# Modelagem Matemático-computacional
# Aula 5

Apresentaremos nesta aula um método mutivariado para reduzir a dimensão de sistemas com muitas dimensões, conhecido como Análise por Componentes Principais (PCA).

## I. MOTIVAÇÃO

Em diversas pesquisas é comum coletar-se uma quantidade imensa de dados do sistema que está sendo estudado, o que inclui uma grande variedade de medidas. Pense no tradicional problema de reconhecimento facial: é preciso coletar informações sobre a posição relativa, o tamanho e a proporção dos olhos, nariz, boca, além de mensurar a luminosidade em diversas posições da face. Comparar faces é um exemplo típico de *reconhecimento de padrões*, ou seja, a partir de uma população de indivíduos (ou objetos), pretende-se econtrar semelhanças entre os mesmos com base em medidas deles extraídas.

Quando o volume de dados é muito grande, a dificuldade de analisar o sistema como um todo torna-se complicada não apenas computacionalmente, mas também dificulta o trabalho do pesquisador de encontrar ferramentas que possam auxiliá-lo nesta análise e na *visualização*. Idealmente, seria conveniente produzir resultados passíveis de visualização 2 e 3D: é muito mais fácil analisar um gráfico do que uma tabela cheia de números!

Vimos nas aulas passadas que as medidas de *correlação* e *covariância* são uma boa forma de estimar a dependência entre duas variáveis aleatórias. Podemos considerar as medidas extraídas de um determinado estudo como variáveis aleatórias. Quando pares de medidas estão correlacionados, uma pode ser obtida a partir da outra, de modo que não se perde informação ao escolher apenas uma delas para a análise do sistema, pois há redundância.

| ID | Altura | Peso | Idade | ID | Altura | Peso | Idade |
|---|---|---|---|---|---|---|---|
| 1 | 1.53 | 48 | 23 | 14 | 1.88 | 82 | 33 |
| 2 | 1.44 | 41 | 29 | 15 | 1.62 | 65 | 31 |
| 3 | 1.72 | 70 | 25 | 16 | 1.96 | 90 | 36 |
| 4 | 1.55 | 42 | 37 | 17 | 1.67 | 68 | 24 |
| 5 | 1.69 | 67 | 22 | 18 | 1.92 | 85 | 35 |
| 6 | 2.00 | 98 | 37 | 19 | 1.46 | 46 | 27 |
| 7 | 1.50 | 41 | 32 | 20 | 2.11 | 122 | 29 |
| 8 | 1.55 | 47 | 38 | 21 | 1.63 | 43 | 24 |
| 9 | 2.11 | 114 | 29 | 22 | 1.73 | 74 | 27 |
| 10 | 1.58 | 48 | 41 | 23 | 1.84 | 78 | 36 |
| 11 | 1.59 | 47 | 25 | 24 | 2.06 | 108 | 29 |
| 12 | 1.53 | 43 | 39 | 25 | 2.07 | 102 | 27 |
| 13 | 1.90 | 86 | 25 | | | | |

Tabela I

Vamos exemplificar esta idéia com os dados de um conjunto de atletas de esportes diferentes. A tabela I apresenta as medidas de peso e altura e a idade de 25 atletas diferentes. Tente juntar grupos de atletas que pertençam a esportes diferentes, encontrando também quantos e quais são os esportes. Em princípio, somente com os dados da tabela esta é uma tarefa difícil (ou demorada).

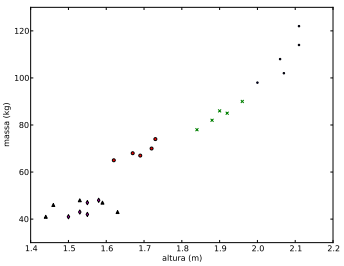
Agora veja os gráficos com os pares de medidas plotados na forma de correlograma (figura 1). Nesta figura cada atleta é mapeado num espaço bidimensional, sendo cada eixo respectivo a uma medida. O esporte que o atleta pratica é indicado pela cor e forma do ponto no gráfico. Círculos azuis são jogadores de basquete, cruzes verdes são jogadores de vôlei, círculos vermelhos são jogadores de futebol, triângulos pretos são ginastas e losangos roxos são jóqueis.

Note que na figura 1a em que temos o correlograma da altura vs. a massa dos atletas há uma tendência de variação conjunta na forma linear. O mesmo não é observado nas figuras 1b e 1c, onde estão os correlogramas de altura vs. idade e peso vs. idade, respectivamente. Em particular, no gráfico 1a é possível notar a formação de alguns grupos. Um aglomerado de pontos no canto esquerdo inferior revela que duas categorias de atletas têm pesos e alturas muito próximos: são jóqueis e ginastas. Observe que, de fato, a altura não tem dependência explícita com a idade, assim como o peso não está ligado à idade dos atletas adultos. Quando a mesma comparação é feita com crianças, este perfil de correlação seria alterado, já que a medida que na idade infantil a criança está em fase de crescimento.

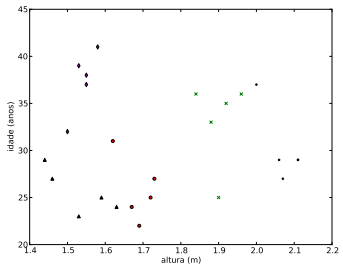
Este exemplo clarifica a idéia de que a covariância entre duas medidas (variáveis aleatórias) pode ser aproveitada para reduzir o número de medidas a serem consideradas na análise de um problema. Em seguida, vamos introduzir os passos da transformação que gera uma redução de dimensões do sistema, isto é, o PCA.

## II. PRINCIPAL COMPONENT ANALYSIS (PCA)

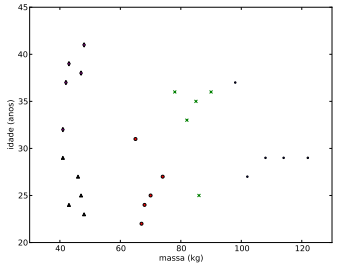
Análise por Componente Principais (do inglês *Principal Component Analysis*), é uma transformação linear multivariada que implementa uma rotação no sistema de eixos originais do sistema. Tais eixos podem ser entendidos como as medidas do sistema. Por exemplo quando se mede a posição de uma partícula no espaço dimensional, associa-se um ponto as suas coordenadas $(x_p, y_p, z_p)$. Para uma partícula livre, cada um dos $x_p$,




a)




b)




c)

Figura 1: Correlogramas entre a) altura e peso, b) altura e idade, c) peso e idade.

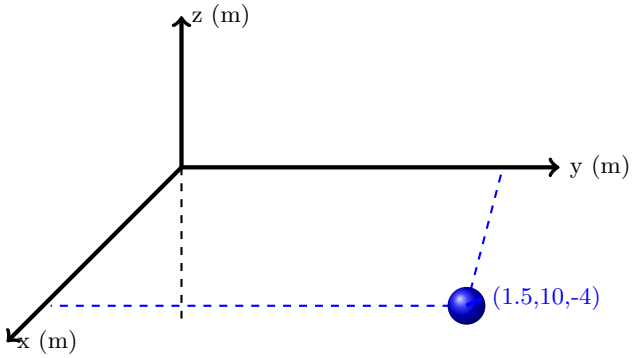



Figura 2: Posição de uma partícula livre no espaço 3D.

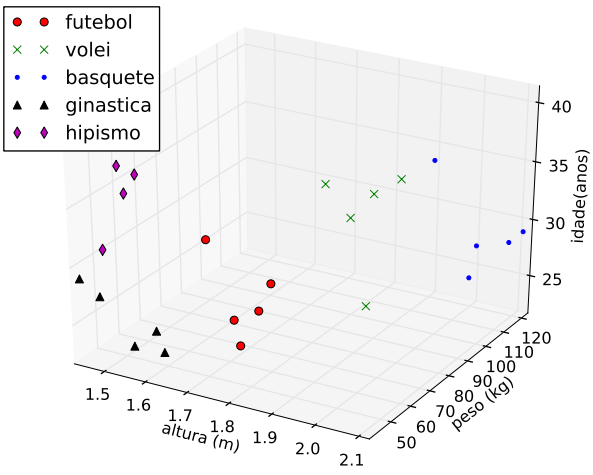



Figura 3: Espaço 3D de medidas dos atletas.

$y_p$ e $z_p$ pode assumir valores reais entre $(-\infty, +\infty)$. Uma posição válida seria $(x_p = 1.5m, y_p = 10m, z_p = -4m)$ (figura 2). No caso de um número $M$ de medidas mais gerais (não necessariamente de posição), associamos a cada eixo uma medida, de modo que cada indivíduo é mapeado no espaço M-dimensional como um ponto. No exemplo anterior, com as medidas de altura, peso e idade de atletas, poderíamos fazer as associações $x \rightarrow altura$, $y \rightarrow peso$ e $z \rightarrow idade$ (figura 3. Para $M > 3$ é impossível que este mapeamento seja visualizado.

O objetivo do PCA é obter um novo sistema de coordenadas com $P < M$ dimensões, de modo que a dispersão dos dados seja máximizada ao longo dos $P$ eixos, ao mesmo tempo em que estas $P$ dimensões permitam representar o sistema sem perda considerável de informação do sistema original. Para duas variáveis correlacionadas de forma linear com eixos $x, y$, a mudança de eixos acarreta na escolha de uma única dimensão com uma dispersão associada. Na figura 4 está ilustrado como a mudança

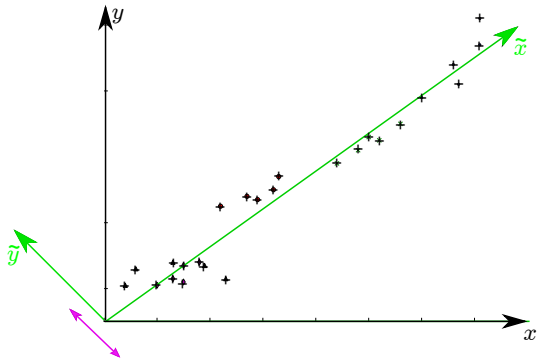



Figura 4: Exemplo de mudança dos eixos através do PCA, de modo que o primeiro eixo contém praticamente toda a variância do sistema.

leva a um novo sistema $\tilde{x}, \tilde{y}$.

Um fato interessante é que alguns sistemas concentram aproximadamente 99.5% da variação nos dois primeiros eixos dos PCA, como por exemplo em áudio e imagens. Logo, o PCA está associado à *compactação* (compressão) dos dados.

Idealmente, o valor de $P$ deve ser escolhido de modo que a nova projeção represente 75% da variância dos dados.

O algoritmo para implementar o PCA deve seguir um conjunto de passos que será explicado logo abaixo. Considere que o sistema sobre o qual o PCA será aplicado contém $N$ observações (indivíduos) de $M$ variáveis aleatórias (medidas). Matricialmente isto equivale a dizer que temos um conjunto de dados $\vec{X}$, no qual cada componente é uma variável aleatória $X_i$ com $N$ realizações.

### Passos

1. Obter a matriz de covariância $K$ entre os pares de medidas, ou seja:

$$K = \begin{bmatrix} cov(X_1, X_1) & cov(X_1, X_2) & \dots & cov(X_1, X_M) \\ cov(X_2, X_1) & cov(X_2, X_2) & \dots & cov(X_2, X_M) \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_M, X_1) & cov(X_M, X_2) & \dots & cov(X_M, X_M) \end{bmatrix}$$

   onde cada elemento $cov(X_i, X_j)$ é a covariância calculada entre as duas medidas $X_i$ e $X_j$. Observe que $K$ é quadrada, real e simétrica.

2. Obter os autovalores $\vec{\lambda}$ e os autovetores associados $V$ da matriz $K$, ordenados de modo decrescente, isto é, ordenar os autovalores e os respectivos autovetores, de modo que cada linha da matriz $V$ contenha um autovetor, sendo que a primeira linha contém o autovetor associado ao maior autovalor.

3. Aplicar a transformação linear:
   $\tilde{\vec{X}} = V\vec{X}$,

   conhecida como transformação de *Karhunen-Loève*.

   Observe que o PCA é aplicado sobre esta transformação, desprezando-se os eixos com menor dispersão, ou seja, escolhendo-se de $P < M$ componentes para reduzir a dimensão do sistema. Os autovetores associados aos maiores autovalores são as direções com mais informação. Ou seja, são as direções menos correlacionadas, também chamadas de *componentes principais* que agem como pesos, permitindo selecionar automaticamente as medidas que são mais importantes para a representação do sistema.

No exemplo dos atletas, o PCA com apenas duas componentes está ilustrado na figura 5.

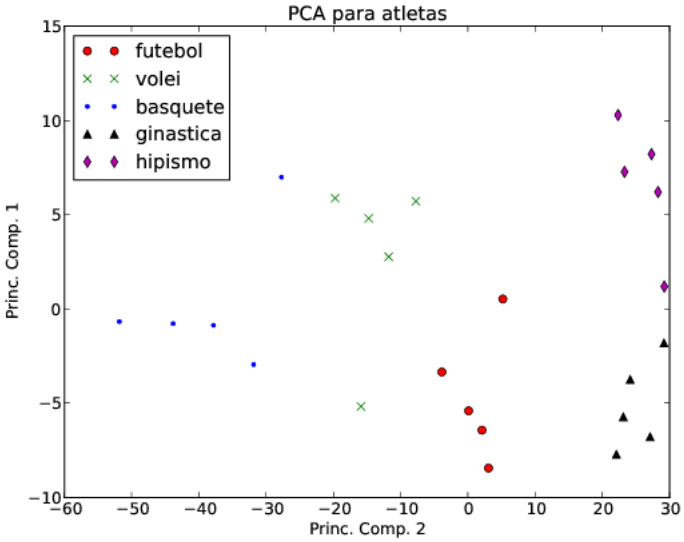



Figura 5: PCA aplicado sobre os dados dos atletas. Duas componentes representaram mais de 99.9% da variância total dos dados.

### Referências

[1] COSTA, L. da F., CESAR Jr, R.M. Shape analysis and classification: theory and pratice, Boca Raton, FL, CRC Press; 2001.

[2] SMITH, L. I., A tutorial on Principal Component Analysis.