Podemos ver em R^n ?

Carlos Eduardo Pedreira

PESC - COPPE

2019

¿Donde están los datos? En R^n

Tenemos entonces 2 posibilidades:

- Ir al R^n y clasificar ahí
- Traer los datos para el R^2

Ventajas de traer a 2-D:

- Es el usuario el que toma la decisión, no el 'sistema'.
- Se puede añadir información especifica
- Los resultados parecen mas 'convincentes' cuando uno puede 'verlos'

Ventajas de decidir en n dimensiones:

- No tener de traer a 2-D
- Hay siempre perdidas para traer de n-D a 2-D. El espacio em 2-D es muy muy muy pequeño para la información que tenemos en n-D

Existen muchas posibilidades para traer los dados de n-D a 2-D:

- PCA -Principal Component Analisys
- MDS - Multidimensional Scaling
- t-SNE Stochastic Neighborhood Environment
- NCA -Neighbourhood Component Analisys
- Otros

Porque (e quando) queremos ‘ver’ em R^n?

Porque:

Frequentemente, é interessante ter uma ferramenta de suporte a decisão para auxiliar na tarefa de classificação. **Busca-se que a decisão final seja tomada pelo usuário e não pelo ‘sistema’**.

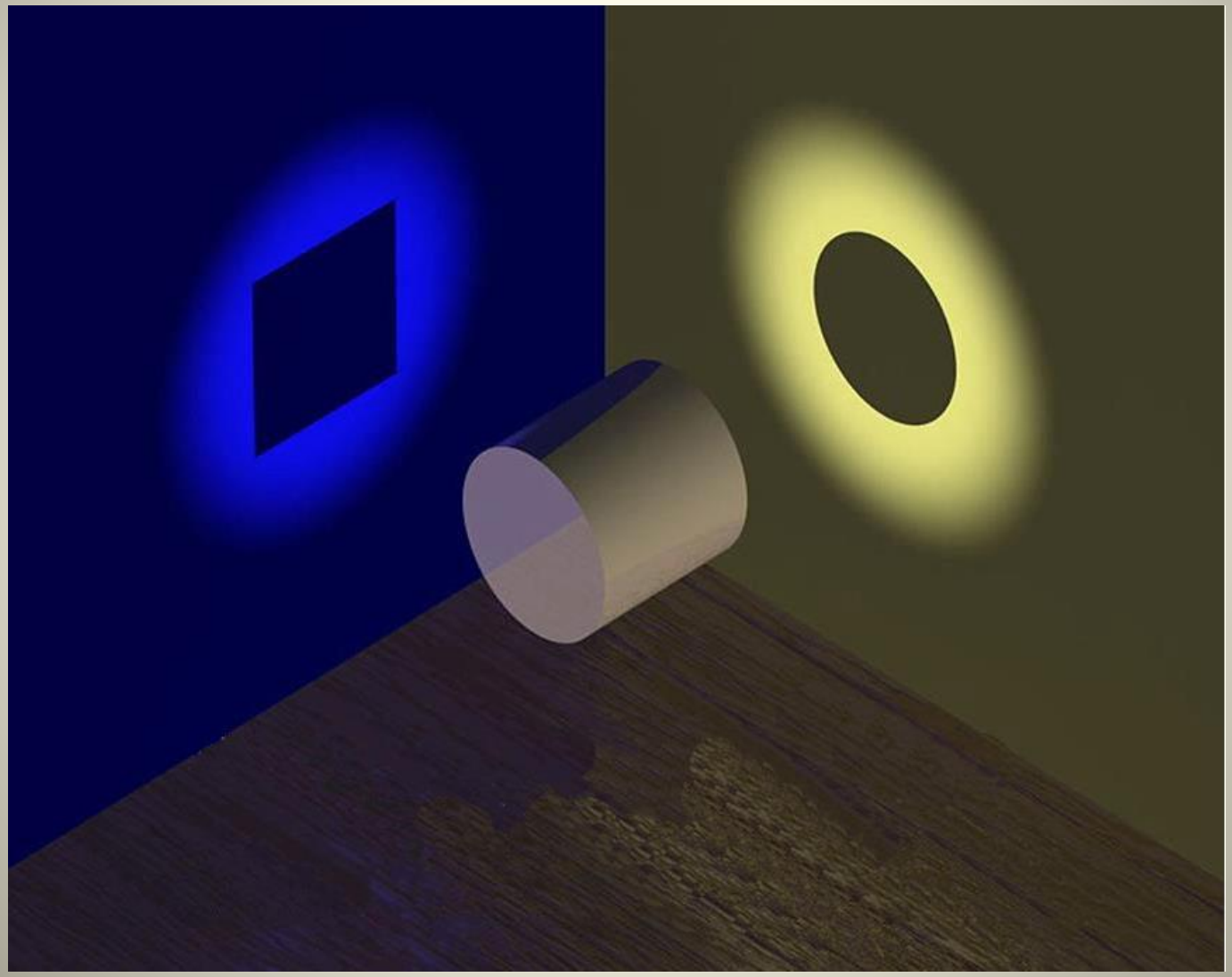
Quando:

- Não se quer classificar automaticamente por **razões éticas ou legais** e.g. diagnósticos médicos.
- Existe **informação adicional** difícil de ser modelada mas relevante de ser incluída.

Directamente en n-D (sin traer a 2-D)

- NCA - Neighbourhood Component Analisys
- SVM - Support Vector Machine
- Redes Neuronales
- Métodos Locales-Globales

Como se projeta = Como se vê



Escolhendo critérios para projetar os dados

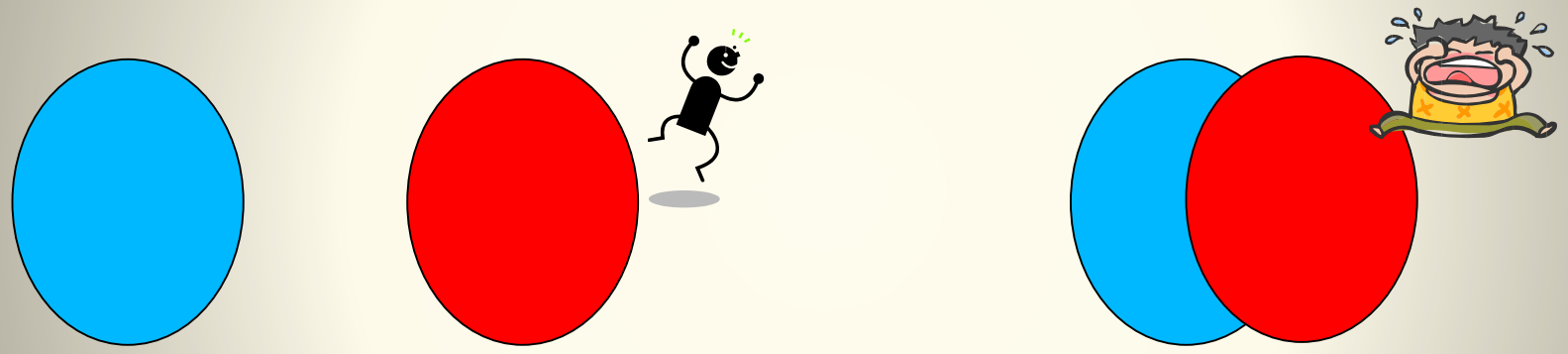
- Minimizar o *erro médio quadrático de reconstrução*.
- Buscar preservar a topologia ou a estrutura de distância no espaço projetado $\Re^2$.
- Produzir agrupamentos concentrados e bem separados no espaço projetado.

Bom para classificação!

Classificação de padrões

Queremos agrupamentos que sejam:

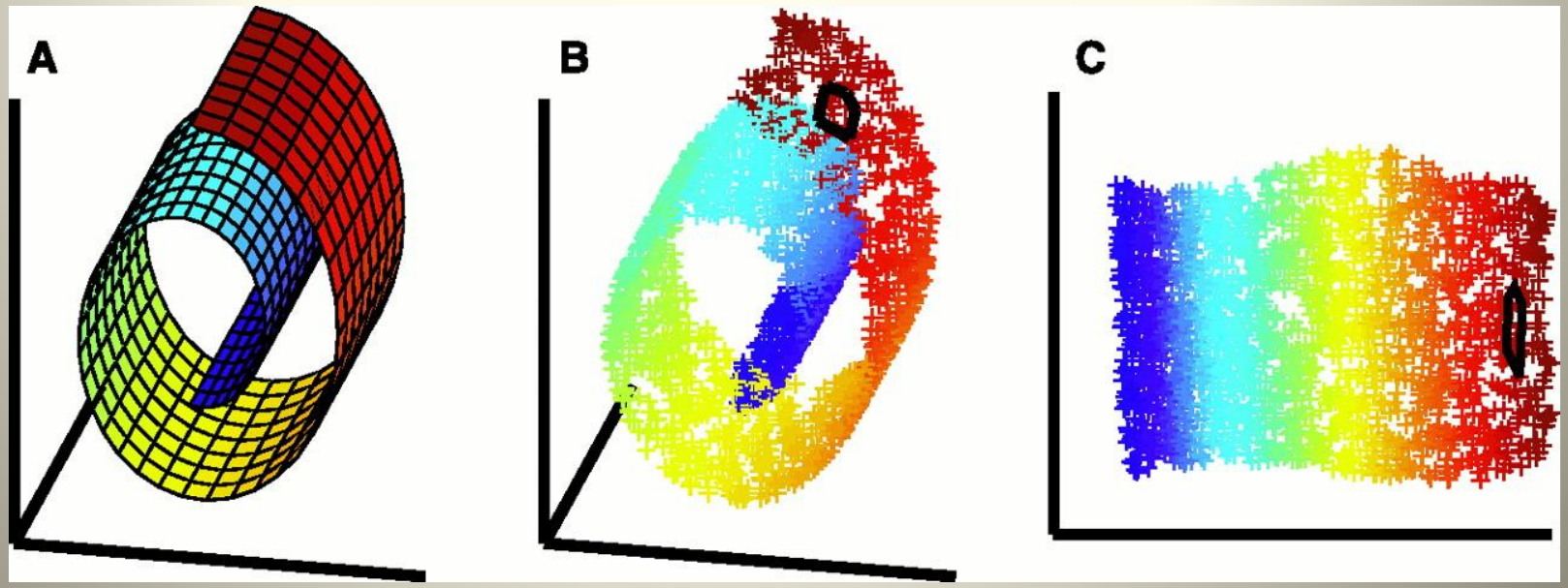
1) O mais separados possível



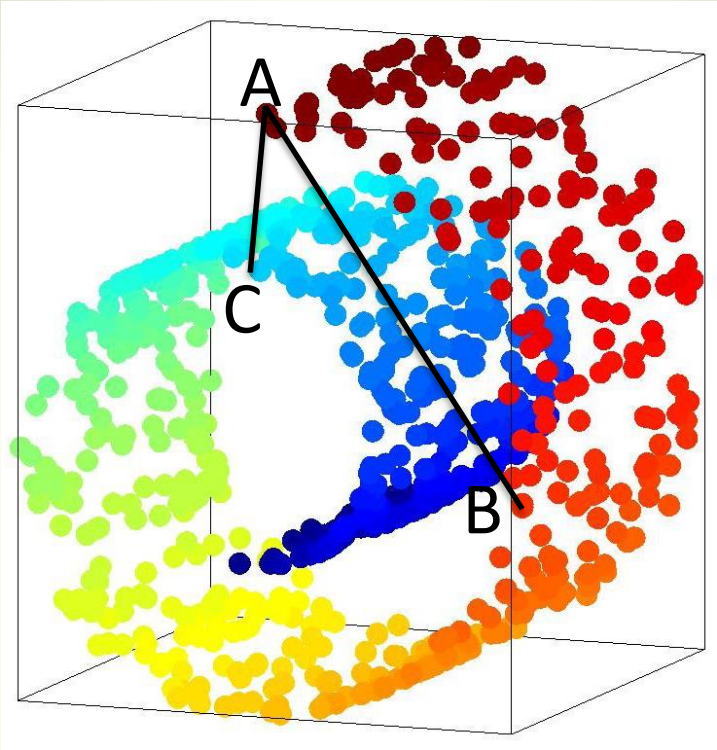
2) O mais concentrados possível

Manifold Learning: Desenrolando o Rocambole

A ideia central é **revelar uma ´dimensão intrínseca´** dos dados usando uma **métrica** baseada no menor caminho em um **grafo de vizinhos mais próximos**.



Se usarmos a distância Euclidiana, $D_{AC} < D_{AB}$



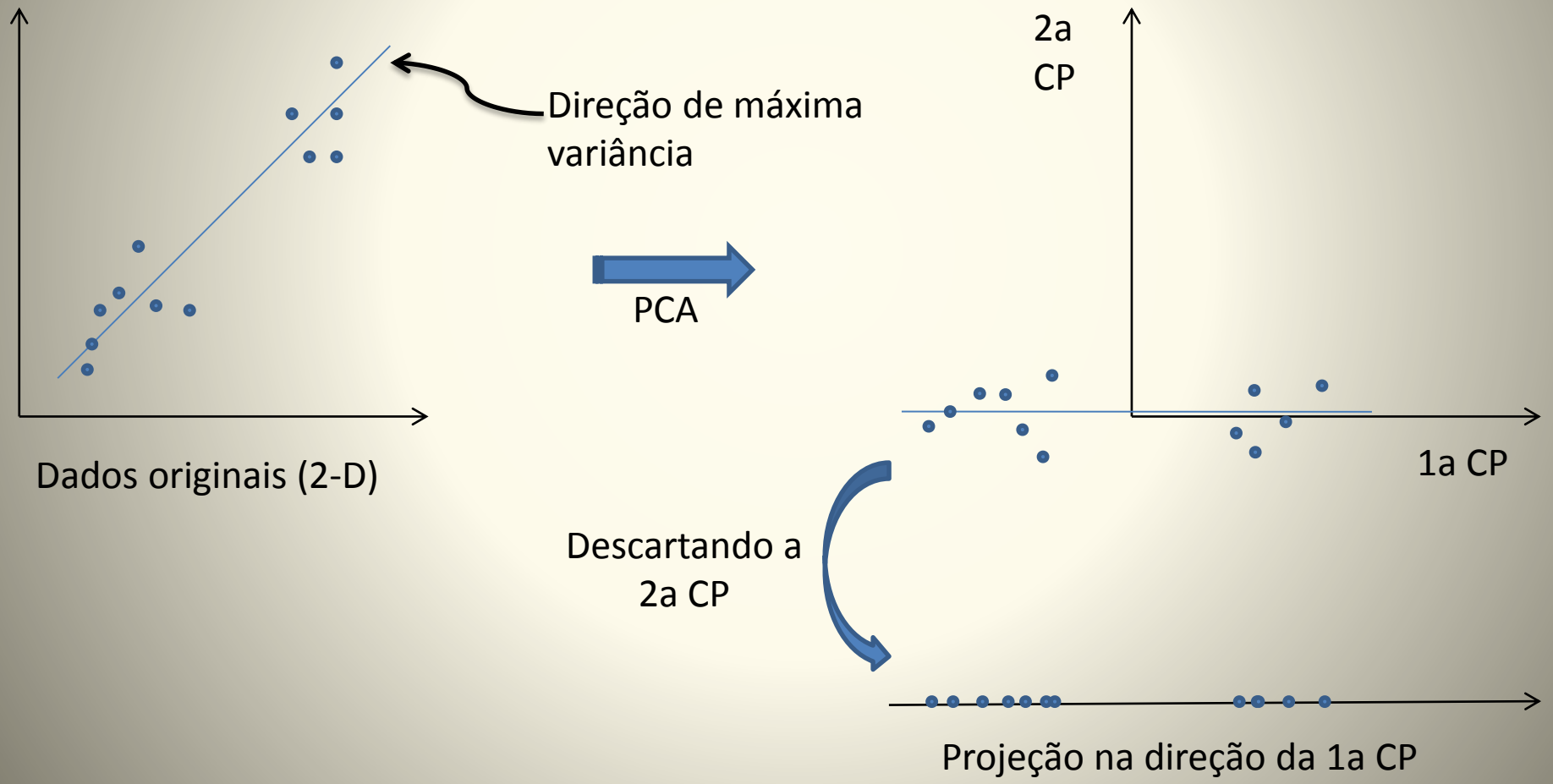
a estrutura real dos dados seria ocultada

PCA

Projeções nas componentes pricipais (transformada de Karhunen) **retêm o máximo da variação** presente nos dados no espaço original ($\Re^n$).

Como estamos interessados em '**visualização**', iremos direcionar a atenção à **primeira e segunda componentes**.

Vamos, por simplicidade, considerar uma projeção $\Re^2 \rightarrow \Re$ (normalmente estaríamos interessados em reduzir de $\Re^n \rightarrow \Re^2$)



Porque usar PCA ?
(dispersão como critério)

- Porque a solução do problema de otimização envolvido é bem conhecida. Existem alguns algoritmos bastante testados para esta finalidade.
- Porque funciona bastante bem em muitas situações.

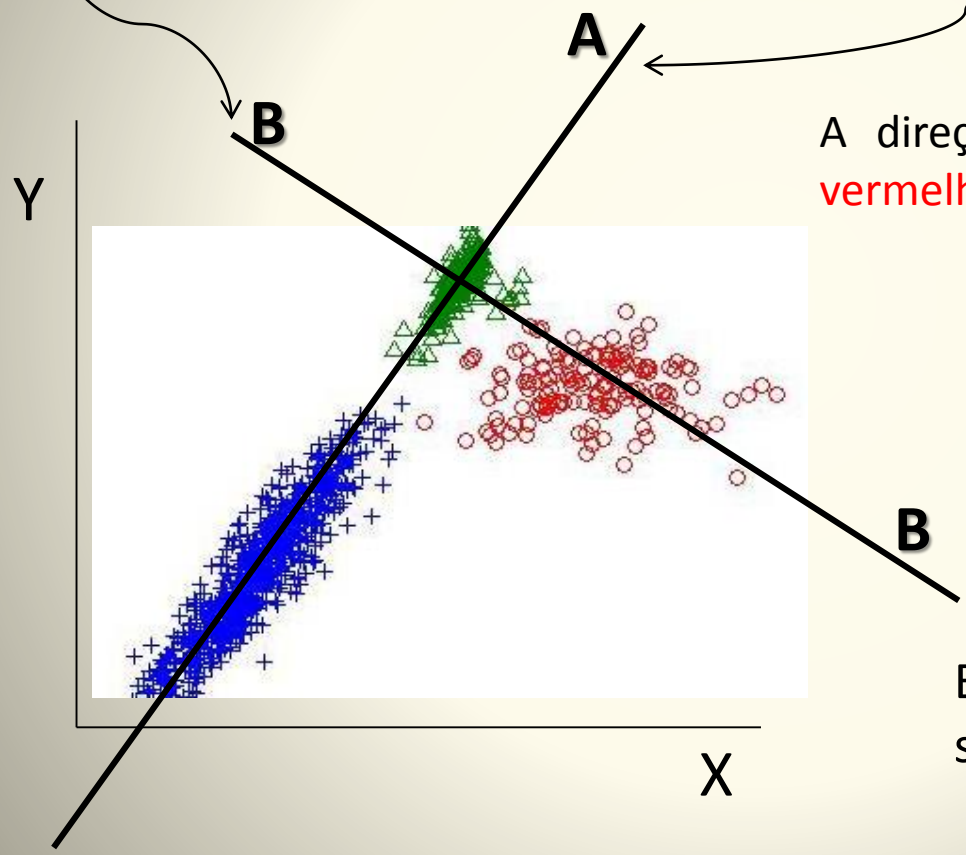
Mas não tão bem quanto gostaríamos ...

Porque?

Quando PCA vai mal para classificação

A direção **B** seria um desastre para agrupamentos azul e verde

Agrupamentos azul e verde se separam muito bem na direção **A**



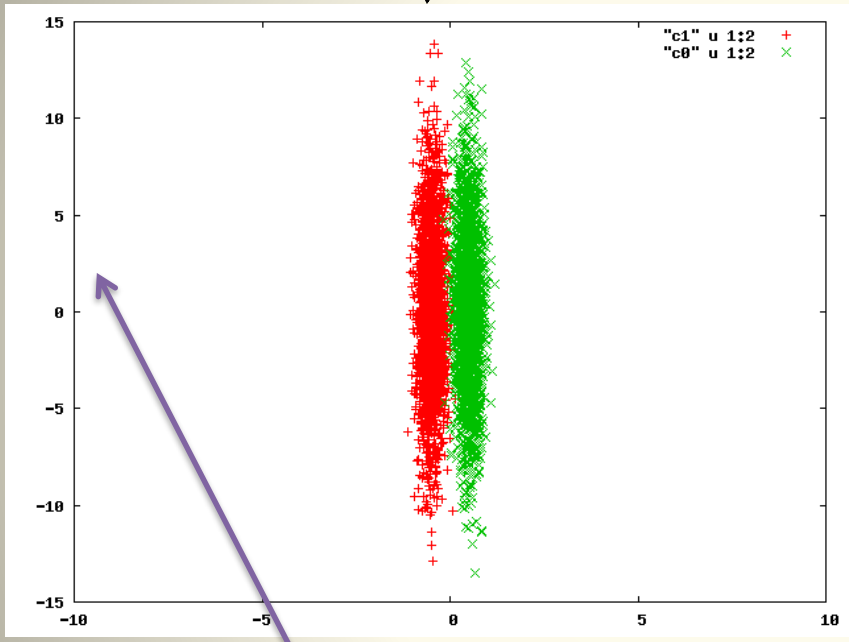
A direção **A** também é boa para azul e vermelho

Mas não tão boa para agrupamentos verde e vermelho

Estes seriam melhor separados na **B**

Quando PCA vai mal para classificação

Esta direção seria melhor



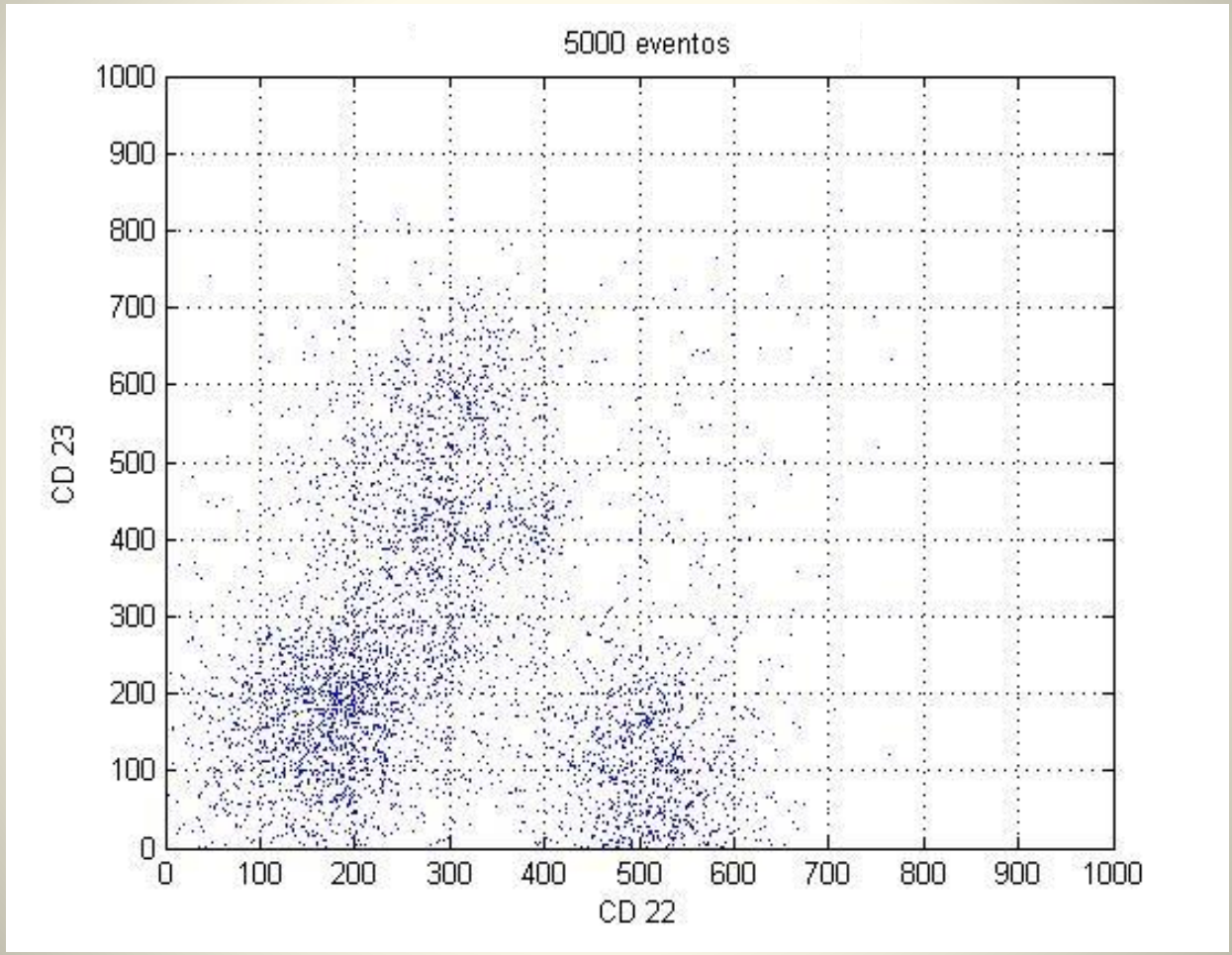
A direção de máxima variância não separa os dados de nenhuma maneira.

Neighbourhood Componet Analysis (NC A)

Criado por Geofrey Hinton (Premio Turing de 2019)

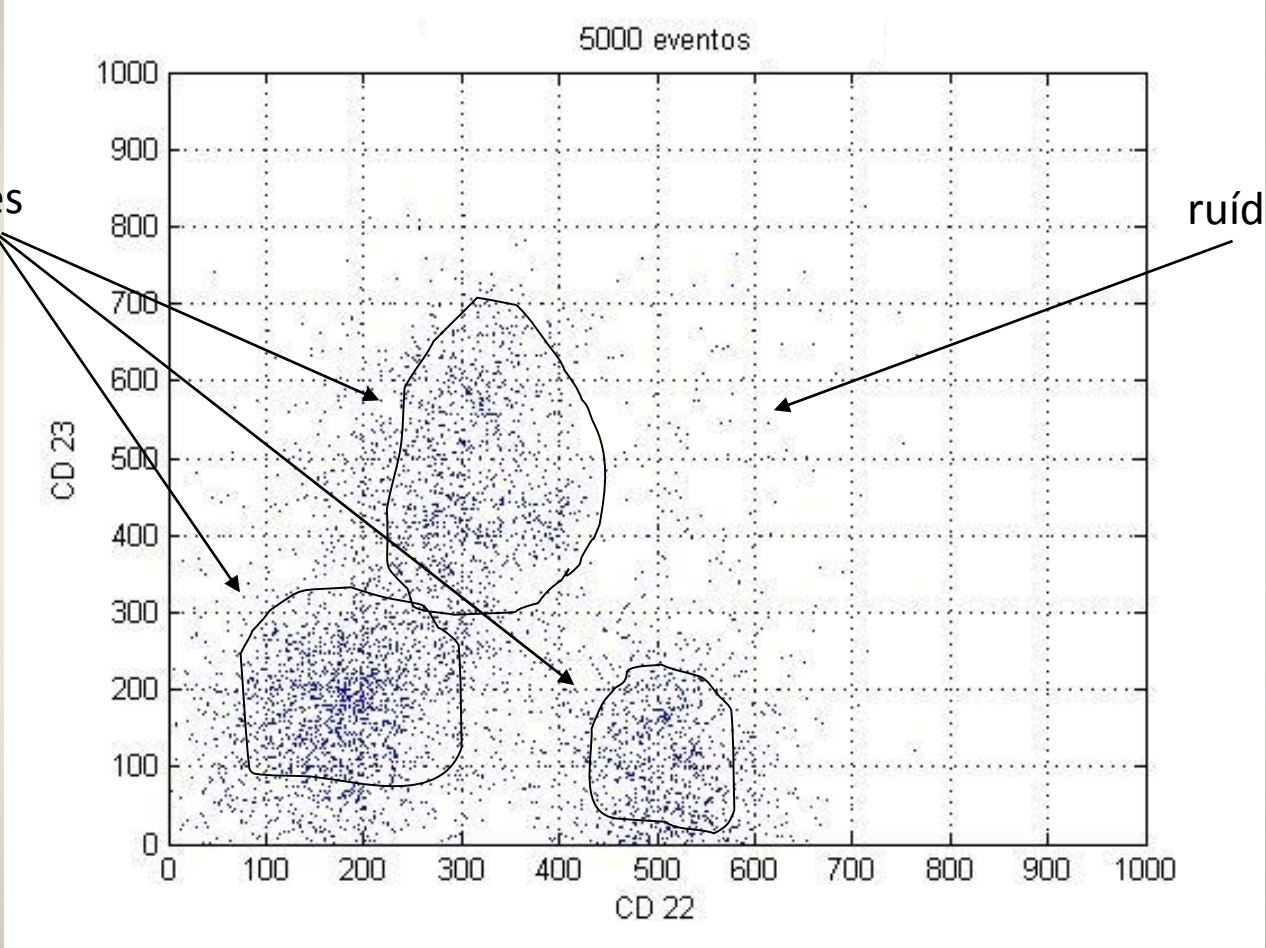
Clase correcta	Clase estimada	Probabilidades			
1.0000	1.0000	0.9801	0	0.0199	0
1.0000	1.0000	1.0000	0	0	0
1.0000	1.0000	1.0000	0	0	0
1.0000	1.0000	1.0000	0	0	0
4.0000	4.0000	0	1.0000	0	0
4.0000	4.0000	0	1.0000	0	0
4.0000	4.0000	0	1.0000	0	0
6.0000	6.0000	0	0	0	1.0000
6.0000	6.0000	0	0	0	1.0000
6.0000	6.0000	0	0	0	1.0000
6.0000	6.0000	0	0	0	1.0000
5.0000	5.0000	0	0	1.0000	0
5.0000	5.0000	0	0	1.0000	0
5.0000	5.0000	0	0	1.0000	0
5.0000	5.0000	0	0	0.9999	0.0001

Um problema de classificação

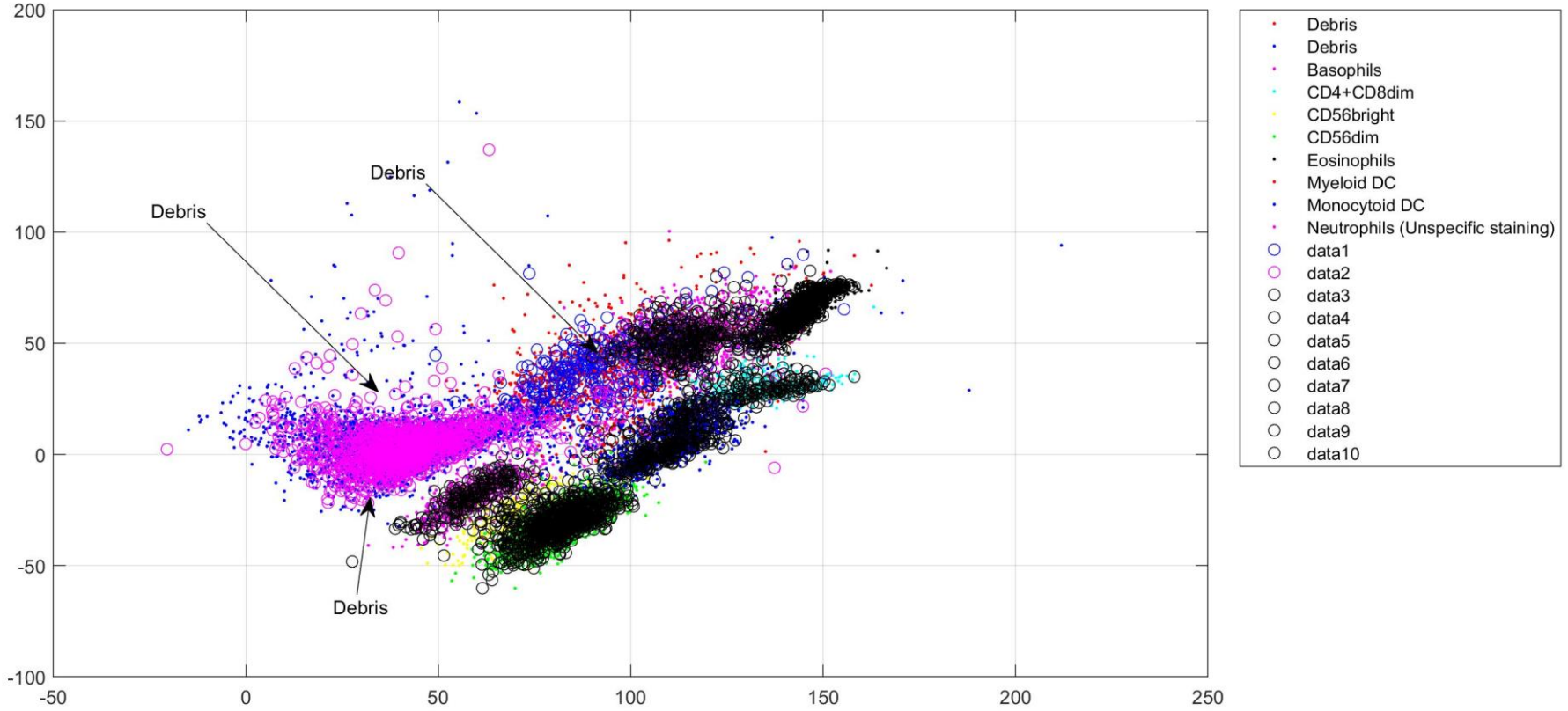


mas onde estão os grupos?

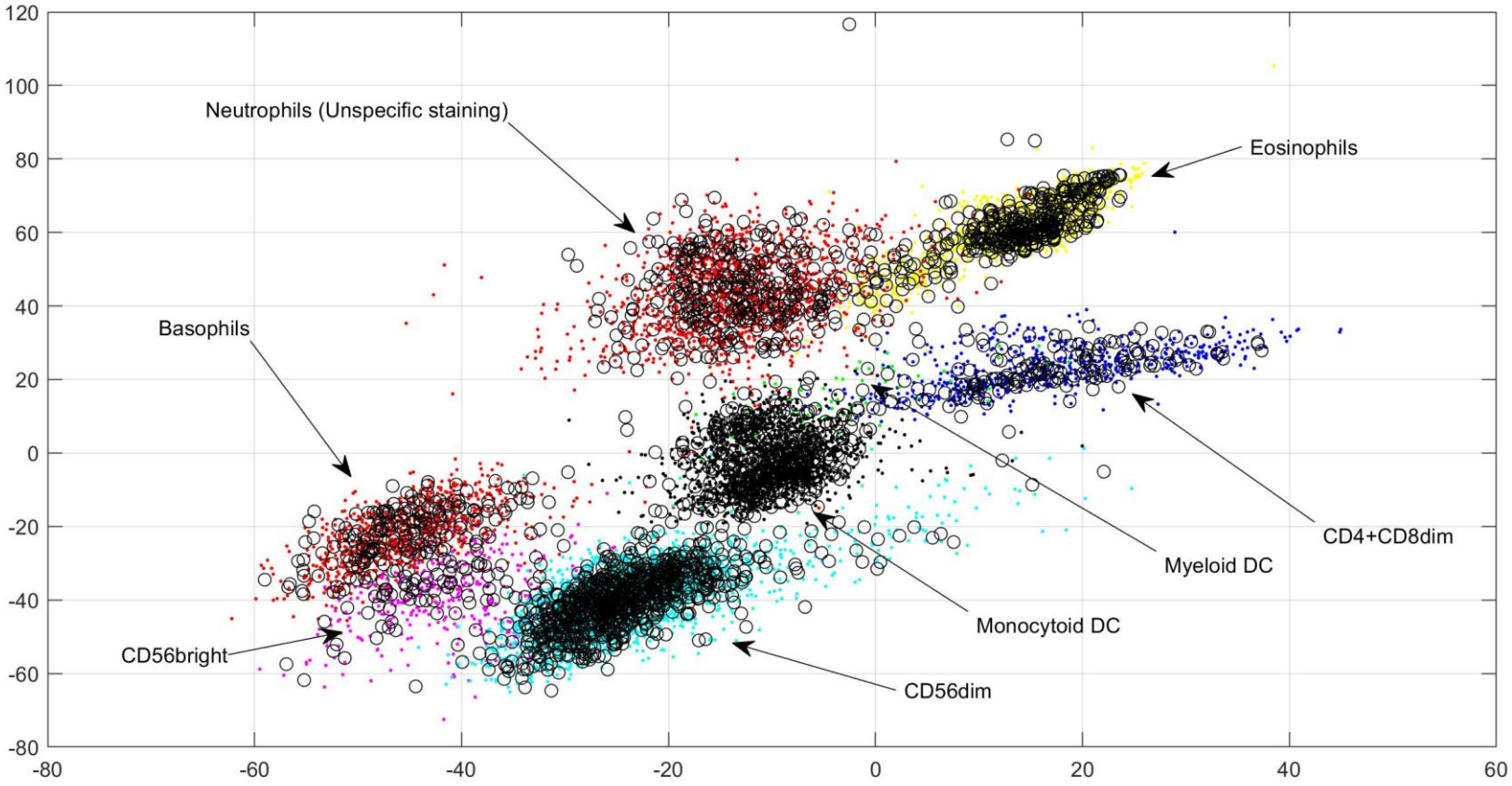
Populações celulares

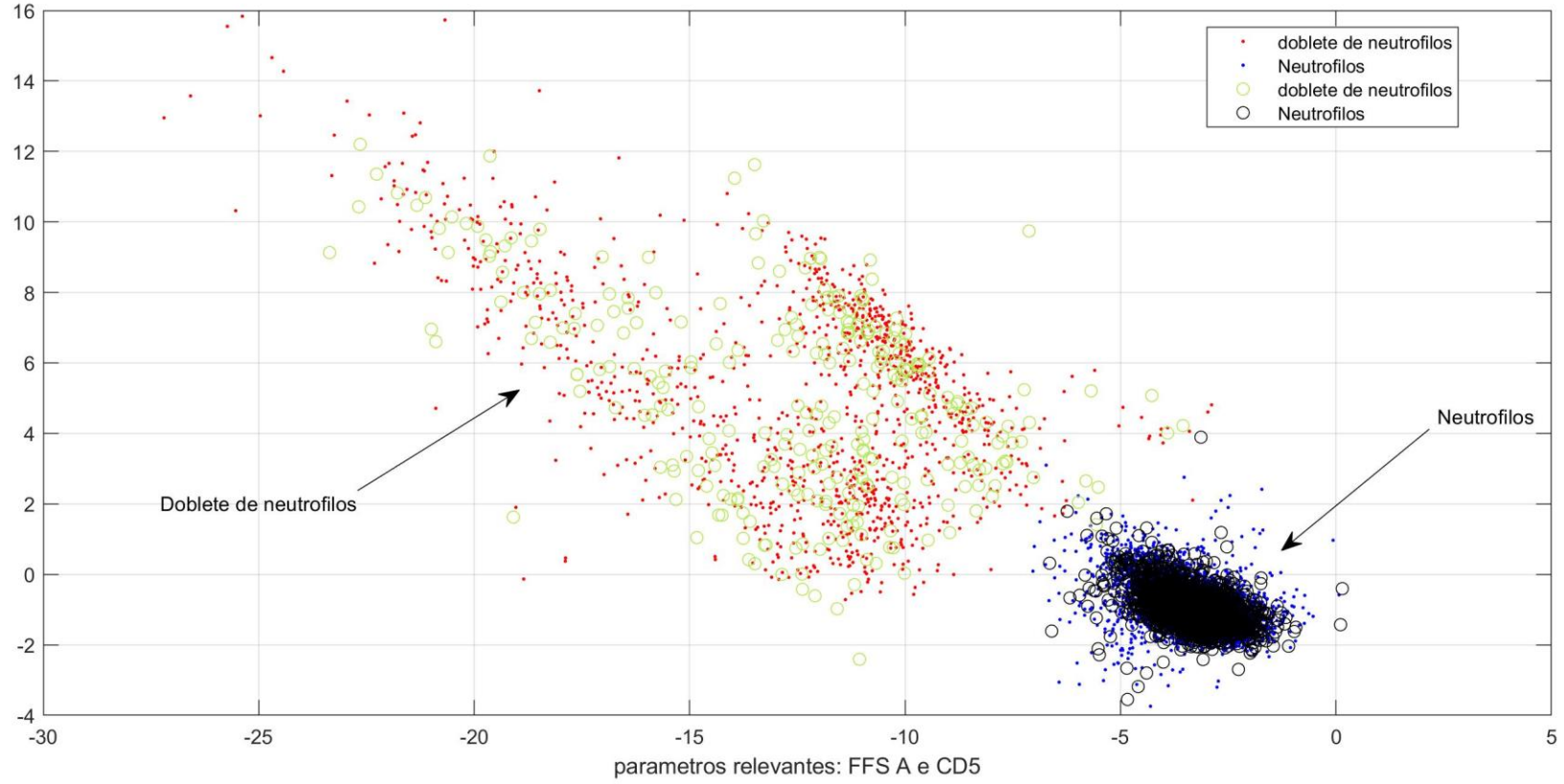


ruído


Debris
Debris
Debris
Debris
Debris
Basophils
CD4+CD8dim
CD56bright
CD56dim
Eosinophils
Myeloid DC
Monocytoid DC
Neutrophils (Unspecific staining)
data1
data2
data3
data4
data5
data6
data7
data8
data9
data10
200
150
100
50
0
-50
-100
-50
0
50
100
150
200
250

Classificação de populações via NCA




doblete de neutrofilos
Neutrofilos
doblete de neutrofilos
Neutrofilos
Doblete de neutrofilos
Neutrofilos
parametros relevantes: FFS A e CD5
16
14
12
10
8
6
4
2
0
-2
-4
-30
-25
-20
-15
-10
-5
0
5