Action	Katz, Omri	Hoffman, Shawr	Thompson, Bre	19	No	NicholasCage.gif
28	1992	95	Alan & Naomi	Drama	Haas, Lukas	Aquino, Vanessa	Vanwagenen, S	3	No	NicholasCage.gif

Representação de dados multidimensionais

- Mapear o espaço nD para o espaço 2D/3D da imagem
- Abordagens diferentes de mapeamento (Keim, 1996)
 - Técnicas iconográficas
 - Baseadas em ícones e glifos
 - Técnicas orientadas a pixel
 - Mapeamento direto para pixels na imagem
 - Técnicas de projeção geométrica
 - Projeção para coordenadas num domínio espacial

Técnicas de projeção geométricas

- Os dados são mapeados para representações visuais, através de algum tipo de **projeção geométrica**
 - Gráficos 2D tradicionais
 - Matriz de scatter plots
 - Coordenadas paralelas
 - Coordenadas radiais
 - **Projeções multidimensionais**

Projeções multidimensionais

- Diversas abordagens para redução de dimensionalidade (DR= *dimensionality reduction*)
- Família de técnicas capazes de produzir gráficos baseados em pontos (*scatterplots*) que preservam (dis)similaridade entre os elementos
 - O conceito de **(dis)similaridade** depende do tipo de dados e relações entre as instâncias, desde relações de ordem até distâncias no espaço multidimensional

Projeções multidimensionais

- Diferentes classificações, dependendo dos autores ...

Por exemplo:

- Técnicas lineares

- Novas dimensões são calculadas como combinações lineares das dimensões originais respeitando a propriedade

Seja $\Phi: D \rightarrow M$, então $\Phi(au+bv) = a\Phi(u) + b\Phi(v)$

- Técnicas não-lineares

- Novas dimensões são calculadas usando diferentes métodos, por exemplo, preservando a similaridade (num sentido mais amplo) existente do espaço original no espaço de dimensão reduzida

Exemplo de classificação

Class			Name	Complexity	Local?
Multidimensional Scaling	Metric		Classical Scaling (Young and Householder, 1938) ISOMAP (Tenenbaum, 1998) Sammon’s mapping (Sammon, 1969) CCA (Demartines and Herault, 1997)	$O(n^3)$ $O(n^2)$ $O(n^2)$ $O(n^2)$	✓
	Non-Metric		Kruskal (Kruskal, 1964)	$O(n^2)$	
Force-Directed Placement			Mass-spring model (Eades, 1984)	$O(n^3)$	✓
			Charlmers (Chalmers, 1996)	$O(n^2)$	✓
			Hybrid Model (Morrison et al., 2002)	$O(n\sqrt{n})$	✓
			Force Scheme (Tejada et al., 2003)	$O(cn^2)$	✓
Dimensionality Reduction	Linear	2do order	PCA (Jolliffe, 2002)	$O((n \times m)^3)$	
			Kernel-PCA (Schlkopf et al., 1999)	$O(n^3)$	
	SVD (Demmel and Y, 1997)		$O(n^3)$		
	Anchored Least Stress (ALS) (Wise, 1999)				
		Higher order	Projection Pursuit (Posse, 1995)	$O()$	
	Non-Linear	LLE (Roweis and Saul, 2000)	$O(n^2)$		
		Fastmap (Faloutsos and Lin, 1995)	$O(n)$		
LSP (Paulovich et al., 2008)		$O(nk^2 + n^2)$			
PLMP (Paulovich et al., 2010)		$O(n)$			
PLP (Paulovich et al., 2011)		$O(n)$	✓		
	LAMP (Joia et al., 2011a)	$O(n)$	✓		

Table 3.1 Classification of the multidimensional projection techniques. Where n and k represent the total number of instances and the number samples, c is the number of iterations, and m is the dimensionality of the high-dimensional space.

Extraída de Poco-Medina (2013) baseada em Paulovich(2008)

Exemplo de class

Métodos que consideram medidas de distâncias ou dissimilaridades entre os elementos do conjunto e criam uma representação desse conjunto no espaço visual

Class			Name		Local?
Multidimensional Scaling	Metric		Classical Scaling (Young and Householder, 1938) ISOMAP (Tenenbaum, 1998) Sammon's mapping (Sammon, 1969) CCA (Demartines and Herault, 1997)	$O(n^3)$ $O(n^2)$ $O(n^2)$ $O(n^2)$	✓
	Non-Metric		Kruskal (Kruskal, 1964)	$O(n^2)$	
Force-Directed Placement			Mass-spring model (Eades, 1984)	$O(n^3)$	✓
			Charlmers (Chalmers, 1996)	$O(n^2)$	✓
			Hybrid Model (Morrison et al., 2002)	$O(n\sqrt{n})$	✓
			Force Scheme (Tejada et al., 2003)	$O(cn^2)$	✓
Dimensionality Reduction	Linear	2do order	PCA (Jolliffe, 2002)	$O((n \times m)^3)$	
			Kernel-PCA (Schlkopf et al., 1999)	$O(n^3)$	
			SVD (Demmel and Y, 1997)	$O(n^3)$	
			Anchored Least Stress (ALS) (Wise, 1999)		
		Higher order	Projection Pursuit (Posse, 1995)	$O()$	
	Non-Linear		LLE (Roweis and Saul, 2000)	$O(n^2)$	
			Fastmap (Faloutsos and Lin, 1995)	$O(n)$	
			LSP (Paulovich et al., 2008)	$O(nk^2 + n^2)$	
			PLMP (Paulovich et al., 2010)	$O(n)$	
			PLP (Paulovich et al., 2011)	$O(n)$	✓
			LAMP (Joia et al., 2011a)	$O(n)$	✓

Table 3.1 Classification of the multidimensional projection techniques. Where n and k represent the total number of instances and the number samples, c is the number of iterations, and m is the dimensionality of the high-dimensional space.

Extraída de Poco-Medina (2013) baseada em Paulovich(2008)

Exemplo de classificação

Class			Name	Complexity	Local?
Multidimensional Scaling	Metric		Classical MDS (Crisp, 1968) ISOMAP (Tenenbaum and Semmes, 1998) Sammon's Mapping (Sammon, 1969) CCA (Crisp, 1968)		✓
	Non-Metric		Kruskal (Kruskal, 1964)	$O(n^2)$	
Force-Directed Placement			Mass-spring model (Eades, 1984)	$O(n^3)$	✓
			Charlmers (Chalmers, 1996)	$O(n^2)$	✓
			Hybrid Model (Morrison <i>et al.</i> , 2002)	$O(n\sqrt{n})$	✓
			Force Scheme (Tejada <i>et al.</i> , 2003)	$O(cn^2)$	✓
Dimensionality Reduction	Linear	2do order	PCA (Jolliffe, 2002)	$O((n \times m)^3)$	
			Kernel-PCA (Schlkopf <i>et al.</i> , 1999)	$O(n^3)$	
			SVD (Demmel and Y, 1997)	$O(n^3)$	
			Anchored Least Stress (ALS) (Wise, 1999)		
		Higher order	Projection Pursuit (Posse, 1995)	$O()$	
	Non-Linear		LLE (Roweis and Saul, 2000)	$O(n^2)$	
			Fastmap (Faloutsos and Lin, 1995)	$O(n)$	
			LSP (Paulovich <i>et al.</i> , 2008)	$O(nk^2 + n^2)$	
			PLMP (Paulovich <i>et al.</i> , 2010)	$O(n)$	
			PLP (Paulovich <i>et al.</i> , 2011)	$O(n)$	✓
			LAMP (Joia <i>et al.</i> , 2011a)	$O(n)$	✓

Métodos baseados em otimização de funções de stress ou sistemas massa-mola, métodos baseados em forças

Table 3.1 Classification of the multidimensional projection techniques. Where n and k represent the total number of instances and the number samples, c is the number of iterations, and m is the dimensionality of the high-dimensional space.

Extraída de Poco-Medina (2013) baseada em Paulovich(2008)

Exemplo de classificação

Class			Name	Complexity	Local?
Multidimensional Scaling	Metric		Classical Scaling (Young and Householder, 1938) ISOMAP (Tenenbaum, 1998) Sammon's mapping (Sammon, 1969) CCA (Demartines and Herault, 1997)	$O(n^3)$ $O(n^2)$ $O(n^2)$ $O(n^2)$	✓
	Non-Metric		Kruskal		
Force-Directed Placement			Mass Charl Hybr Force		✓ ✓ ✓ ✓
Dimensionality Reduction	Linear	2do order	PCA Kerne SVD (Demmel and Y, 1997) Anchored Least Stress (ALS) (Wise, 1999)	$O(n^3)$	
		Higher order	Projection Pursuit (Posse, 1995)	$O()$	
			LLE (Roweis and Saul, 2000) Fastmap (Faloutsos and Lin, 1995) LSP (Paulovich et al., 2008) PLMP (Paulovich et al., 2010) PLP (Paulovich et al., 2011) LAMP (Joia et al., 2011a)	$O(n^2)$ $O(n)$ $O(nk^2 + n^2)$ $O(n)$ $O(n)$ $O(n)$	✓ ✓
	Non-Linear				

Métodos baseados numa transformação dos pontos do espaço original em outra representação num espaço de dimensão reduzida, transformação que leva em conta característica(s) dos dados

Table 3.1 Classification of the multidimensional projection techniques. Where n and k represent the total number of instances and the number samples, c is the number of iterations, and m is the dimensionality of the high-dimensional space.

Extraída de Poco-Medina (2013) baseada em Paulovich(2008)

Exemplo de classificação

Class		Name	Complexity	Local?	
Multidimensional Scaling	Metric	Classical Scaling (Young and Householder, 1938) ISOMAP (Tenenbaum, 1998) Sammon's mapping (Sammon, 1969) CCA (Demartines and Herault, 1997)	$O(n^3)$ $O(n^2)$ $O(n^2)$ $O(n^2)$	✓	
	Non-Metric	Kruskal			
Force-Directed Placement		Mass Char Hybr Force		✓ ✓ ✓ ✓	
Dimensionality Reduction	Linear	2do order	PCA (Jolliffe, 2002) Kernel-PCA (Schlkopf et al., 1999) SVD (Demmel and Y, 1997) Anchored Least Stress (ALS) (Wise, 1999)	$O((n \times m)^3)$ $O(n^3)$ $O(n^3)$	
		Higher order	Projection Pursuit (Posse, 1995)	$O()$	
	Non-Linear		LLE (Roweis and Saul, 2000) Fastmap (Faloutsos and Lin, 1995) LSP (Paulovich et al., 2008) PLMP (Paulovich et al., 2010) PLP (Paulovich et al., 2011) LAMP (Joia et al., 2011a)	$O(n^2)$ $O(n)$ $O(nk^2 + n^2)$ $O(n)$ $O(n)$ $O(n)$	✓ ✓

Apenas relações lineares são preservadas, ou seja, considera que os dados tem uma estrutura linear no espaço original

Apenas relações lineares são preservadas, ou seja, considera que os dados tem uma estrutura linear no espaço original

Table 3.1 Classification of the multidimensional projection techniques. Where n and k represent the total number of instances and the number samples, c is the number of iterations, and m is the dimensionality of the high-dimensional space.

Extraída de Poco-Medina (2013) baseada em Paulovich(2008)

Exemplo de classificação

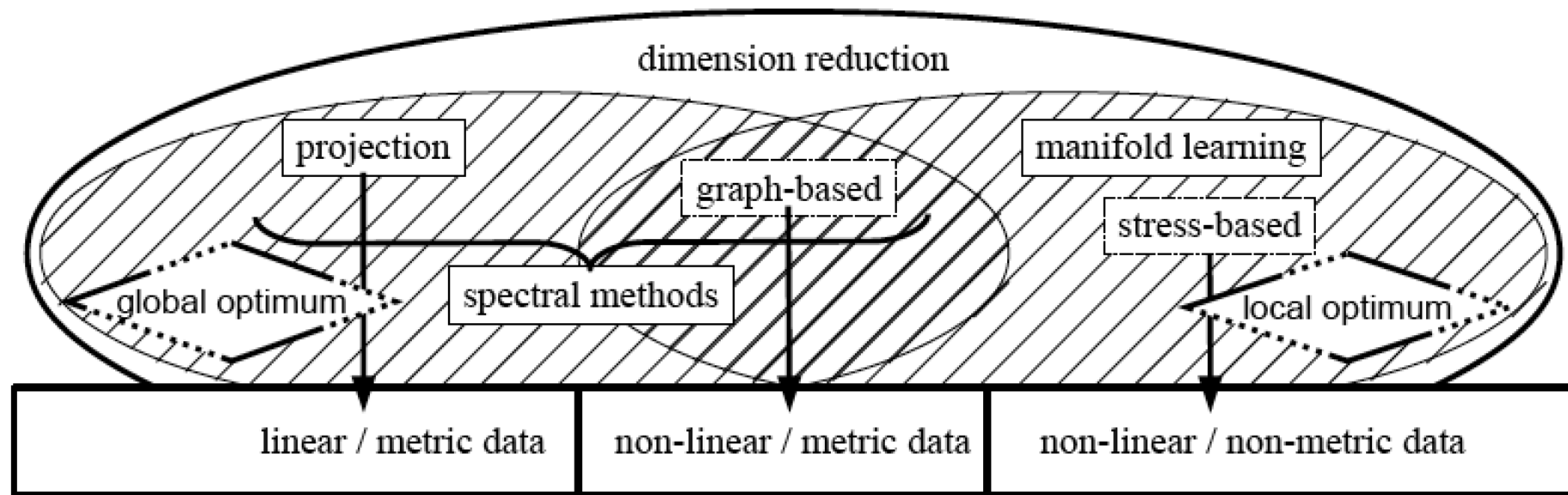
Class			Name	Complexity	Local?
Multidimensional Scaling	Metric		Classical Scaling (Young and Householder, 1938) ISOMAP (Tenenbaum, 1998) Sammon's mapping (Sammon, 1969) CCA (Demartines and Herault, 1997)	$O(n^3)$ $O(n^2)$ $O(n^2)$ $O(n^2)$	✓
	Non-Metric		Kruskal		
Force-Directed Placement			Mass Char Hybr Force		✓ ✓ ✓ ✓
Dimensionality Reduction	Linear	2do order	PCA (Jolliffe, 2002) Kernel-PCA (Schlkopf et al., 1999) SVD (Demmel and Y, 1997) Anchored Least Stress (ALS) (Wise, 1999)	$O((n \times m)^3)$ $O(n^3)$ $O(n^3)$	
			Projection Pursuit (Posse, 1995)	$O()$	
		Non-Linear	LLE (Roweis and Saul, 2000) Fastmap (Faloutsos and Lin, 1995) LSP (Paulovich et al., 2008) PLMP (Paulovich et al., 2010) PLP (Paulovich et al., 2011) LAMP (Joia et al., 2011a)	$O(n^2)$ $O(n)$ $O(nk^2 + n^2)$ $O(n)$ $O(n)$ $O(n)$	✓ ✓

Assume que os dados originais **não** estão organizados segundo uma estrutura linear (por exemplo, são “manifolds” desconhecidos)

Table 3.1 Classification of the multidimensional projection techniques. Where n and k represent the total number of instances and the number samples, c is the number of iterations, and m is the dimensionality of the high-dimensional space.

Extraída de Poco-Medina (2013) baseada em Paulovich(2008)

Exemplo de classificação #2

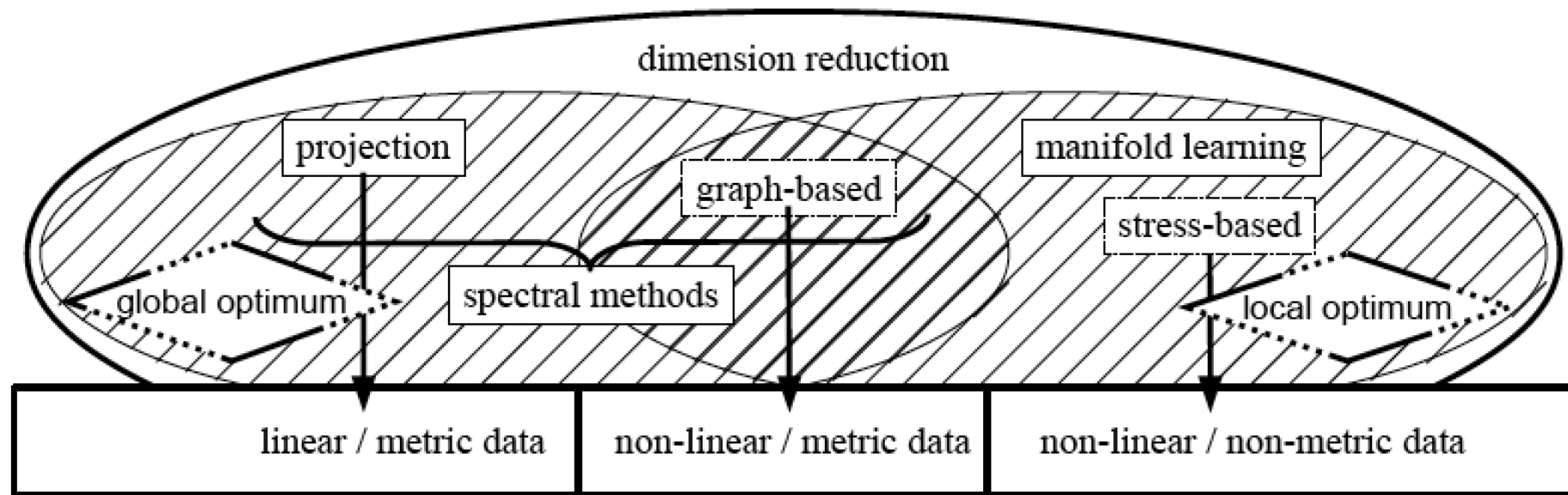


Methods that are solely based on linear transformations are defined as projection techniques.

Methods that are able to ascertain distance relationships in a non-linear data structure are defined as manifold learning techniques.

Extraída de (Engel, Hüttenberger, Hamann, 2011)

Exemplo de classificação #2

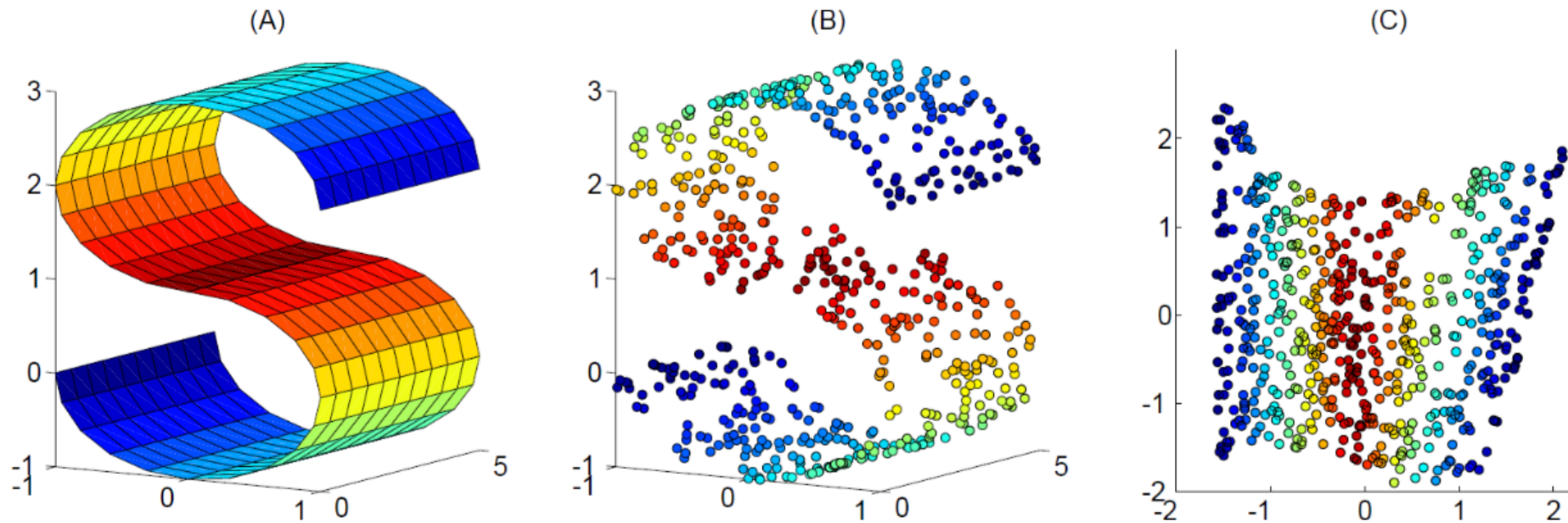


Spectral methods rely on a point-wise distance matrix.

Graph-based methods utilize optimizations of graph theory to learn manifold distances in data space.

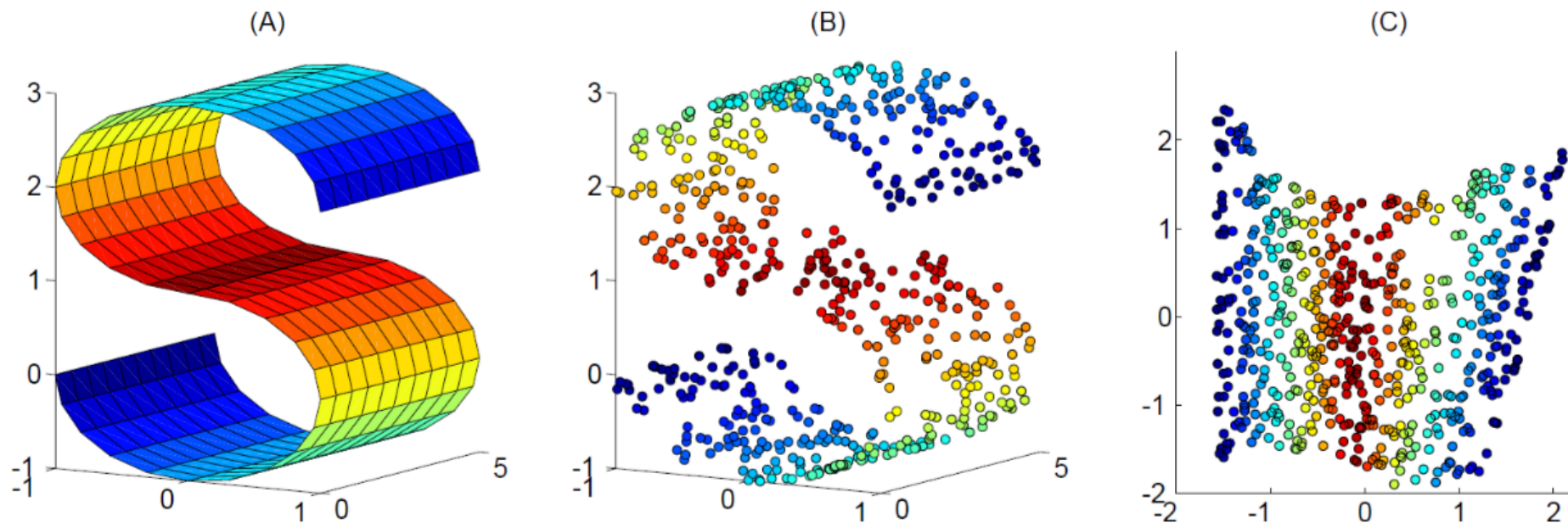
Stress-based focuses on the embedding directly, i.e., learning the mapping that minimizes the mapping error in target space by iterative optimizations of the mapping error (stress)

Exemplo de DR



■ **Figure 2** In this example, data sampled from a non-linear three-dimensional manifold (A) are mapped by a projection-based method (B) and by a manifold learning technique (C). In (B), the projection of the data is a linear transformation that optimally captures Euclidean distances. In (C), distance relationships along the manifold are captured by a non-linear mapping of the data. This figure derives from [21].

Exemplo de DR



Dois **princípios de Gestalt** são usados aqui: **proximidade e similaridade**.

A lei da **proximidade** diz que pontos que estão próximos são processados em tempo de pré-atenção e, assim, intuitivamente percebemos que deve compartilhar alguma característica. A lei da **similaridade** diz que pontos representados por marcadores iguais (cor, símbolo) também indicam compartilhamento de alguma característica, ou seja, os elementos pertencem a um mesmo grupo ou classe.

Exemplo de classificação

Class			Name	Complexity	Local?	
Multidimensional Scaling	Metric		Classical Scaling (Young and Householder, 1938)	$O(n^3)$		
			ISOMAP (Tenenbaum, 1998)	$O(n^2)$		
			Sammon's mapping (Sammon, 1969)	$O(n^2)$		
			CCA (Demartines and Herault, 1997)	$O(n^2)$	✓	
	Non-Metric		Kruskal (Kruskal, 1964)	$O(n^2)$		
Force-Directed Placement			Mass-spring model (Eades, 1984)	$O(n^3)$	✓	
			Charlmers (Chalmers, 1996)	$O(n^2)$	✓	
			Hybrid Model (Morrison <i>et al.</i> , 2002)	$O(n\sqrt{n})$	✓	
			Force Scheme (Tejada <i>et al.</i> , 2003)	$O(cn^2)$	✓	
Dimensionality Reduction	Linear	2do order	PCA (Jolliffe, 2002)	$O((n \times m)^3)$		
				Kernel-PCA (Schlkopf <i>et al.</i> , 1999)	$O(n^3)$	
				SVD (Demmel and Y, 1997)	$O(n^3)$	
				Anchored Least Stress (ALS) (Wise, 1999)		
			Higher order	Projection Pursuit (Posse, 1995)	$O()$	
	Non-Linear		LLE (Roweis and Saul, 2000)	$O(n^2)$		
			Fastmap (Faloutsos and Lin, 1995)	$O(n)$		
			LSP (Paulovich <i>et al.</i> , 2008)	$O(nk^2 + n^2)$		
			PLMP (Paulovich <i>et al.</i> , 2010)	$O(n)$		
			PLP (Paulovich <i>et al.</i> , 2011)	$O(n)$	✓	
		LAMP (Joia <i>et al.</i> , 2011a)	$O(n)$	✓		

Table 3.1 Classification of the multidimensional projection techniques. Where n and k represent the total number of instances and the number samples, c is the number of iterations, and m is the dimensionality of the high-dimensional space.

Extraída de Poco-Medina (2013) baseada em Paulovich(2008)

Técnicas lineares: exemplo PCA e MDS

- ***Principal Component Analysis (PCA)***
 - Pearson (1901): buscava linhas e planos que melhor se adequassem a um conjunto de pontos em um espaço n-dimensional. Conceito de Componente Principal (PC)
 - Hotelling (1933): buscava encontrar um conjunto menor de variáveis (dimensões) que expressassem as n variáveis originais. Procurou maximizar essas “componentes principais” no sentido da variância das variáveis originais.
- Baseado na identificação das combinações lineares de variáveis que melhor identifiquem a variabilidade dos dados

PCA

- Dado um conjunto de dados representado por uma matriz de m elementos (pontos), cada um com n atributos (dimensões)
 - PCA sumariza esse conjunto como pontos num espaço formado por eixos não-correlacionados (componentes principais) que são combinações lineares das n dimensões originais
- As k primeiras componentes “contém” a maior variação do conjunto de dados

Interpretação geométrica

- Os n eixos (no espaço n -dimensional) são rotacionados rigidamente (transformação linear) para novas orientações (eixos componentes principais) tal que:
 - O eixo principal 1 tem a maior variância, o eixo 2 tem a segunda maior variância, etc.
 - A covariância entre cada par de eixos é zero, ou seja, os componentes principais não são correlacionados

PCA – definições iniciais importantes

- Centróide dos pontos: valor médio de cada atributo

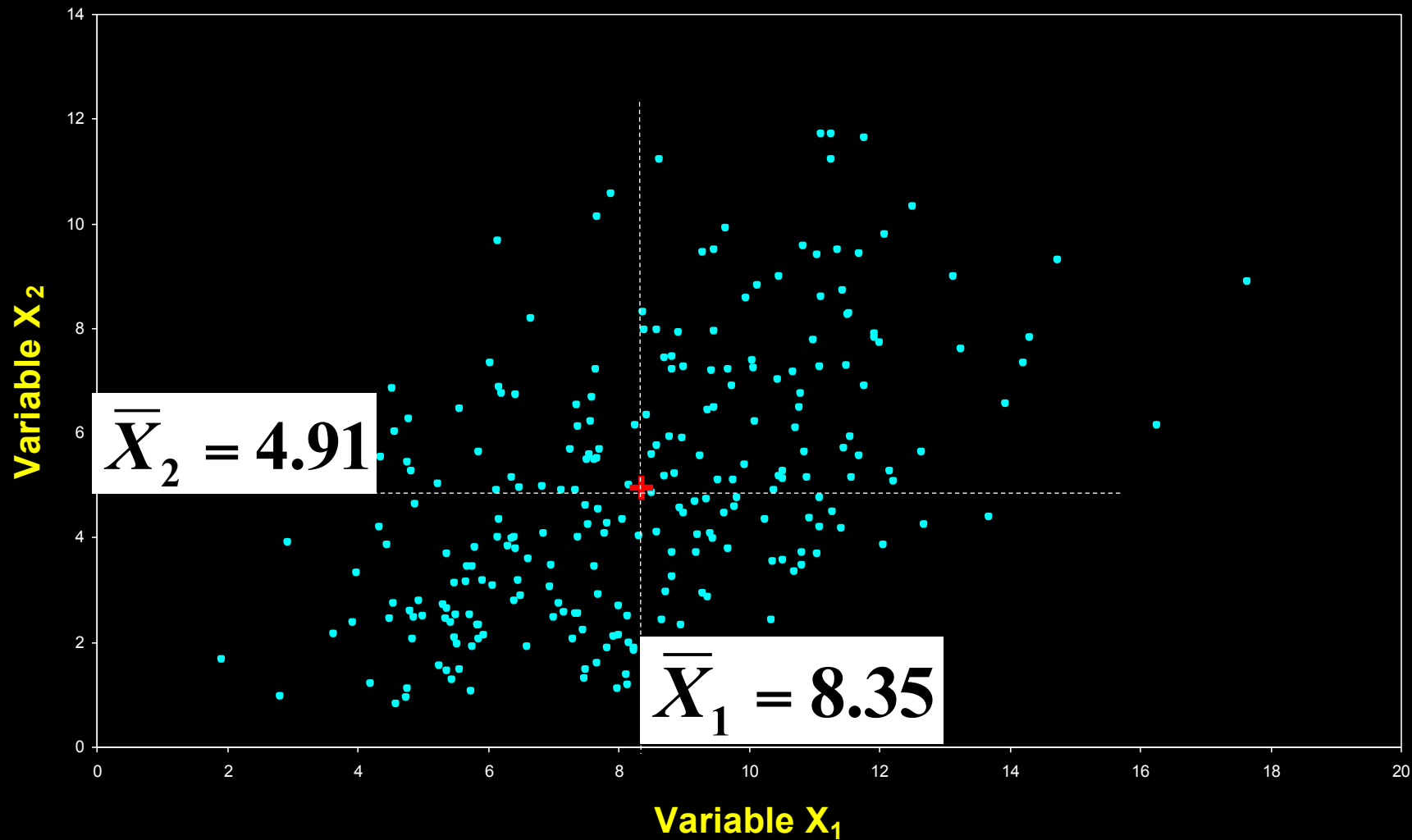
- Variância:
$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n - 1)}$$

- Covariância:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n - 1)}$$

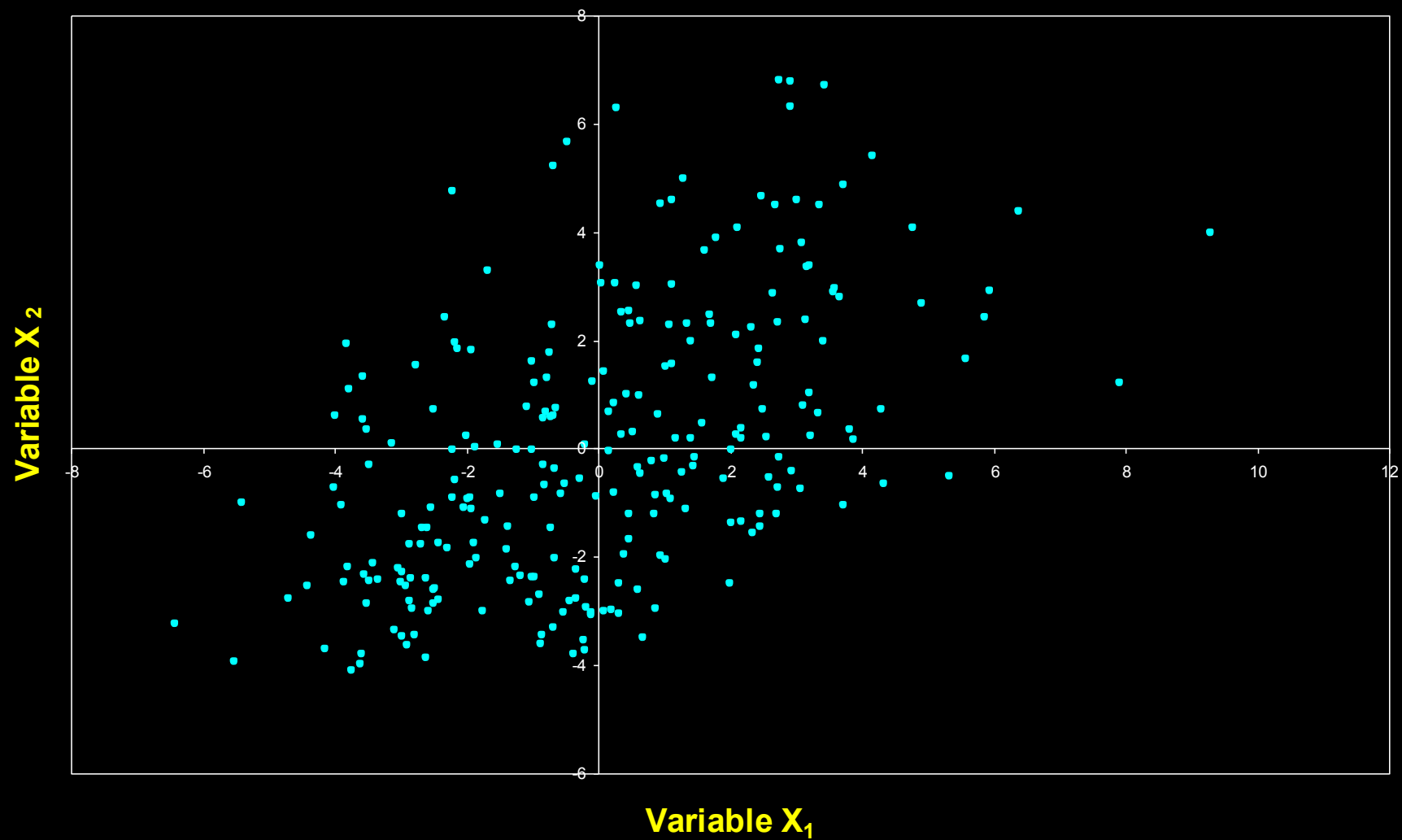
- <https://www.youtube.com/watch?v=F-nfsSq42ow>

Váriaveis X_1 e X_2 tem covariância positiva e variância similar

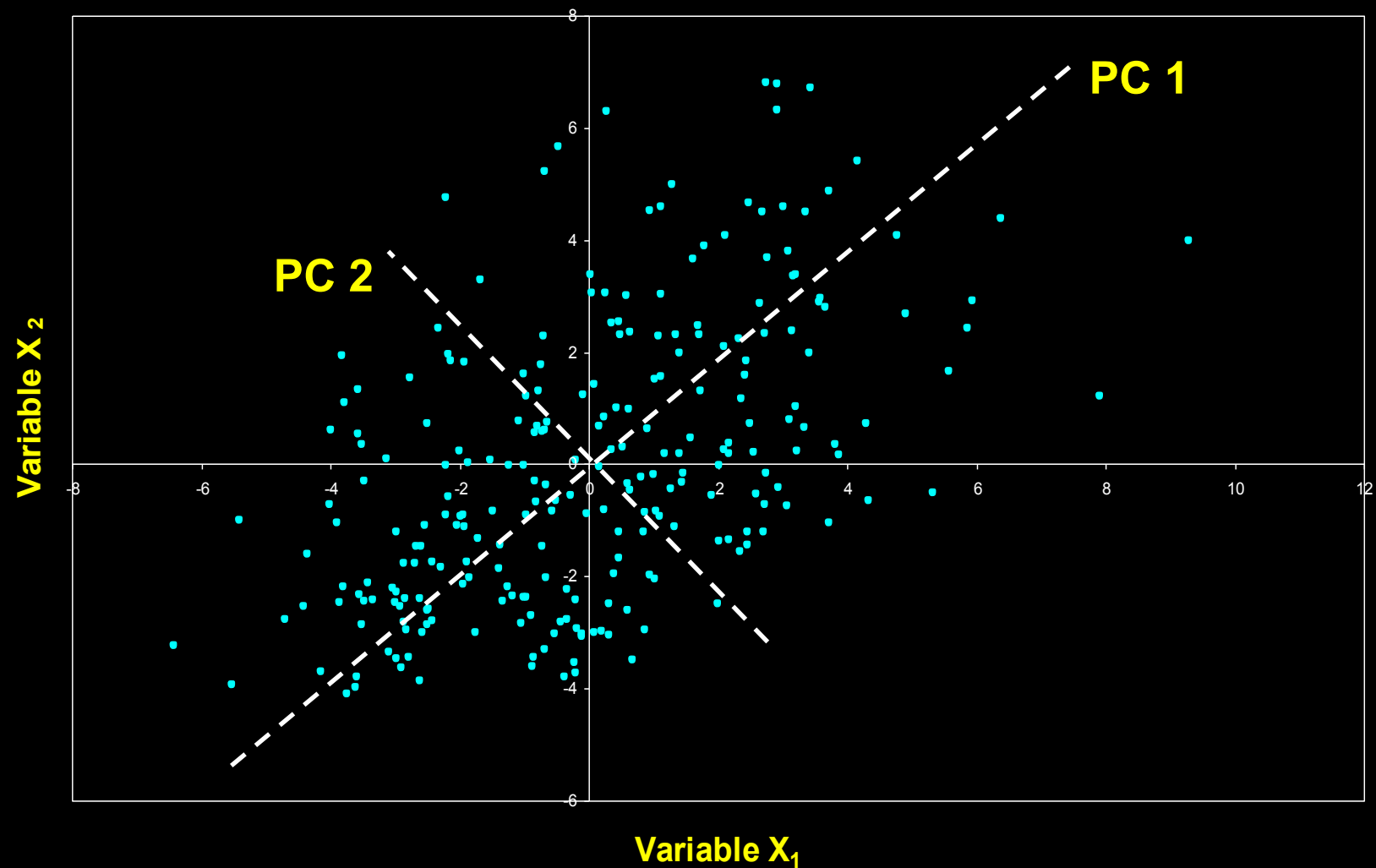


$$V_1 = 6.67; \quad V_2 = 6.24; \quad \text{Cov} = 3.42$$

Dados centralizados (subtraindo o centróide)

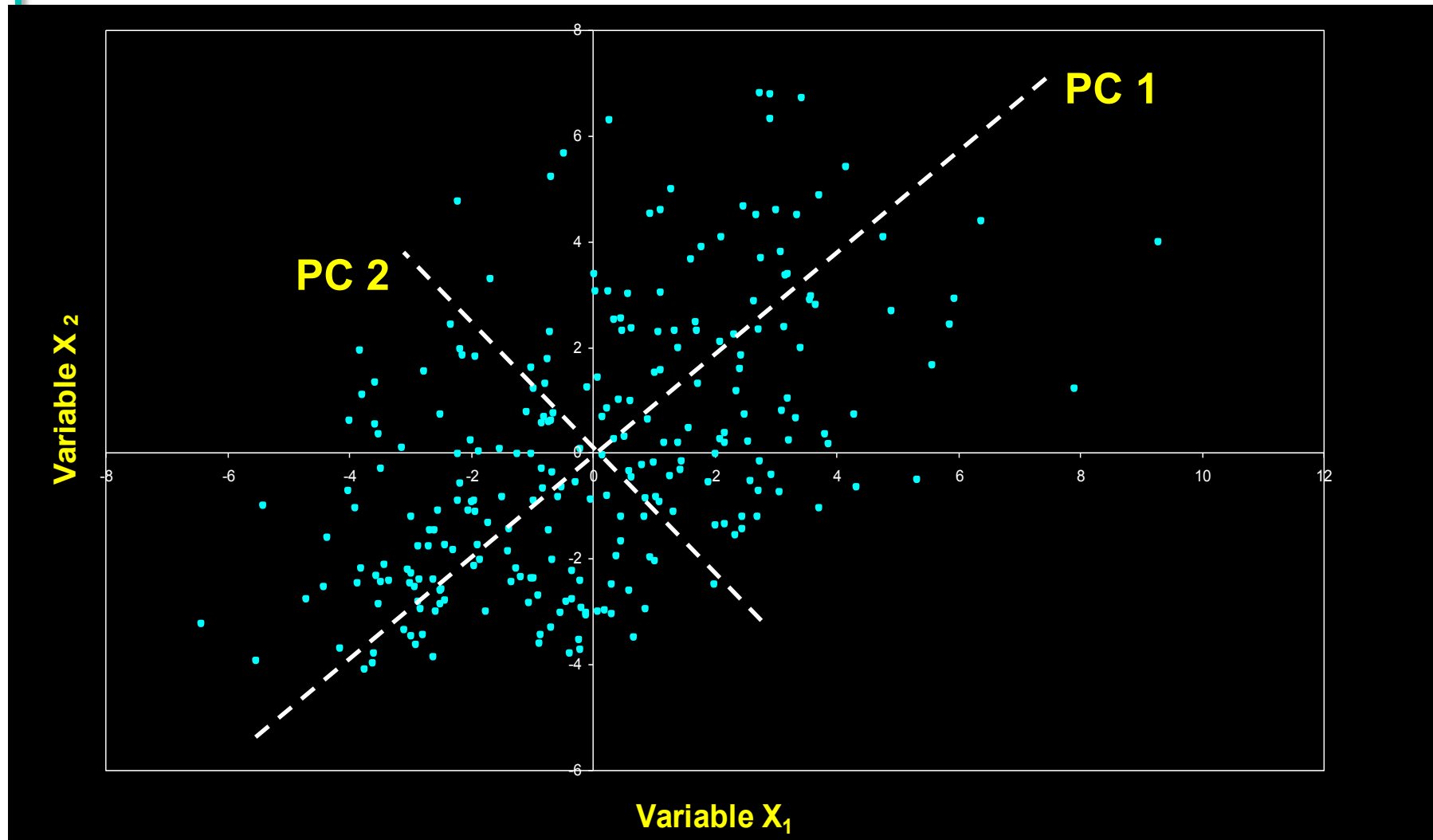


Cálculo dos (eixos) Componentes Principais



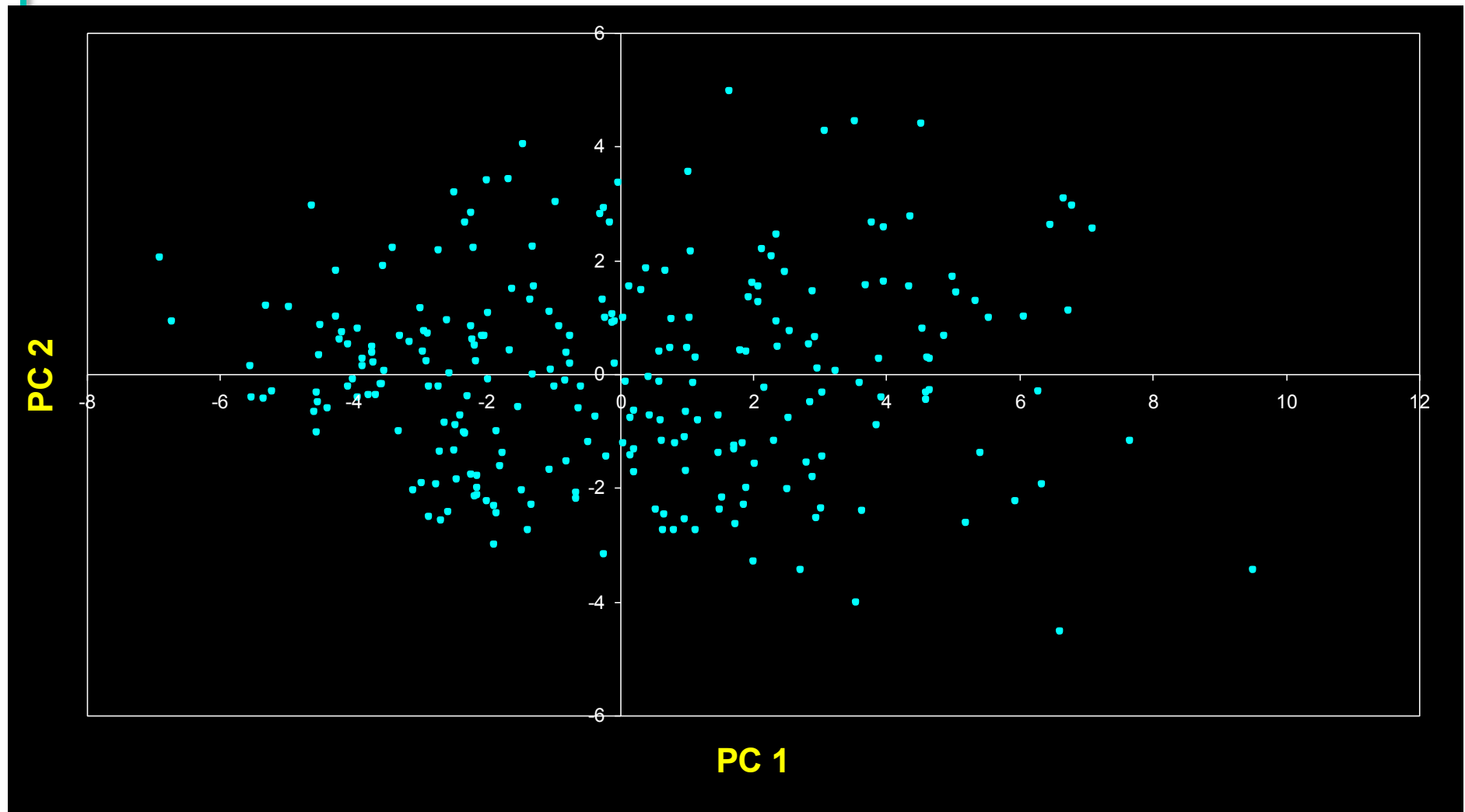
PC1 tem a maior variância (9.88); PC2 a menor (3.03); covariância = 0

Os eixos PCs são rotações rígidas dos originais



PC 1 é simultaneamente a direção da maior variância e a regressão linear do conjunto de pontos

Cálculo dos Componentes Principais



Generalização de PCA para n dimensões

- O uso de covariância só faz sentido se os atributos tem valores na mesma unidade de medida
- Para permitir atributos com unidades de medidas diferentes, padroniza-se

$$X'_{im} = \frac{(X_{im} - \bar{X}_i)}{SD_i}$$

média da variável i

desvio padrão da variável i

- Covariâncias entre variáveis assim padronizadas são correlações

$$r_{ij} = \frac{C_{ij}}{\sqrt{V_i V_j}}$$

covariância entre atributos i e j

variâncias dos atributos i e j

Generalização de PCA para n dimensões

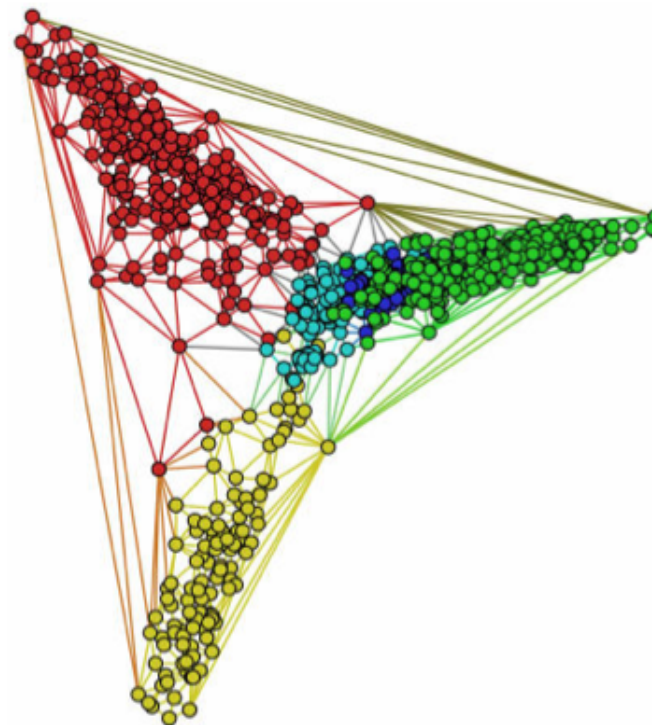
covariância

A		X_1	X_2
	X_1	6.6707	3.4170
	X_2	3.4170	6.2384

correlação

B		X_1	X_2
	X_1	1.0000	0.5297
	X_2	0.5297	1.0000

- Calcula-se os autovetores e os autovalores da matriz de correlação (ou de covariância)
 - Autovalores: variâncias em cada eixo coordenado
 - Autovetores: vetor que aplicando-se a matriz como transformação sofre apenas uma escala com fator=autovalor
- Os autovetores com maiores autovalores são os componentes principais



PCA

Conjunto de dados de 270 artigos (de 4 áreas da CC) com título, autores, afiliação, resumo e referências.

(Paulovich, 2008)

Multi-Dimensional Scaling

Young & Householder, *Psychometrika*, 1938

Shepard, *Psychometrika*, 1962

Kruskal, *Psychometrika*, 1964

Entrada:

- $\{\mathbf{p}_i\}$ N pontos multidimensionais
- $d_{ij} = d(\mathbf{p}_i, \mathbf{p}_j)$ distância (por exemplo, *Euclidiana*)

Saída:

- $\{\mathbf{x}_i\}$ N pontos 2D que melhor preservam “distância”:

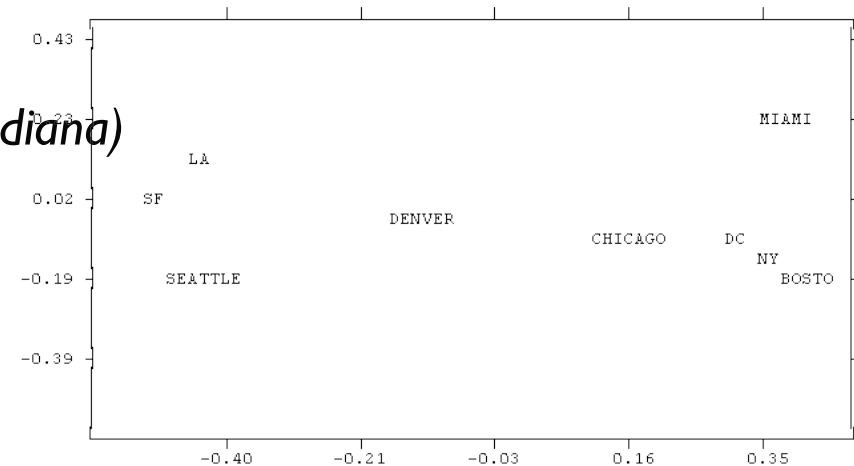
$$\|\mathbf{x}_i - \mathbf{x}_j\| \approx d_{ij}$$

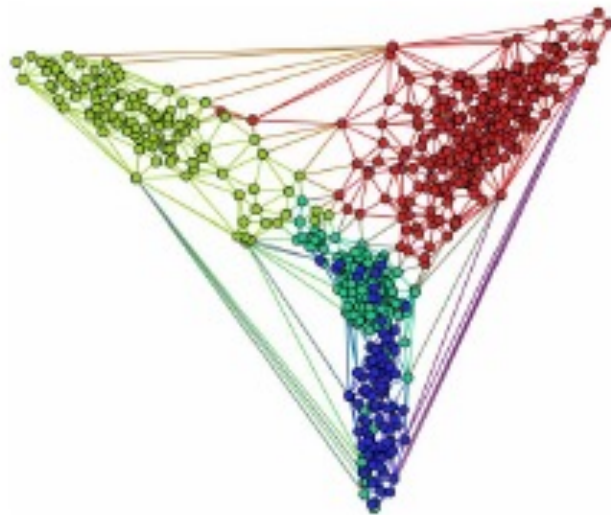
Em linhas gerais:

- Minimiza uma função

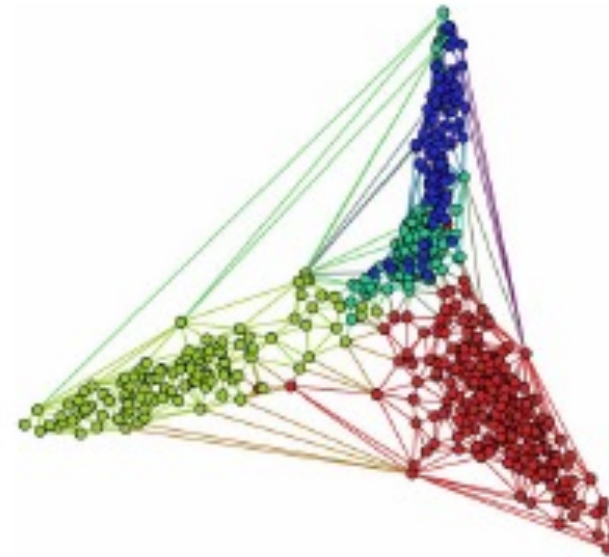
$$s = \sum \sum (\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij})^2$$

	1	2	3	4	5	6	7	8	9
	BOST	NY	DC	MIAM	CHIC	SEAT	SF	LA	DENV
1 BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2 NY	206	0	233	1308	802	2815	2934	2786	1771
3 DC	429	233	0	1075	671	2684	2799	2631	1616
4 MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5 CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6 SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7 SF	3095	2934	2799	3053	2142	808	0	379	1235
8 LA	2979	2786	2631	2687	2054	1131	379	0	1059
9 DENVER	1949	1771	1616	2037	996	1307	1235	1059	0





MDS



PCA

Conjunto de dados de 675 artigos (de 4 áreas da CC) com título, autores, afiliação, resumo e referências.

(Paulovich, 2008)

Técnicas não-lineares

t-SNE (van der Maaten e Hinton, 2008)

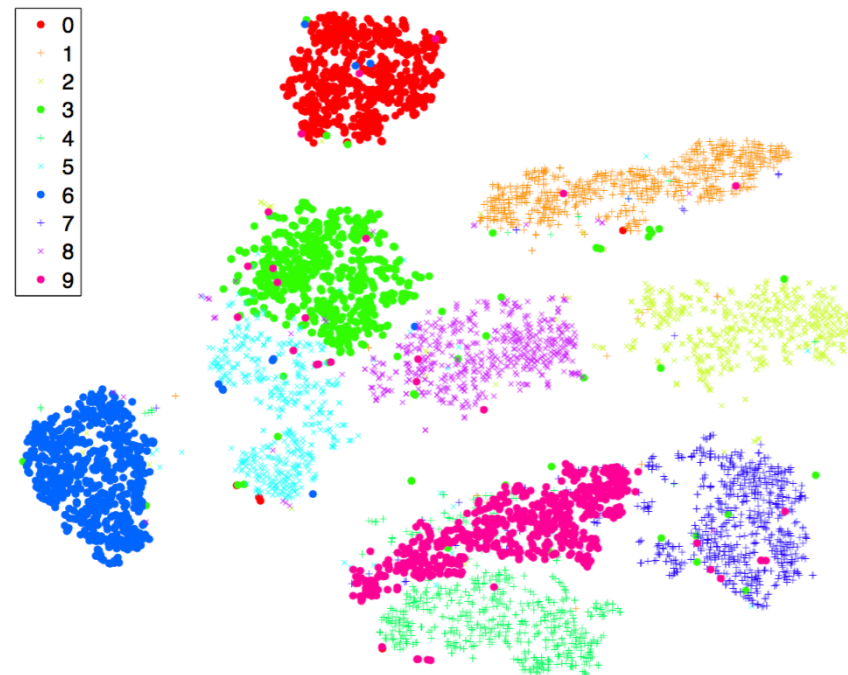
- Variação da técnica original SNE (Stochastic Neighbor Embedding) proposta por Hinton e Roweis (2002)
- *“A way of converting a high-dimensional data set into a matrix of pair-wise similarities for visualizing the resulting similarity data. t-SNE is capable of capturing much of the **local structure** of the high-dimensional data very well, while also revealing **global structure** such as the presence of clusters at several scales.”*

SNE

- Stochastic Neighbor Embedding (SNE) starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities.
- The similarity of data point x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_j would pick x_i as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i .
 - A probabilidade condicional dos pontos no espaço DR deve ser o mais próxima possível da no espaço original
 - Uma função “custo” é otimizada

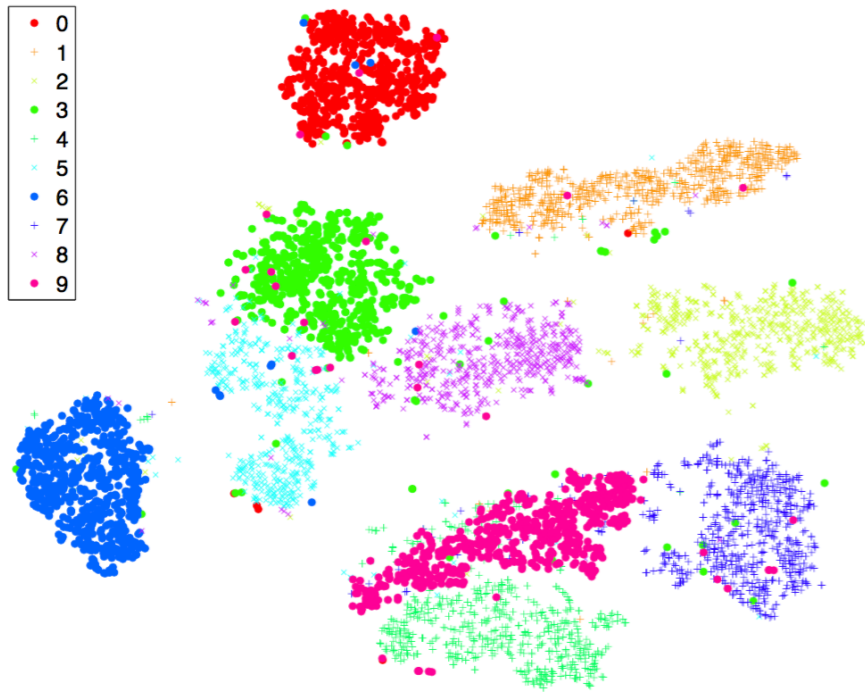
t-SNE: t-Distributed Stochastic Neighbor Embedding

- Usa uma distribuição t-Student para calcular a similaridade no espaço DR e não uma Gaussiana
- Usa uma função custo diferente a ser otimizada

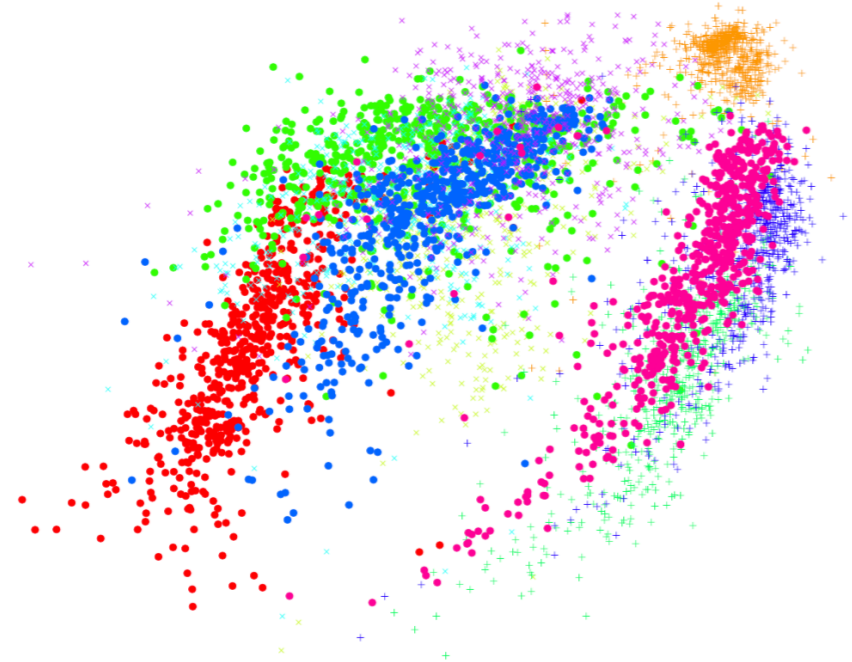


(a) Visualization by t-SNE.

Comparação t-SNE e ISOMAP

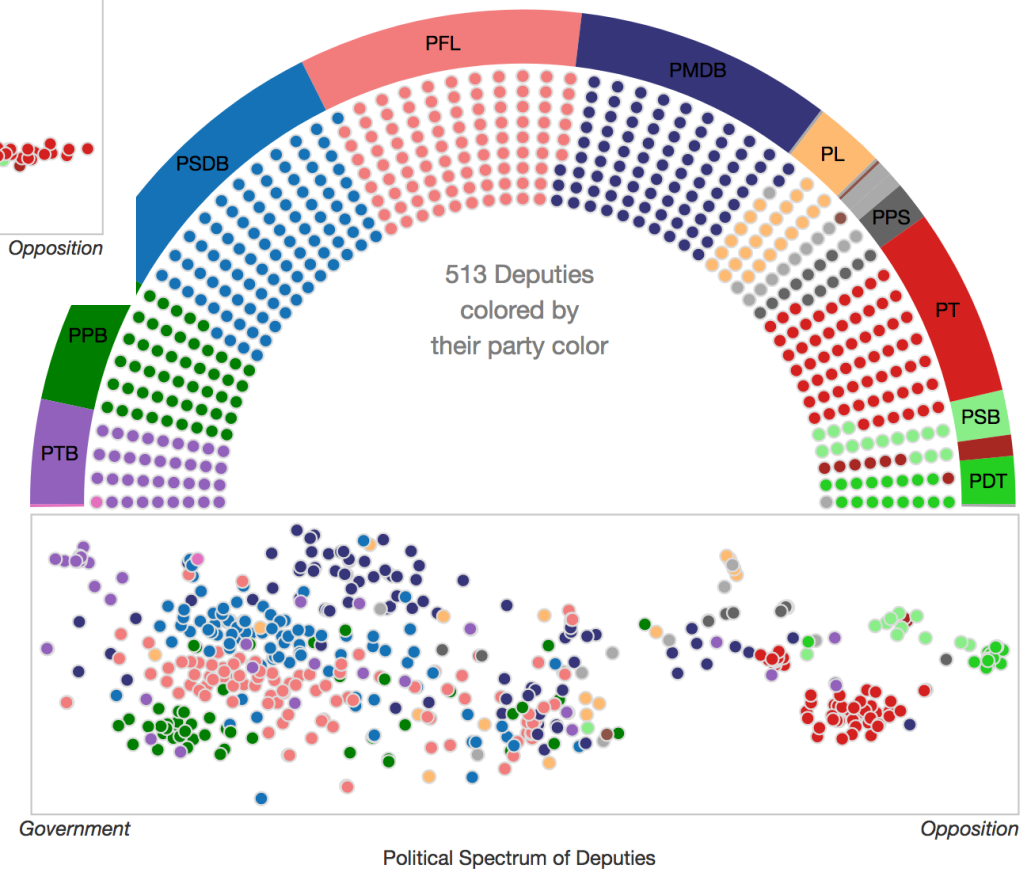
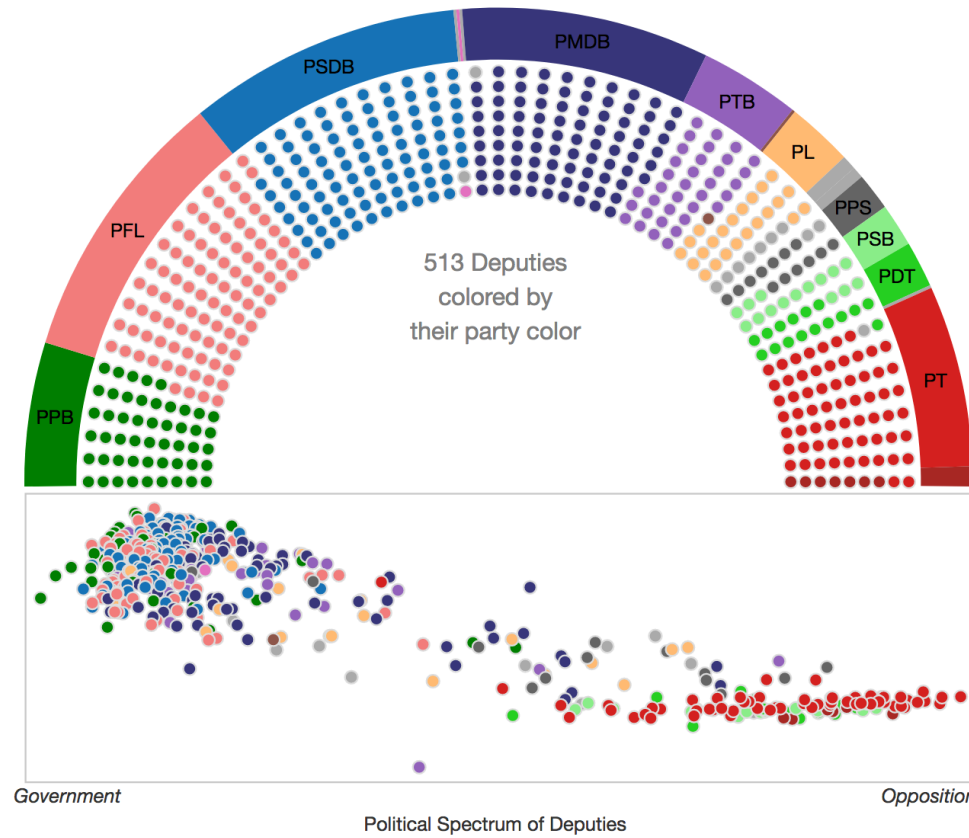


(a) Visualization by t-SNE.

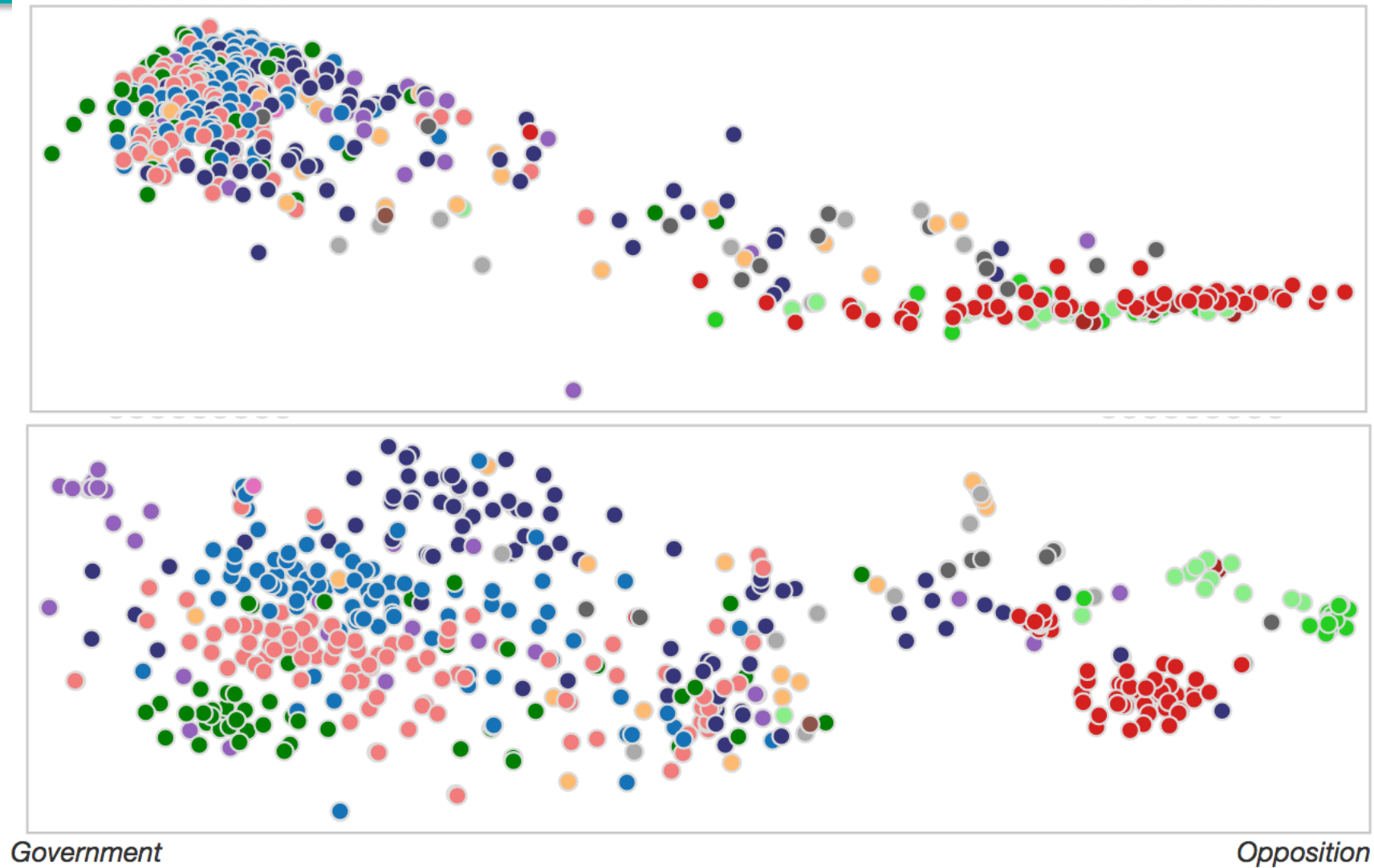


(a) Visualization by Isomap.

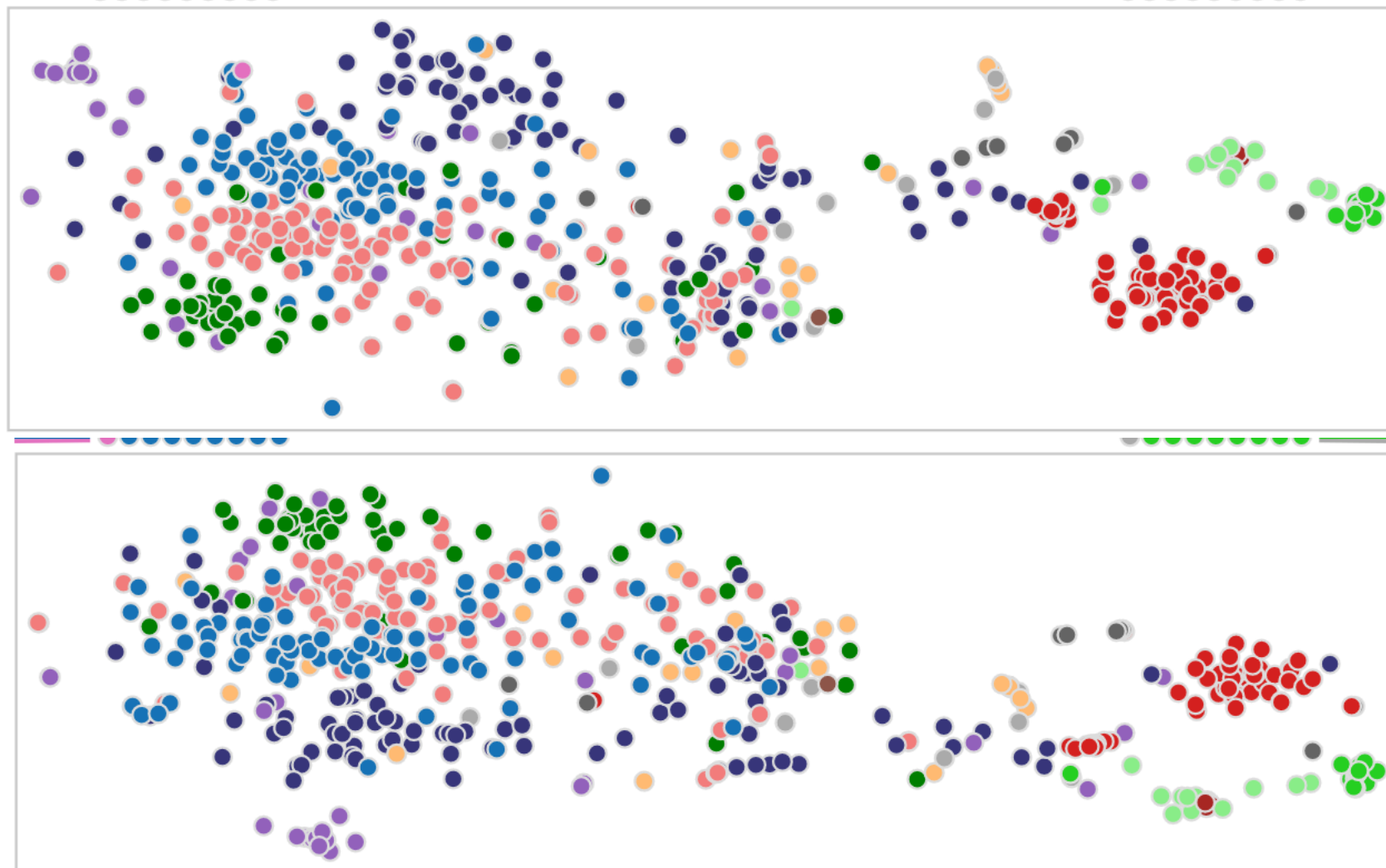
PCA x t-SNE



PCA x t-SNE



Duas execuções de t-SNE



Government

Political Spectrum of Deputies

Opposition

Uma boa (mas breve) introdução ao assunto

- Ward, Grinstein and Keim. Interactive Data Visualization. Cap. 7
- Artigos selecionados estão no Moodle