

Análise de Componentes Principais (PCA)

Álgebra Linear Computacional - Fabricio Murai

Aula anterior

- Revisão sobre posto de matrizes **Posto k ?**
- Relação entre posto e valores singulares **Qual a relação?**
- SVD reduzido, SVD truncado e SVD randomizado **Complexidade?**

Aula de hoje

- O que é PCA?
- Visão geral
- Exemplos
- Fundamentos teóricos do PCA

Análise de componentes principais (PCA)

Técnica de análise de dados baseada em decomposição de matrizes que transforma **conjunto de variáveis possivelmente correlacionadas** em **conjunto menor de variáveis** chamadas **componentes principais**

- Um dos resultados mais importantes de álgebra linear
- Primeiro passo ao analisar grandes datasets
- Outras aplicações: remover ruído de imagens, compressão de dados



PCA: visão geral

- Transforma o espaço vetorial dos dados para reduzir a dimensionalidade de grandes conjuntos de dados.
- Usando uma projeção, dados originais (muitas variáveis) podem ser interpretados facilmente (poucas variáveis)
 - E.g.: encontrar tendências, padrões e outliers

Aula de hoje

- O que é PCA?
- Visão geral
- Exemplos
- Fundamentos teóricos do PCA

Exemplo: automóveis

Dataset $\{x^{(i)}: i=1, \dots, m\}$ de atributos de m tipos de automóveis.

Atributos: velocidade máxima, raio de giro, etc.

Seja $x^{(i)} \in \mathbb{R}^n$, i.e., n atributos.

Suponha que entre atributos, temos vel. max. em km/h e mph. Estes atributos são quase linearmente dependentes (erros de arredondamento).

Ou seja: dados estão aproximadamente em espaço de $n-1$ dimensões!

Como detectar
redundância
automaticamente?

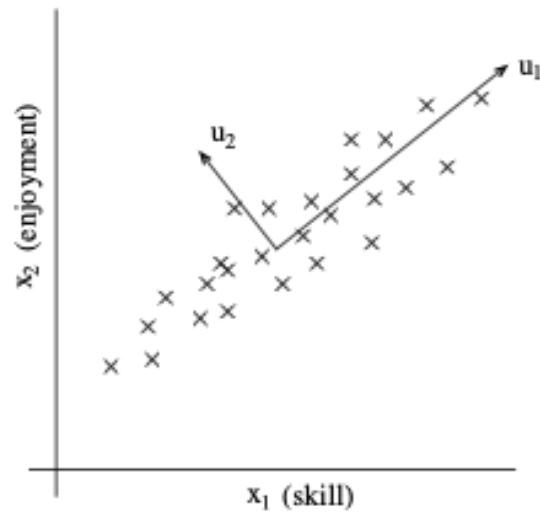
Exemplo: pilotos de helicóptero de controle remoto

$x_1^{(i)}$ é uma medida da habilidade do piloto i

$x_2^{(i)}$ é uma medida de quanto ele gosta de voar

Suponha que apenas pilotos que treinam bastante, aqueles que realmente gostam de voar, se tornam bons pilotos. Logo, x_1 e x_2 estão fortemente correlacionados.

Podemos conjecturar que os dados estão em um eixo "diagonal", com apenas um pouco de ruído saindo do eixo.



Como determinar esta direção u_1 ?

Pré-processamento dos dados

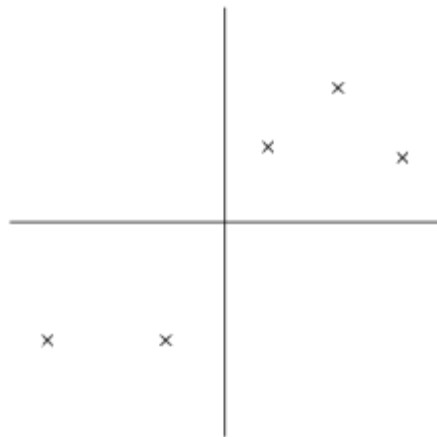
Passos 3-4: garantem que atributos sejam tratados igualmente (e.g., vel. max. e número de assentos)

Primeiro normalizamos a média e a variância:

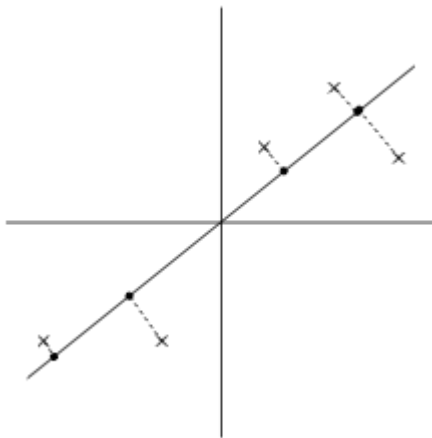
1. Seja $\mu = \frac{1}{m} \sum_{i=1}^m \mathbf{x}^{(i)}$
2. Substitua cada $\mathbf{x}^{(i)}$ por $\mathbf{x}^{(i)} - \mu$
3. Seja $\sigma_j^2 = \frac{1}{m-1} \sum_i (x_j^{(i)})^2$
4. Substitua cada $x_j^{(i)}$ por $x_j^{(i)} / \sigma_j$

Como calcular o maior eixo de variação?

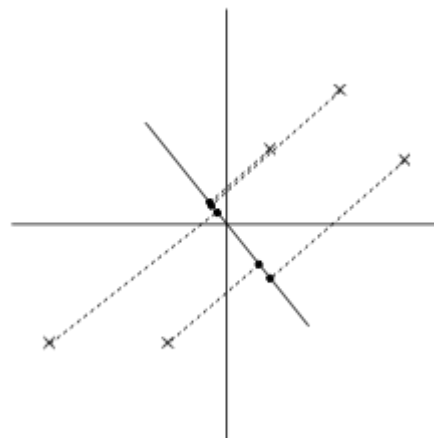
Problema: encontrar vetor unitário u que maximiza a variância dos dados projetados em u .



dados normalizados



muita variação 📈



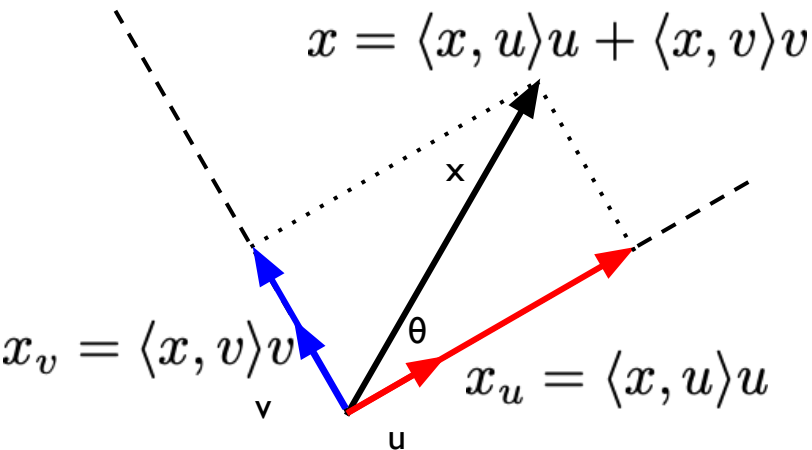
pouca variação 📉

(GAAL) projeções ortogonais

Hora do quiz

Por que "ortogonais"?

Deveriam ser chamadas projeções sobre vetores ortogonais.



$$\begin{aligned}\|x_u\| &= \|x\| \cos \theta \\ &= \|x\| \frac{\langle x, u \rangle}{\|x\| \|u\|} \\ &= \langle x, u \rangle\end{aligned}$$

Na base canônica, $x = (x_1, x_2)$.

Na base $\{u, v\}$, temos

$$x = (\langle x, u \rangle, \langle x, v \rangle)$$

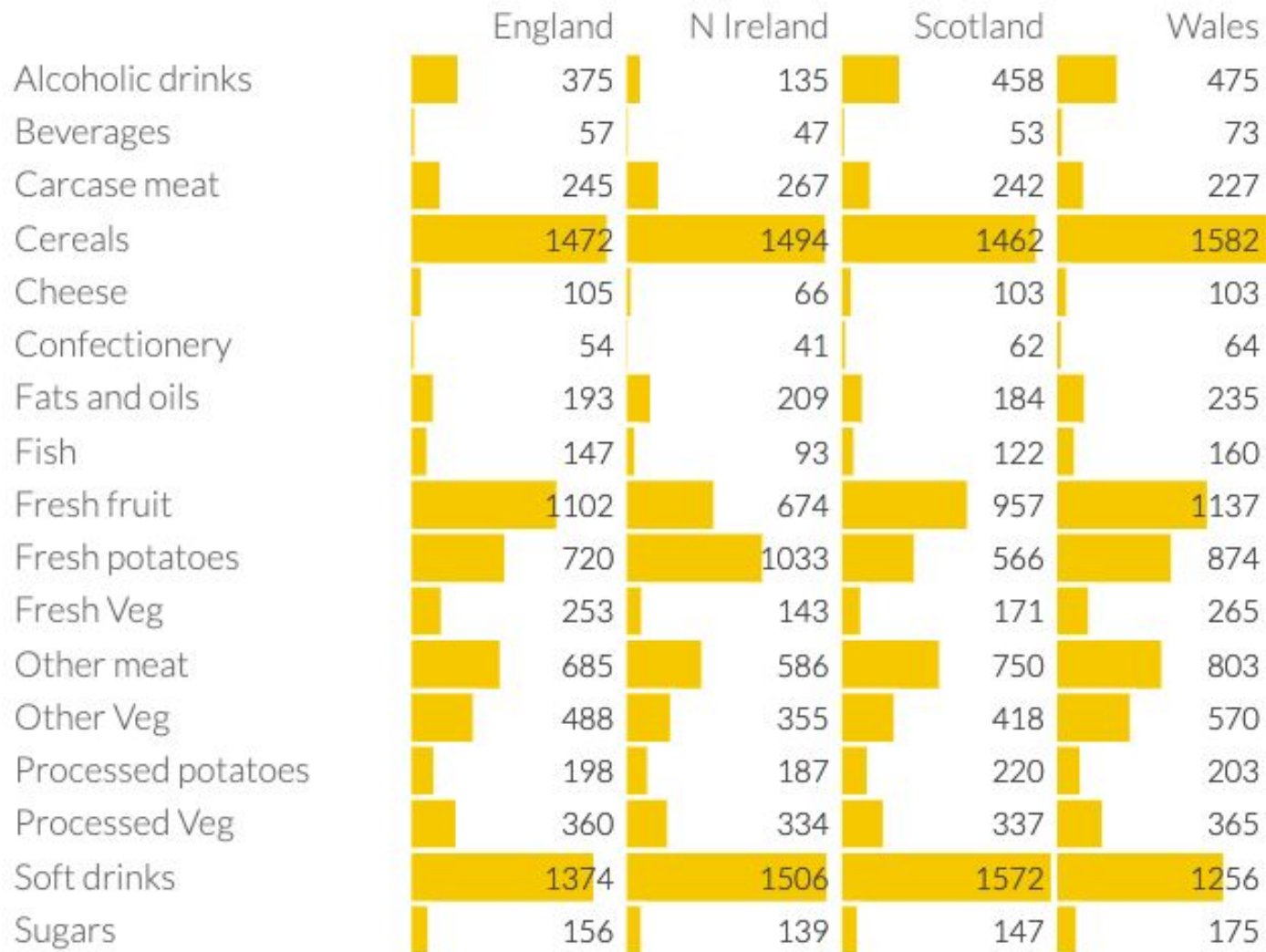
Exemplo de análise de dados multivariada

DEFRA¹ data: consumo (g/pessoa/semana), 17 tipos de comida, quatro países Reino Unido em 1997.

	England	Wales	Scotland	N Ireland
Cheese	105	103	103	66
Carcass meat	245	227	242	267
Other meat	685	803	750	586
Fish	147	160	122	93
Fats and oils	193	235	184	209
Sugars	156	175	147	139
Fresh potatoes	720	874	566	1033
Fresh Veg	253	265	171	143

	England	Wales	Scotland	N Ireland
Other Veg	488	570	418	355
Processed potatoes	198	203	220	187
Processed Veg	360	365	337	334
Fresh fruit	1102	1137	957	674
Cereals	1472	1582	1462	1494
Beverages	57	73	53	47
Soft drinks	1374	1256	1572	1506
Alcoholic drinks	375	475	458	135
Confectionery	54	64	62	41

¹Department for Environment, Food and Rural Affairs

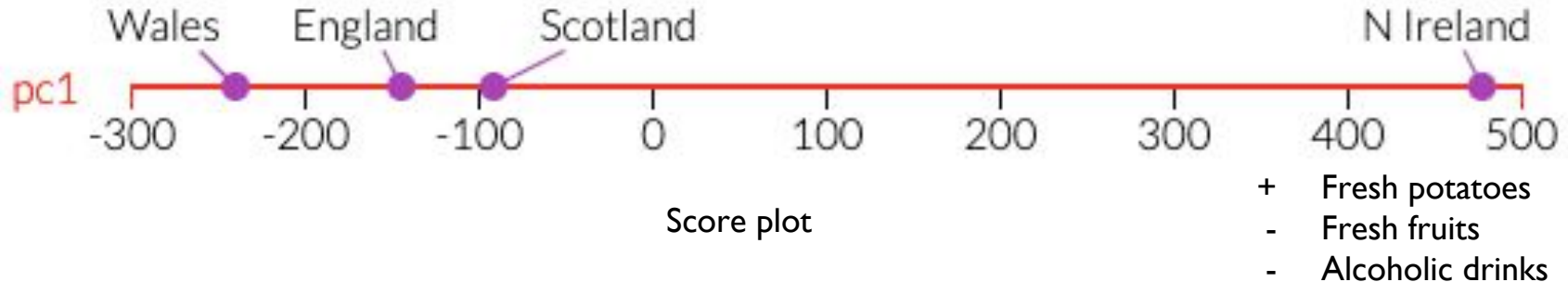


Como analisar
estes dados?
Existem
correlações
entre os países?
Plotar todos de
uma vez? Plotar
variáveis 2 a 2?

PCA: primeira coordenada principal

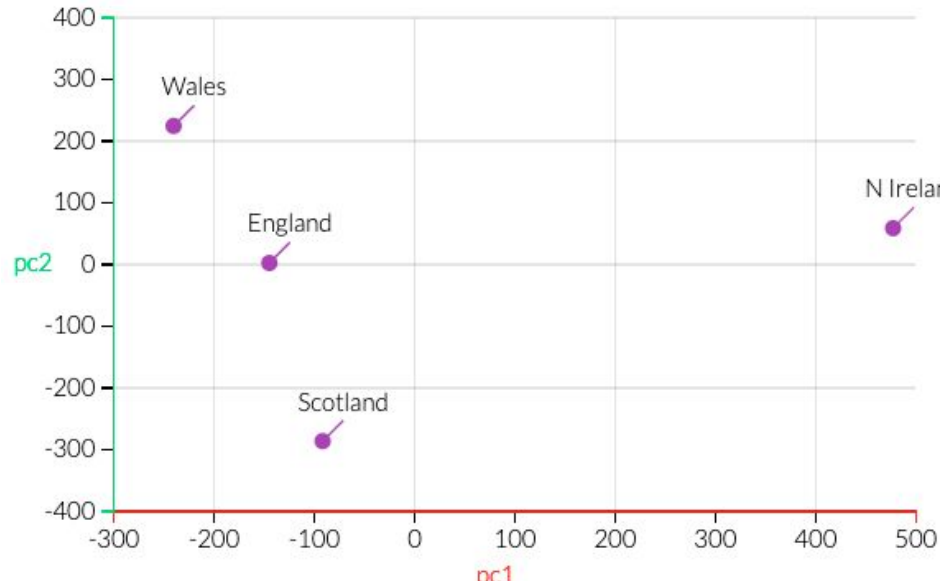
O objetivo do PCA é identificar novo conjunto de eixos ortogonais de coord. a partir dos dados. No primeiro passo, PCA encontra a **primeira componente principal (PCI)**, direção que maximiza variância através das 17 coords.

Após encontrar PCI, projetamos os dados neste novo eixo. Como?



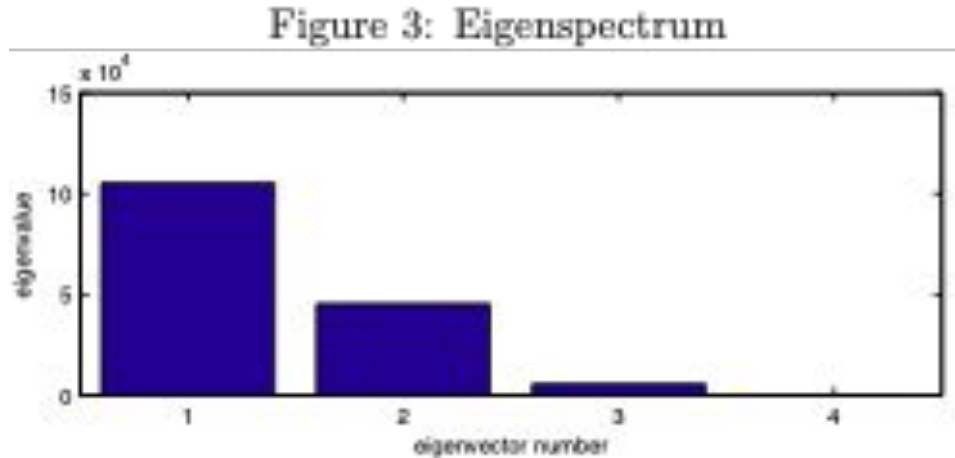
PCA: segunda componente principal

- PC2 é ortogonal a PC1
- Direção de maior variância dos dados, dentre direções ortogonais a PC1
- De novo, projetar coordenadas em PC2



PCA: variância explicada

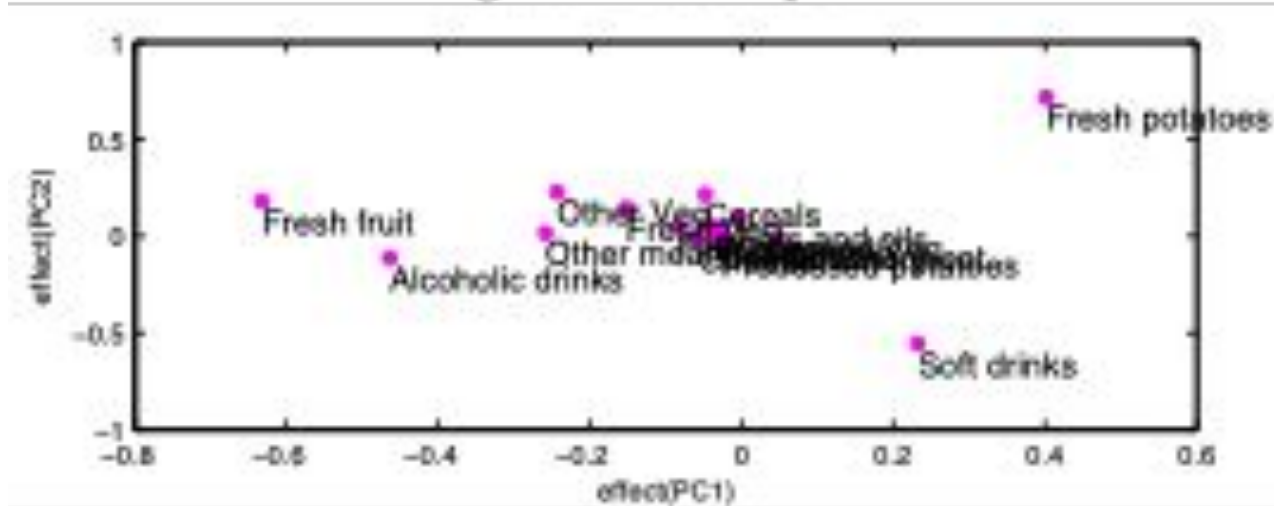
PCA também retorna info sobre contribuição de cada PC para a variância total dos dados: 67% por PC1, 97% por PC1+PC2



PCA: composição de cada PC

Também obtemos influência de cada variável original nos PCs

Figure 4: Load plot



Aula de hoje

- O que é PCA?
- Visão geral
- Exemplos
- Fundamentos teóricos do PCA

Fundamentos teóricos do PCA

Problema resolvido pelo PCA:

podemos obter outra base que é combinação linear da base original e re-expressar dados de maneira **ótima**?

Seja X uma matriz $m \times n$ onde colunas = observações, linhas = vars

Queremos transformar X em matriz $Y_{m \times n}$ usando $P_{m \times m}$:

$$Y = PX$$

(Mudança
de base)

Mudança de base $Y=PX$

Sejam p_1, p_2, \dots, p_m os vetores linha de P e x_1, \dots, x_n os vetores coluna de X . Temos

$$PX = \begin{pmatrix} P x_1 & P x_2 & \dots & P x_n \end{pmatrix} = \begin{pmatrix} p_1 \cdot x_1 & p_1 \cdot x_2 & \dots & p_1 \cdot x_n \\ p_2 \cdot x_1 & p_2 \cdot x_2 & \dots & p_2 \cdot x_n \\ \vdots & \vdots & \ddots & \vdots \\ p_m \cdot x_1 & p_m \cdot x_2 & \dots & p_m \cdot x_n \end{pmatrix} = Y$$

As colunas de X estão sendo projetadas nas linhas de P . Logo, as linhas de P $\{p_1, p_2, \dots, p_m\}$ são uma nova base para colunas de X . Linhas de P serão as direções das PCs.

Otimização das PCs: maximizar variância

PCA considera variância dos dados originais. Tenta descorrelacionar dados usando como base direções em que variância é maximizada.

Seja variável Z com média μ . Amostras de Z : $\mathbf{z}=\{z_1, z_2, \dots, z_n\}$.

Variância amostral de Z :

$$\text{var}(\mathbf{z}) = \frac{1}{n-1} \sum_{i=1}^n (z_i - \mu)^2$$

Subtraindo-se a média, ou seja, fazendo $\mathbf{r}=\{r_1, \dots, r_n\}$, onde $r_i = z_i - \mu$, temos

$$\text{var}(\mathbf{r}) = \frac{1}{n-1} \sum_{i=1}^n r_i^2 = \frac{1}{n-1} \mathbf{r} \mathbf{r}^\top$$

Otimização das PCs: maximizar variância

Seja $\mathbf{s} = (s_1, \dots, s_n)$ um segundo vetor de medições com média zero, podemos generalizar a ideia anterior para covariância de \mathbf{r} e \mathbf{s} .

Interpretação: quanto duas variáveis mudam simultaneamente.

$$\text{cov}(\mathbf{r}, \mathbf{s}) = \frac{1}{n-1} \mathbf{r} \mathbf{s}^\top$$

Murai, não entendi o que é a covariância de
duas variáveis. :(

Otimização das PCs: maximizar variância

Generalizando para matrix $X_{m \times n}$ de dados (m vars, n observações)

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{pmatrix} \in \mathbb{R}^{m \times n}, \quad \mathbf{x}_i^T \in \mathbb{R}^n$$

\mathbf{x}_i : vetor de n samples da i-ésima variável

$$C_X = \frac{1}{n-1} X X^T = \frac{1}{n-1} \begin{pmatrix} \mathbf{x}_1 \mathbf{x}_1^T & \mathbf{x}_1 \mathbf{x}_2^T & \cdots & \mathbf{x}_1 \mathbf{x}_m^T \\ \mathbf{x}_2 \mathbf{x}_1^T & \mathbf{x}_2 \mathbf{x}_2^T & \cdots & \mathbf{x}_2 \mathbf{x}_m^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_m \mathbf{x}_1^T & \mathbf{x}_m \mathbf{x}_2^T & \cdots & \mathbf{x}_m \mathbf{x}_m^T \end{pmatrix} \in \mathbb{R}^{m \times m}$$

Matriz de
covariância

variâncias

Voltando a transformação $Y=PX$

Perguntas que precisamos responder:

- Quais são propriedades desejáveis da matriz transformada Y ?
- Qual a relação com a matriz de covariância C_Y ?

Covariância mede quão bem correlacionadas são 2 variáveis.

Suposição fundamental do PCA: variáveis em Y são descorrelacionadas.

Porém, variâncias tão grandes quanto possível (variância pequena = ruído). Logo:

- Maximizar sinal, medido pela variância (maximizar entradas da diagonal)
- Minimizar covariância entre variáveis (minimizar outras entradas)

Conclusão: C_Y deve ser matriz diagonal!

Derivando P

Suposição extra: p_1, p_2, \dots, p_m são ortonormais.

Considere a fórmula da matriz de covariância C_Y e $Y=PX$:

$$C_Y = \frac{1}{n-1}YY^T = \frac{1}{n-1}(PX)(PX)^T = \frac{1}{n-1}(PX)(X^TP^T) = \frac{1}{n-1}P(XX^T)P^T$$

$$\text{i.e. } C_Y = \frac{1}{n-1}PSP^T \quad \text{where} \quad S = XX^T \quad \begin{matrix} S_{m \times m} \\ \text{é} \\ \text{simétrica} \end{matrix}$$

Fato: toda matriz simétrica S é ortonormalmente diagonalizável, ou seja, $S = EDE^T$, onde E são autovetores ortonormais e D é diagonal com autovalores

Derivando P

Escolhendo as linhas de P como autovetores de S, temos $P=E^T$.
Substituindo na expressão da matriz de covariância

$$\begin{aligned}C_Y &= \frac{1}{n-1} \mathbf{P} \mathbf{S} \mathbf{P}^T \\ &= \frac{1}{n-1} \mathbf{E}^T (\mathbf{E} \mathbf{D} \mathbf{E}^T) \mathbf{E}\end{aligned}$$

$\mathbf{E}_{m \times m}$ é ortonormal

Logo,

$$C_Y = \frac{1}{n-1} \mathbf{D}$$

D traz a importância de cada componente; maior variância dá origem a PC1, 2a. maior a PC2, etc

Passo-a-passo

1. Calcular autovalores e autovetores de $S=XX^T$.
2. Ordenar os autovalores em ordem decrescente e colocá-los em D
3. Construir matriz ortonormal E colocando os autovetores associados a cada autovalor na mesma ordem (1o. autovetor na 1a. coluna, 2o. autovetor na 2a. coluna, etc)

Alcançamos o objetivo de diagonalizar a matriz de covariância dos dados transformados.