Principal Component Analysis (PCA)

Tiago Cardoso Botelho

Departamento de Economia
Universidade of São Paulo

6 de novembro de 2020

Plano de Ataque

Motivação
Um Exemplo

Preferências Sociais

Vamos considerar o seguinte **problema de escolha social**: quatro amigos vão a um *shopping center* e precisam decidir onde irão almoçar. Eles reduziram as escolhas a quatro opções viáveis e suas preferências podem ser sintetizadas no quadro abaixo.

	Casa RAW	McDonald's	Galeto's	Animal Chef
Alice	10	1	2	7
Bia	7	2	1	10
Carlos	2	9	7	3
Diego	3	6	10	2

PCA

Alguns questionamentos:

- Será que conseguimos **visualizar** este objeto de dimensão alta ($m = 4$) em apenas duas dimensões?
- O que será que está por trás dessas valorações? Conseguimos as **decompor** em $k < m$ componentes?

PCA

A resposta é **sim** para ambas as perguntas. Vamos definir:

$$\overline{\mathbf{x}} = (5.5, 4.5, 5, 5.5),$$

que tem como componentes as **médias aritméticas** das valorações em cada restaurante. Em seguida, vamos **aproximar** cada pessoa (cada linha) pelo vetor:

$$\overline{\mathbf{x}} + \theta_1^i \cdot \mathbf{v}_1 + \theta_2^i \cdot \mathbf{v}_2$$

onde $\mathbf{v}_1 = (3, -3, -3, 3)$ e $\mathbf{v}_2 = (1, -1, 1, -1)$. O sobrescrito i indexa o *decision-maker* (DM).

Pattern Recognition

Os vetores $\mathbf{v}_1$ e $\mathbf{v}_2$ *não* são arbitrários. Intuitivamente, o que vocês acham que estes vetores deveriam **sintetizar**?

Pattern Recognition

Resposta: eles representam:

1. o grau de **veganismo** de cada DM;
2. o grau de preocupação com **alimentação saudável** de cada DM.

Componentes Principais

Por ora, vamos nos contentar em simplesmente dizer que os coeficientes de cada DM são:

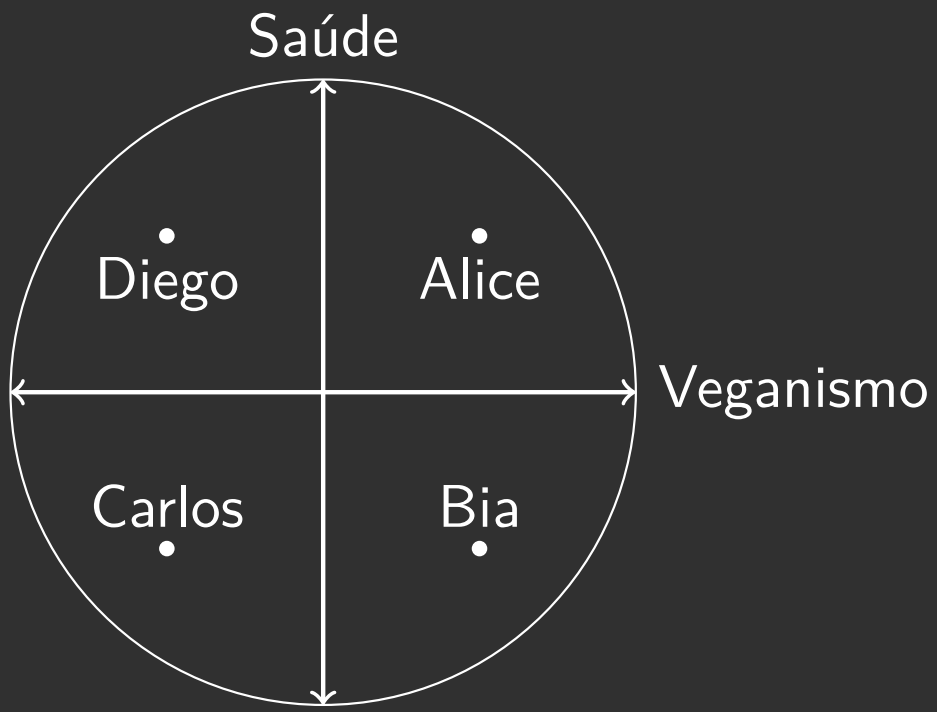
	θ_1	θ_2
Alice	1	1
Bia	1	-1
Carlos	-1	-1
Diego	-1	1

Alice é vegana e saudável $(1, 1)$. Note que, para a Alice, temos

$$\overline{\mathbf{x}} + \mathbf{v}_1 + \mathbf{v}_2 = (9.5, 0.5, 3, 7.5),$$

que é aproximadamente o vetor de valorações dela.

Visualização



Observações

Observe que:

- A **primeira** e a **quarta** componentes de $\mathbf{v}_1$ são positivas e comparativamente grandes. Em particular, elas são positivamente correlacionadas; ora, isto é claro, pois, os restaurantes 1 (Casa RAW) e 4 (Animal Chef) são veganos!
- A **segunda** e a **terceira** componentes de $\mathbf{v}_1$ são negativas e comparativamente grandes (em módulo). Em particular, elas são positivamente correlacionadas (entre si) e negativamente correlacionadas (com as outras componentes). Mas é claro! Elas estão associadas a opções carnívoras!

Praticando

Exercício 1.

Compute os vetores de aproximação para Bia, Carlos e Diego.

Exercício 2.

Repita a análise do *frame* anterior para a componente $\mathbf{v}_2$ das preferências individuais.

Formulação do Problema

PCA: Objetivo

Em PCA, temos m vetores $\mathbf{x}_1, ..., \mathbf{x}_m$ de dimensão n.

Queremos escrevê-los como uma combinação linear de $k < m$ vetores $\mathbf{v}_1, ..., \mathbf{v}_k$ de dimensão n para **aproximá-los** de alguma maneira.

$$\mathbf{x}_i = \sum_{j=1}^{k} a_{i,j} \cdot \mathbf{v}_j.$$

PCA: Primeiro Passo

1. Da visualização que produzimos no exemplo, é claro que gostaríamos que os vetores $\mathbf{x}_i$ estivessem centrados na origem, ou seja, que eles satisfizessem:

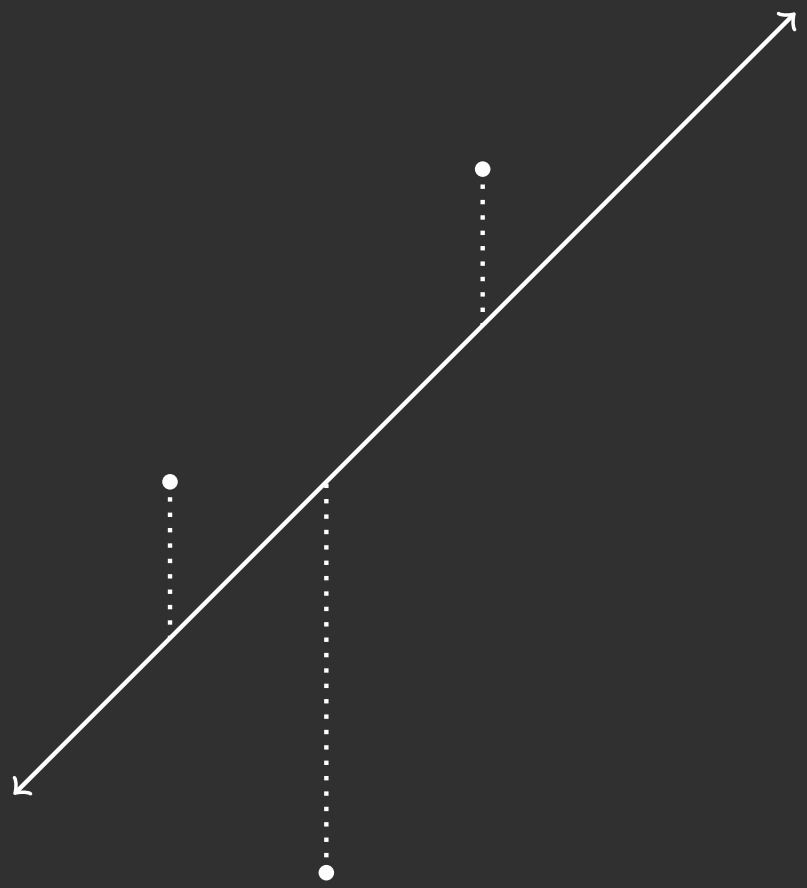
$$\sum_{i=1}^{m} \mathbf{x}_i = \mathbf{0}.$$

Se este já não for o caso, simplesmente subtraia

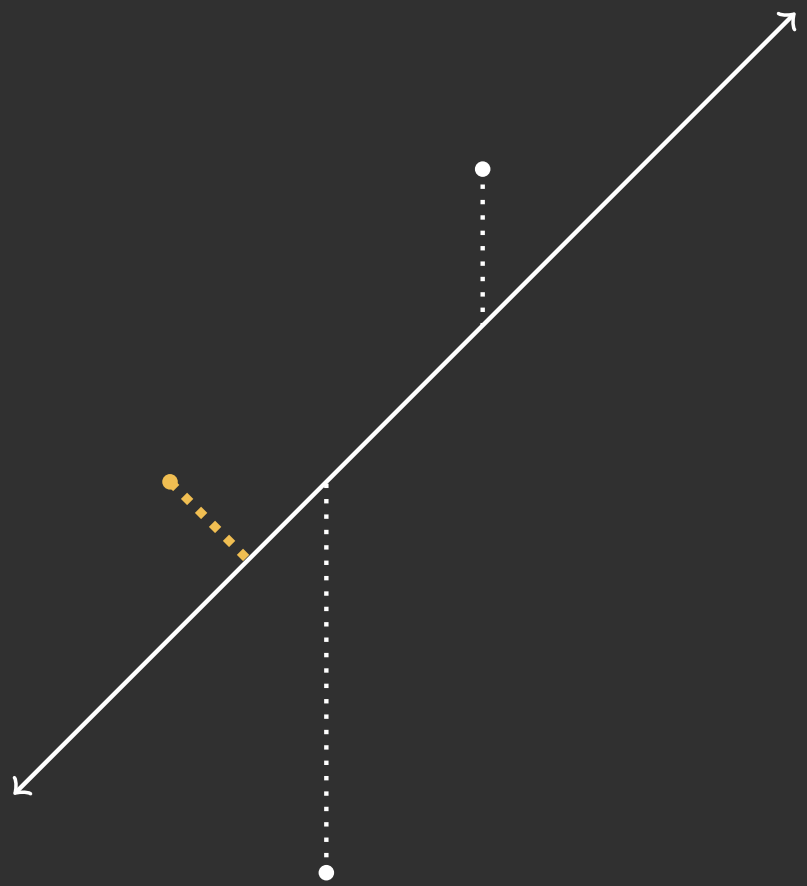
$$\overline{\mathbf{x}} = \frac{1}{m} \cdot \sum_{i=1}^{m} \mathbf{x}_i$$

de cada um deles.

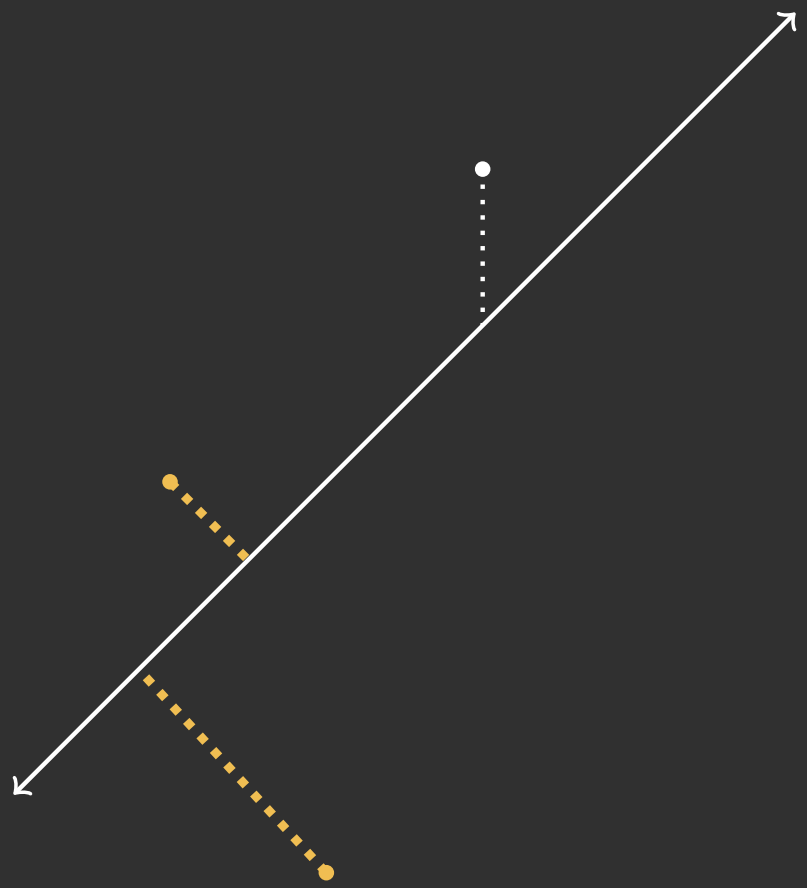
Regressão Linear v. PCA



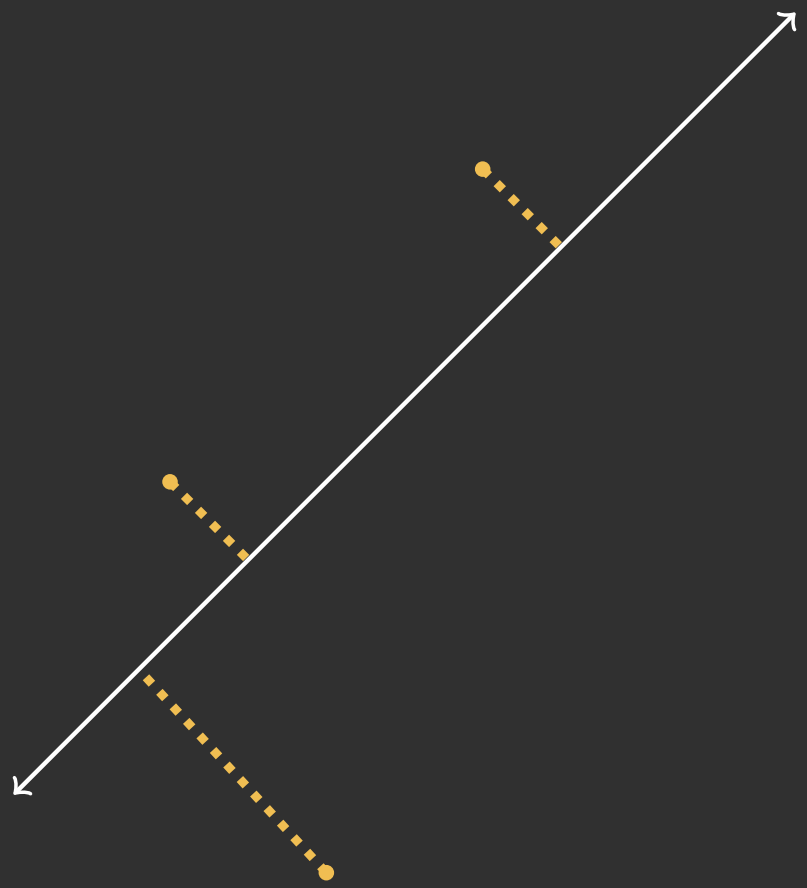
Regressão Linear v. PCA



Regressão Linear v. PCA



Regressão Linear v. PCA



Regressão Linear v. PCA

- Regressão linear minimiza a soma do **quadrado dos erros na vertical**.
- PCA minimiza a soma dos **quadrados dos erros na perpendicular**.

PCA: Segundo Passo

2. Já subtraímos a média dos vetores, de maneira a centrá-los no vetor nulo. Agora, vamos dividir cada coordenada pelo desvio padrão (amostral, é claro) naquela coordenada:

$$s_j = \sqrt{\sum_{i=1}^{m} \mathbf{x}_{i,j}^2}.$$

Com isso conseguimos mitigar a sensibilidade do PCA às unidades de medida de cada coordenada. **Convença-se disso**.

PCA: Terceiro Passo

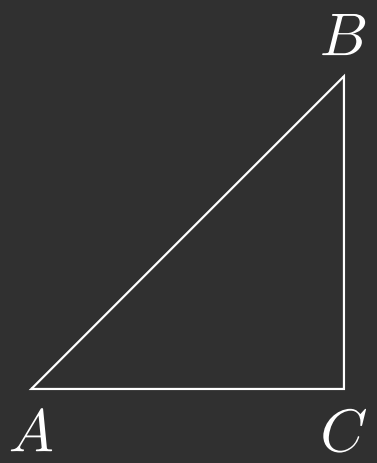
3. Vamos discutir aqui apenas o caso $k = 1$ para fixarmos as ideias. Cabe a você generalizar na prova. Queremos um vetor **unitário** $\mathbf{v}$ que minimiza:

$$\frac{1}{m} \cdot \sum_{i=1}^{m} \mathrm{d}(\mathbf{x}_i, \text{span } \mathbf{v})^2,$$

onde span $\mathbf{v}$ representa o subespaço gerado por $\mathbf{v}$.

PCA: Terceiro Passo

Construa o seguinte triângulo engenhoso:



onde AB $= ||\mathbf{x}_i||$, BC $= d(\mathbf{x}_i, \text{span } \mathbf{v})$ e AC $= \langle \mathbf{x}_i, \mathbf{v} \rangle$. Temos:

$$\text{AB}^2 = \text{AC}^2 + \text{BC}^2,$$

logo:

$$||\mathbf{x}_i||^2 = d(\mathbf{x}_i, \text{span } \mathbf{v})^2 + \langle \mathbf{x}_i, \mathbf{v} \rangle^2.$$

PCA: Terceiro Passo

(...)

Logo, minimizar a expressão do terceiro passo equivale a maximizar:

$$\frac{1}{m} \cdot \sum_{i=1}^{m} \langle \mathbf{x}_i, \mathbf{v} \rangle^2.$$

Então estamos maximizando a **variância** da projeção do ponto sobre o subespaço gerado. Isto quer dizer que acreditamos que a variabilidade nos dados armazena alguma informação valiosa (é um **sinal**, e não apenas **ruído**).

Praticando

Exercício 3

Implemente o algoritmo descrito acima em `Python`. Você pode usar o `numpy` se quiser, mas não é exatamente necessário.

Obrigado!