Análise de Componentes Principais (PCA)

Álgebra Linear Computacional - Fabricio Murai

Créditos: material baseado nas notas de aula de Mark Richardson e de Andrew Ng (CS229)

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?

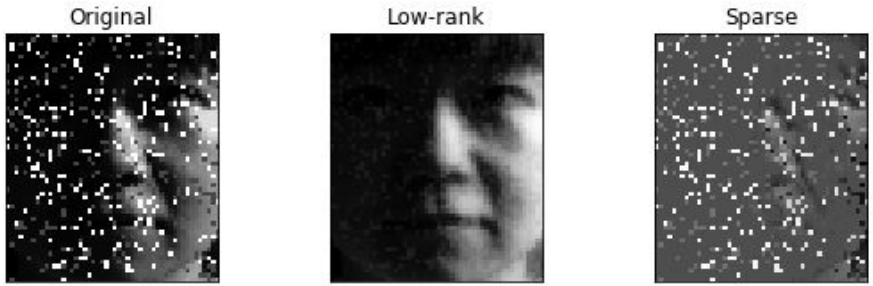
Aula de hoje

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

Atributos: velocidade máxima, raio de giro, etc.

Seja $x^{(i)} \in R^n$, 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!

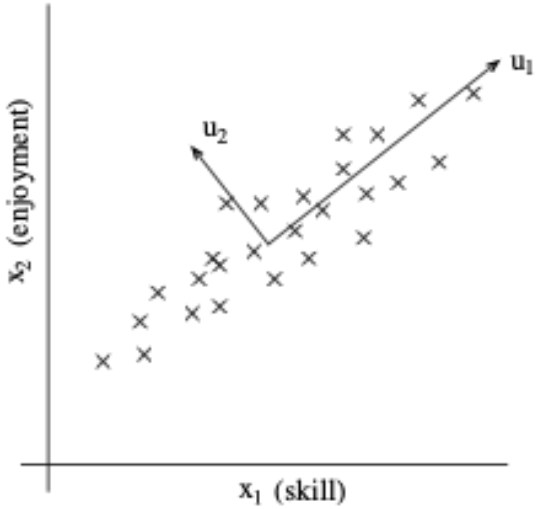
Como detectar redundância automaticamente?

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

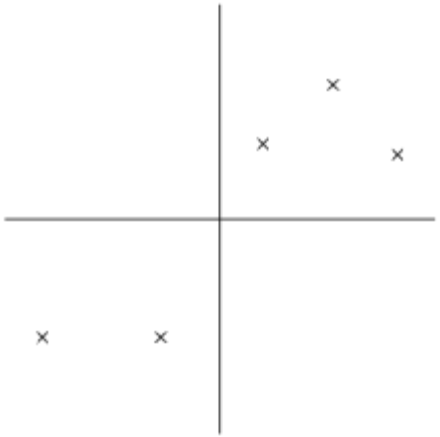
Primeiro normalizamos a média e a variância:

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 $x_j^{(i)}/\sigma_j$

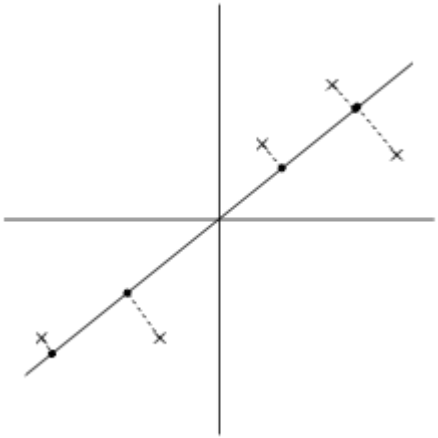
Passos 3-4: garantem que atributos sejam tratados igualmente (e.g., vel. max. e número de assentos)

Como calcular o maior eixo de variação?

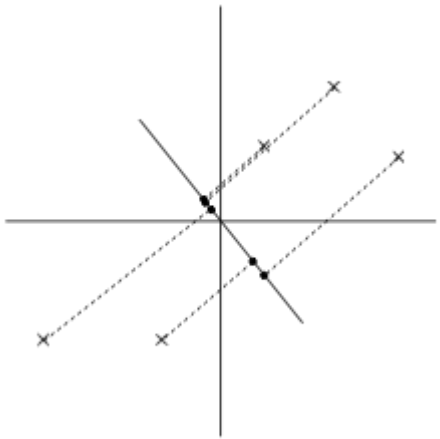
Problema: encontrar vetor unitário u que maximiza a variância dos dados projetados em u. **O que é uma projeção?**



dados normalizados



muita variação 👍



pouca variação 👎

(GAAL) projeções ortogonais

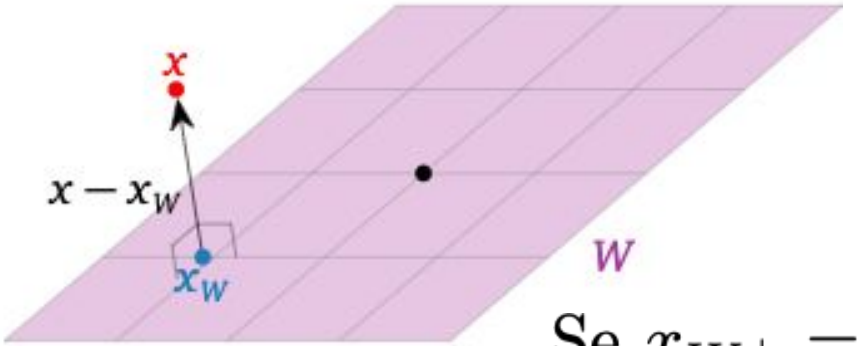
Seja W um subespaço do $\mathbf{R}^n$ e x um vetor em $\mathbf{R}^n$.

Queremos descobrir o vetor x_W mais próximo a x em W.

O vetor x_W é chamado **projeção ortogonal** de x em W.

Decomposição ortogonal

Dizer que o vetor x_W é o vetor mais próximo a x em W equivale a dizer que a diferença $x - x_W$ é ortogonal a W.



Se $x_{W^\perp} = x - x_W$, então $x = x_W + x_{W^\perp}$, onde x_W está em W e $x_{W^\perp}$ está em $W^\perp$.

Decomposição ortogonal

Teorema. Seja W um subespaço do $\mathbf{R}^n$ e x um vetor em $\mathbf{R}^n$. Então podemos escrever x de forma única como

$$x = x_W + x_{W^\perp} \tag{1}$$

Onde x_W é o vetor mais próximo a x em W e $x_{W\perp}$ está em $W^\perp$.

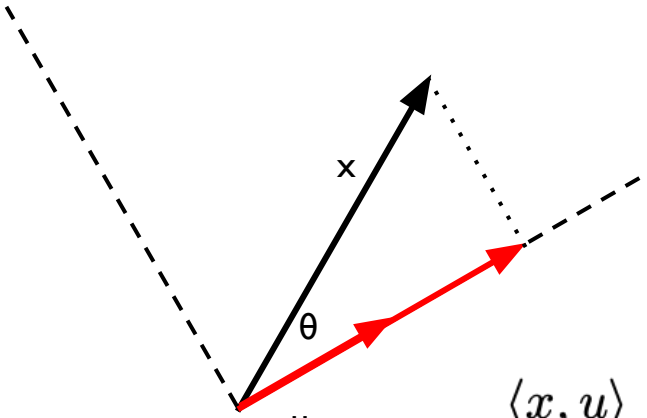
Definição. Equação (1) é chamada decomposição ortogonal de x com relação à W; vetor x_W é a projeção ortogonal de x em W.

Projeção em uma linha

Hora do quiz

$$\|x_u\|_2 = \|x\|_2 \cos\theta$$
$$= \|x\|_2 \frac{\langle x, u\rangle}{\|x\|_2\|u\|_2}$$
$$= \frac{\langle x, u\rangle}{\|u\|_2}$$

$$x_u = \|x_u\|_2 \frac{u}{\|u\|_2}$$
$$= \frac{\langle x, u\rangle}{\|u\|_2^2} u$$



$$x_u = \frac{\langle x, u\rangle}{\|u\|_2^2} u$$

Caso especial: quando u é unitário, temos $x_u = \langle x,u\rangle u$

Projeção em plano ou hiperplano

Seja W o espaço gerado por $\{v_1,...,v_m\}$ e A a matriz cujas colunas são $\{v_1,...,v_m\}$.
Para calcular a decomposição ortogonal de x em relação a W:

1. Calcule a matriz A^TA e o vetor $b=A^Tx$
2. Resolva $A^TAc = b$ e encontre o vetor desconhecido c.
3. O sistema é sempre consistente, escolha uma solução c:

$x_W = Ac$ $\qquad$ $x_{W\perp} = x-x_W$

$$x_W = A(A^\top A)^{-1}A^T x$$

Mudança da base canônica para base B

Seja $B=\{v_1,...,v_n\}$ uma base de R^n. As coordenadas do vetor x em relação à B, $a = [a_1,...,a_n]^T = [x]_B$ são obtidas a partir de

$a_1v_1 + a_2v_2 + \ldots + a_nv_n = x$ ou seja $[v_1\ v_2\ \ldots\ v_n][x]_B = [x]$

Seja P a matriz cujas colunas são os vetores da base

$$P = [v_1\ v_2\ \ldots\ v_n]$$

Então para qualquer vetor x em temos

$$P[x]_B = x \quad \text{e} \quad [x]_B = P^{-1}x$$

O que acontece se P_{nxn} for ortonormal?

Fonte: https://imada.sdu.dk/~marco/DM559/Slides/dm559-lec8.pdf

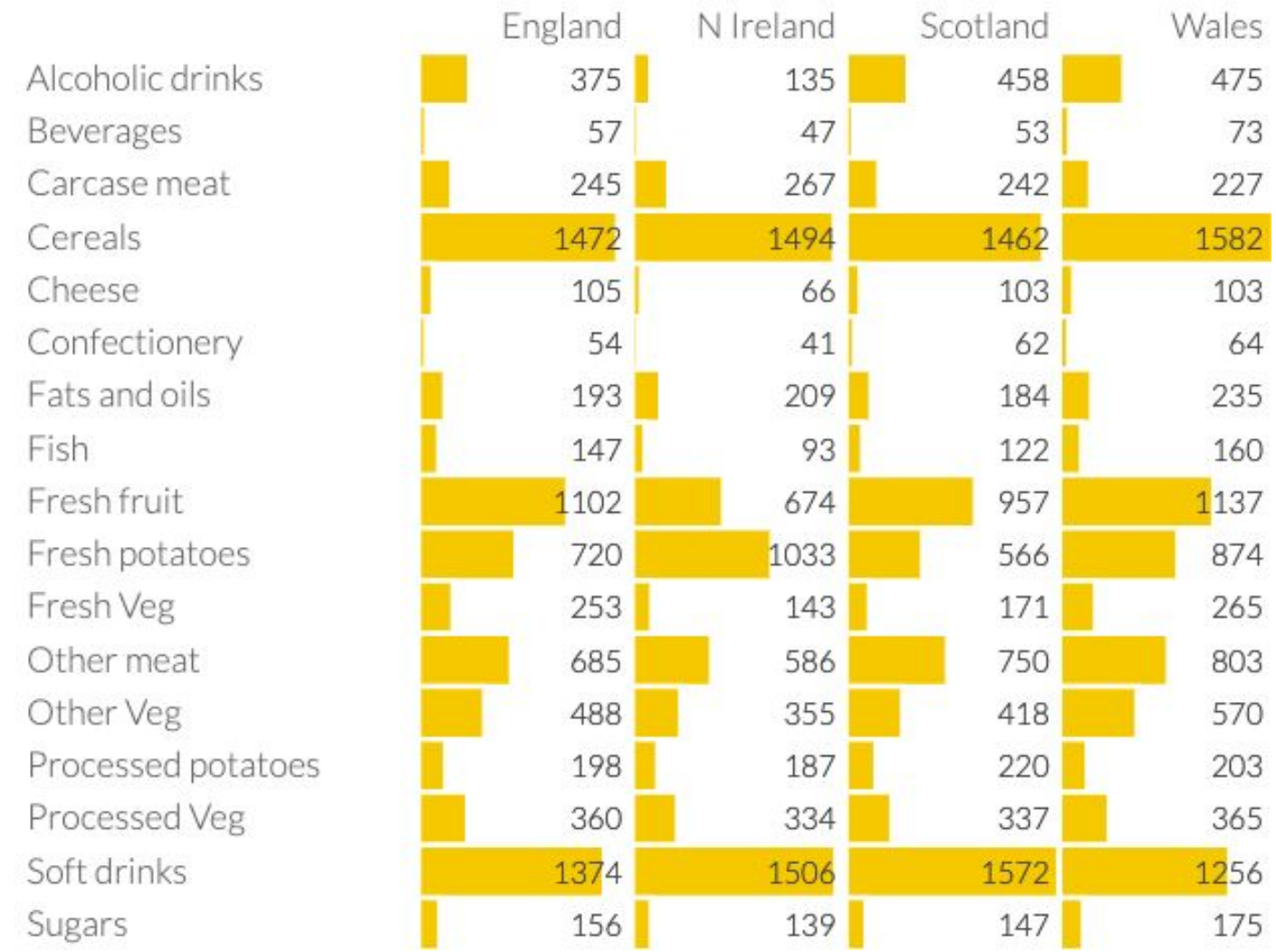
Exemplo de análise de dados multivariada

DEFRA[1] 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

[1]Department for Environment, Food and Rural Affairs

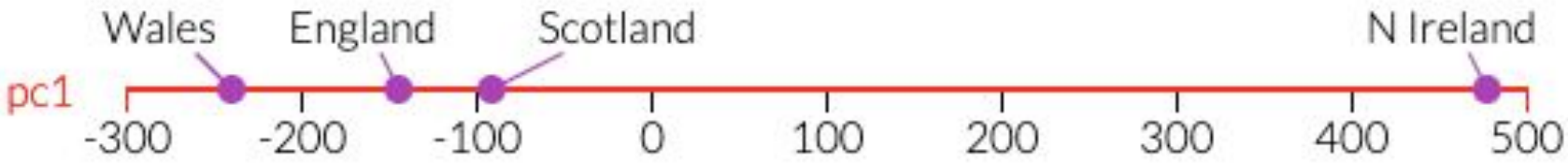


Como analisar estes dados? Existem correlações entre os países? Plotar todos de uma vez? Plotar variáveis 2 a 2?

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 (PC1)**, direção que maximiza variância através das 17 coords.

Após encontrar PC1, projetamos os dados neste novo eixo. Como?

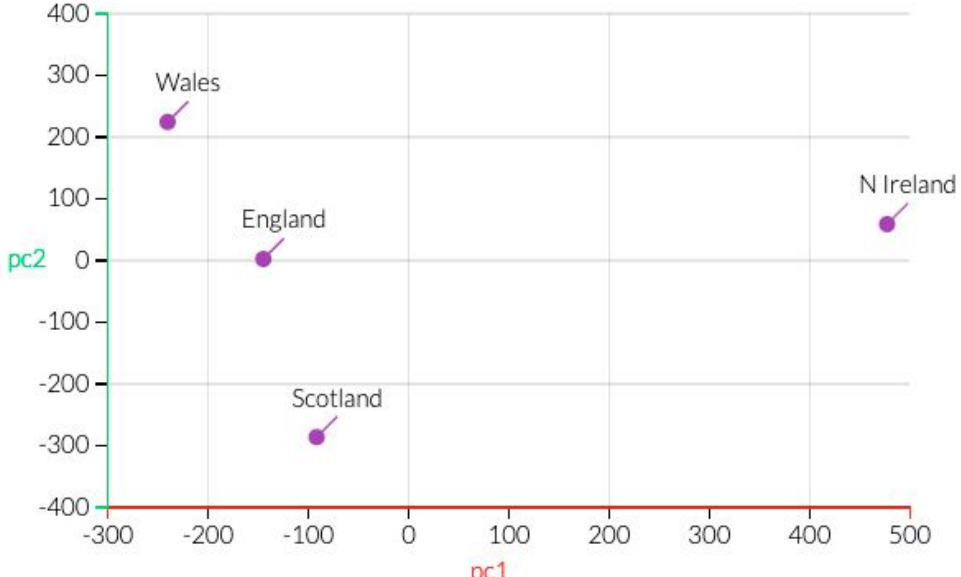


Score plot

\+ Fresh potatoes
\- Fresh fruits
\- Alcoholic drinks

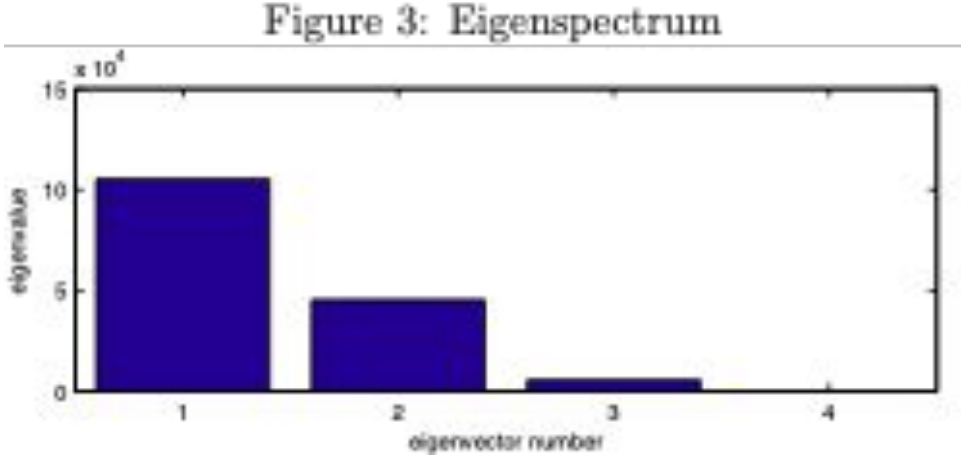
PCA: segunda componente principal

- 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

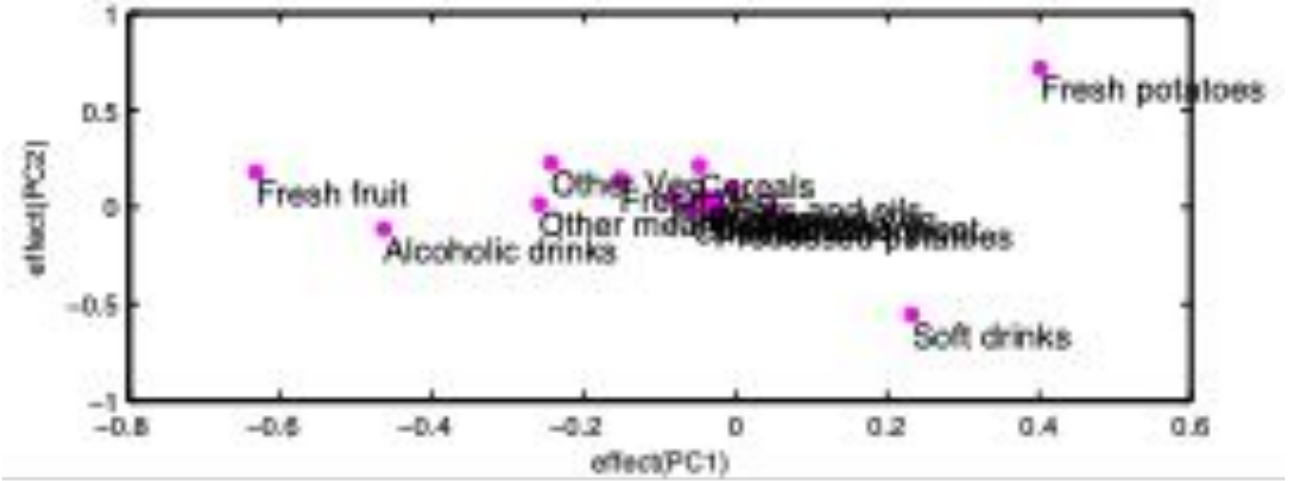
PCA também retorna info sobre contribuição de cada PC para a variância total dos dados: 67% por PC1, 97% por PC1+PC2



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 x n onde colunas = observações, linhas = vars
Queremos transformar X em matriz Y_{mxn} usando P_{mxm}:

$$Y = PX$$

(Mudança de base)

Mudança de base Y=PX

Sejam p_1, p_2, …, p_m os vetores linha de P e x_1, …, xn os vetores coluna de X. Temos

$$PX = \begin{pmatrix} Px_1 & Px_2 & \dots & Px_n \end{pmatrix} = \begin{pmatrix} p_1.x_1 & p_1.x_2 & \cdots & p_1.x_n \\ p_2.x_1 & p_2.x_2 & \cdots & p_2.x_n \\ \vdots & \vdots & \ddots & \vdots \\ p_m.x_1 & p_m.x_2 & \cdots & p_m.x_n \end{pmatrix} = Y$$

As colunas de X estão sendo projetadas nas linhas de P. Logo, as linhas de P {p_1, p_2, …, 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 mu. Amostras de Z: z={z1,z2,...,zn}.
Variância amostral de Z:

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

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

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

Otimização das PCs: maximizar variância

Seja s = (s_1,...,s_n) um segundo vetor de medições com média zero, podemos generalizar a idéia anterior para covariância de r e s.

Interpretação: quanto duas variáveis mudam simultaneamente.

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

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

Otimização das PCs: maximizar variância

Generalizando para matrix X_{mxn} de dados (m vars, n observações)

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_m} \end{pmatrix} \in \mathbb{R}^{m \times n}, \qquad \mathbf{x_i}^T \in \mathbb{R}^n$$

x_i: vetor de n samples da i-ésima variável

$$\mathbf{C_X} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T = \frac{1}{n-1} \begin{pmatrix} \mathbf{x_1}\mathbf{x_1}^T & \mathbf{x_1}\mathbf{x_2}^T & \cdots & \mathbf{x_1}\mathbf{x_m}^T \\ \mathbf{x_2}\mathbf{x_1}^T & \mathbf{x_2}\mathbf{x_2}^T & \cdots & \mathbf{x_2}\mathbf{x_m}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_m}\mathbf{x_1}^T & \mathbf{x_m}\mathbf{x_2}^T & \cdots & \mathbf{x_m}\mathbf{x_m}^T \end{pmatrix} \in \mathbb{R}^{m \times m}$$

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: $p_1, p_2, \ldots, p_m$ são ortonormais.

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

$$\mathbf{C_Y} = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T = \frac{1}{n-1}(\mathbf{PX})(\mathbf{PX})^T = \frac{1}{n-1}(\mathbf{PX})(\mathbf{X}^T\mathbf{P}^T) = \frac{1}{n-1}\mathbf{P}(\mathbf{X}\mathbf{X}^T)\mathbf{P}^T$$

$$\text{i.e.} \quad \mathbf{C_Y} = \frac{1}{n-1}\mathbf{P}\mathbf{S}\mathbf{P}^T \quad \text{where} \quad \mathbf{S} = \mathbf{X}\mathbf{X}^T$$

S_{mxm} é simétrica

Fato: toda matriz simétrica S é ortonormalmente diagonálizável, 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 covariâcia

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

E_{mxm} é ortonormal

Logo,

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

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

Passo-a-passo

Equivale a SVD de **X**?

1. Calcular autovalores e autovetores de $\mathbf{S}=\mathbf{X}\mathbf{X}^T$.
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.

Recapitulando

Vimos que a multiplicação

$$Y_{mxn} = P^T_{mxm} X_{mxn}$$

é uma mudança para base formada pelas colunas de P. A coluna y_i é a representação da coluna x_i na nova base. Quando P é ortonormal, esta mudança de base é uma **rotação**.

Note que a multiplicação pela matriz diagonal D_{mxm}

$$y_{mx1} = D_{mxm}\, x_{mx1}$$

irá "escalar" as coordenadas de x conforme os elementos d_{ii}.

Interpretação Geométrica do SVD

Qual o máximo de ||y||? Quando isso acontece?

Multiplicar $\mathbf{M}_{m\times n}$ é uma transformação linear $\mathbb{R}^n \to \mathbb{R}^m$:

$$\mathbf{y}_{m\times 1} = \mathbf{M}_{m\times n}\mathbf{x}_{n\times 1}$$

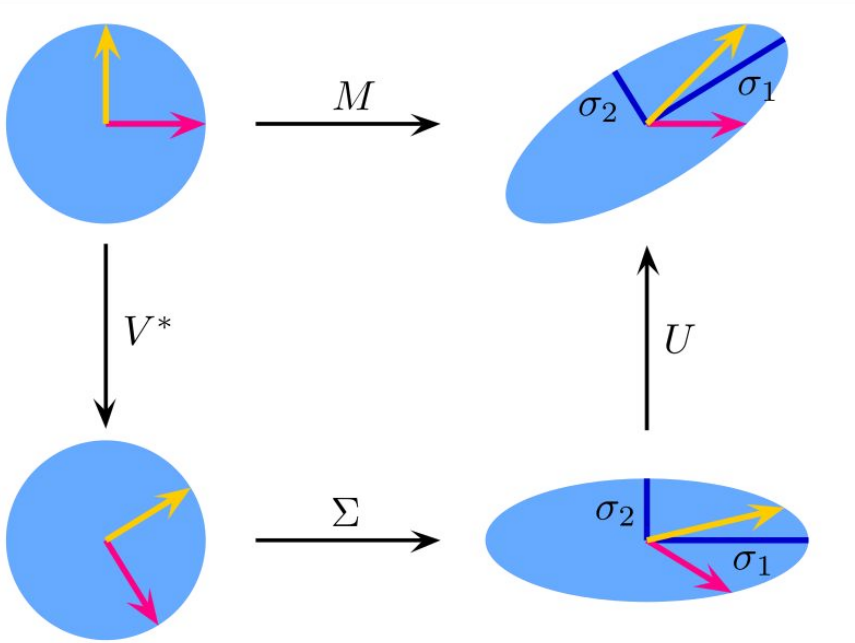
Usando o SVD $M = U\Sigma V^\top$, temos

$$\mathbf{y} = \mathbf{U\Sigma V}^\top\mathbf{x}$$

$$\mathbf{y} = \mathbf{U\Sigma} \times [\mathbf{x}]_\mathbf{V}$$

$$\mathbf{U}^\top\mathbf{y} = \mathbf{\Sigma} \times [\mathbf{x}]_\mathbf{V}$$

$$[\mathbf{y}]_\mathbf{U} = \mathbf{\Sigma} \times [\mathbf{x}]_\mathbf{V}$$



$$M = U \cdot \Sigma \cdot V^*$$