Linear Discriminant Analysis

Renato Assunção

DCC-UFMG

2020

Linear Discriminant Analysis

- Em 1936, (Sir) Ronald A. Fisher, o maior estatístico que já existiu e um dos maiores geneticistas do mundo, criou uma regra de classificação muito popular até hoje.
- É chamada de LDA: Linear Discriminant Analysis.
- Mas, espere um pouco...
- Se temos a regra ótima (Optimal Bayes Classifier), por quê aprender uma regra antiga e diferente?
- Por dois motivos: (a) o LDA está conectado com a regra ótima. (b) ele fornece uma abordagem bem diferente para o nosso problema de classificação: ele o vê como um problema de *separação* de populações.

Função Discriminante de Fisher

- *Linear Discriminant Analysis (LDA)* para duas classes.
- $\mathbf{X} = (X_1, X_2, \ldots, X_p)$
- Fisher: Vamos criar um índice univariado (escalar) calculando

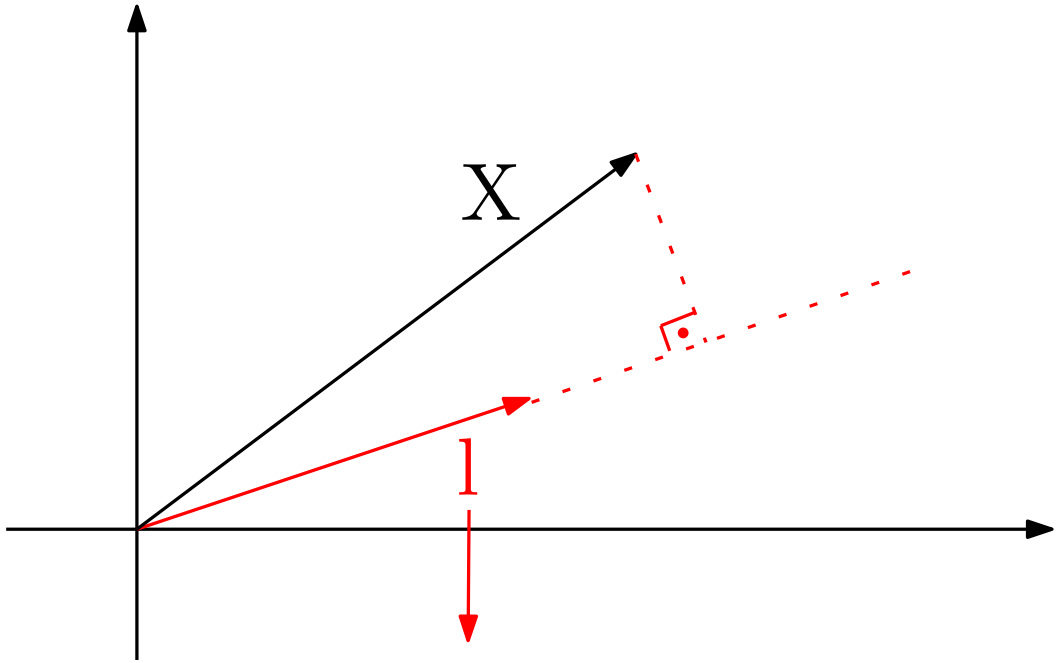
$$\begin{aligned} Y &= \boldsymbol{\ell}^t \mathbf{X} \\ &= \ell_1 X_1 + \ell_2 X_2 + \ldots + \ell_p X_p \end{aligned}$$

onde $\boldsymbol{\ell}^t = (\ell_1, \ell_2, \ldots, \ell_p)$ é um vetor de constantes.

- Queremos escolher o vetor ℓ de forma que:

$$\begin{cases} Y's \text{ de pop1} \\ Y's \text{ de pop2} \end{cases} \longrightarrow \text{o mais separado possivel}$$

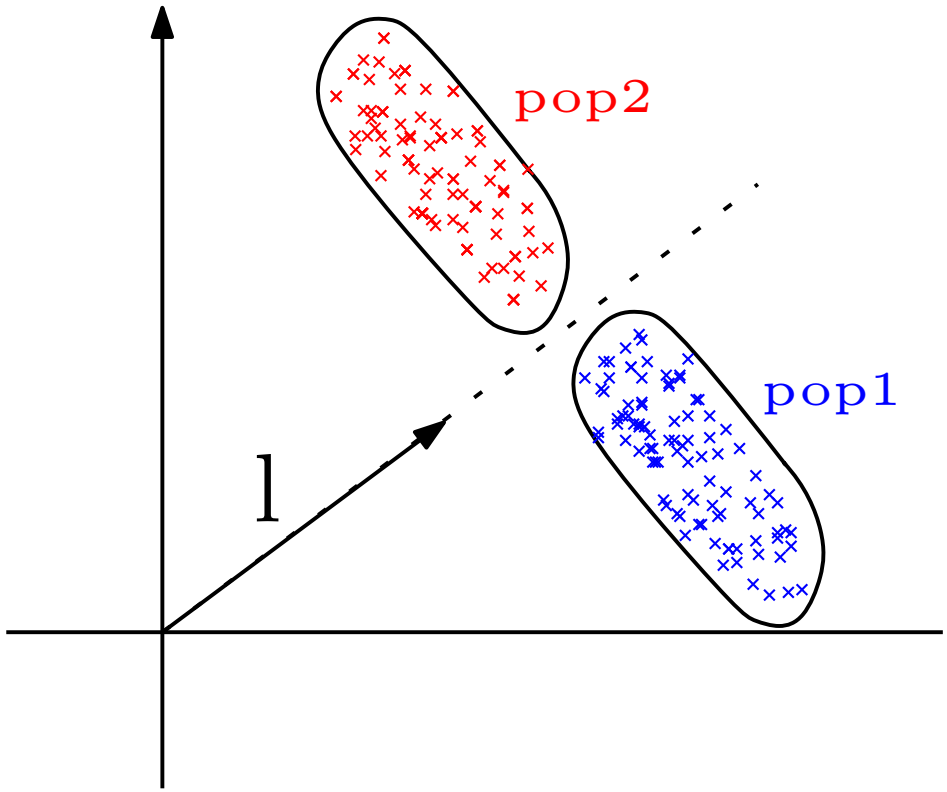
- Projeção ortogonal de $\mathbf{X}$ em ℓ:



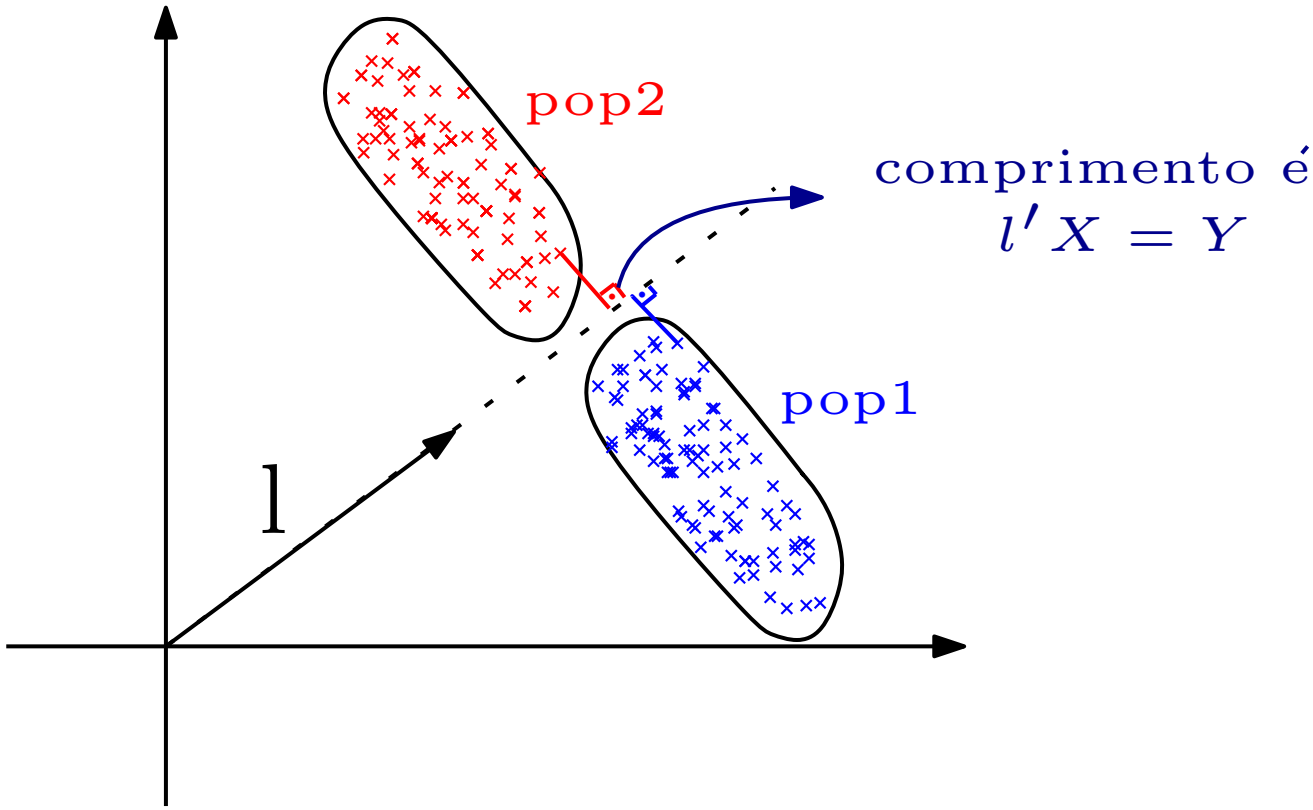
$$\frac{X'l}{\|l\|^2} \cdot l = (X'l) \cdot l, \text{ se } \|l\|^2 = 1$$

$$\downarrow$$

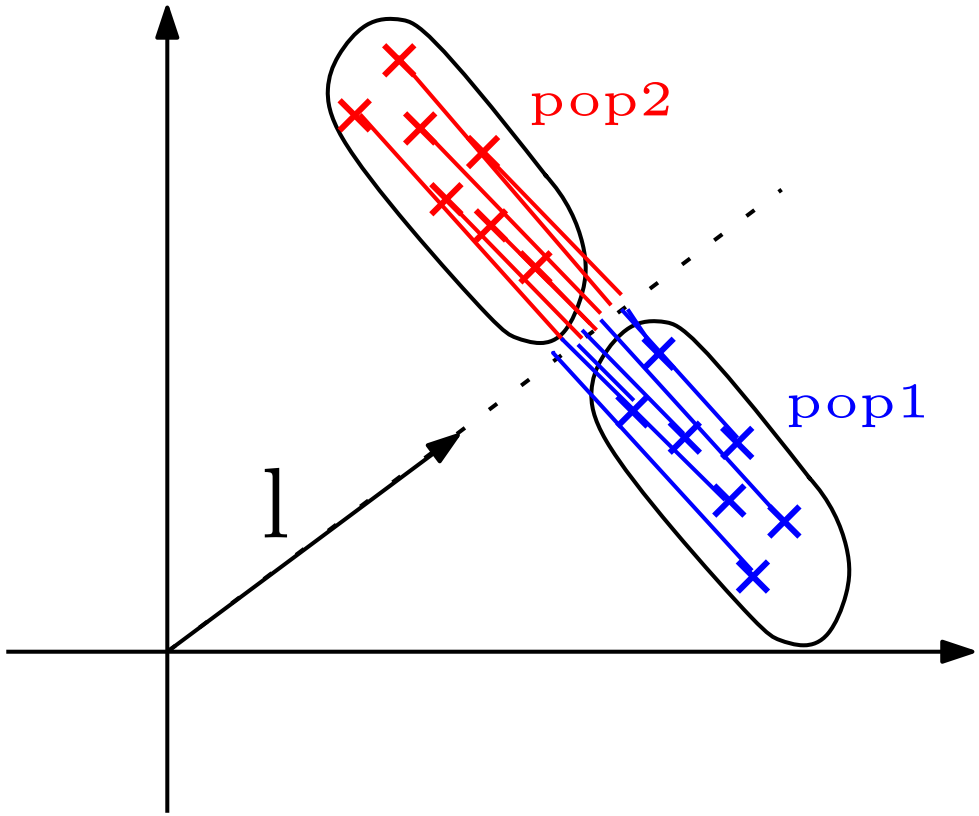
"tamanho das projecoes"



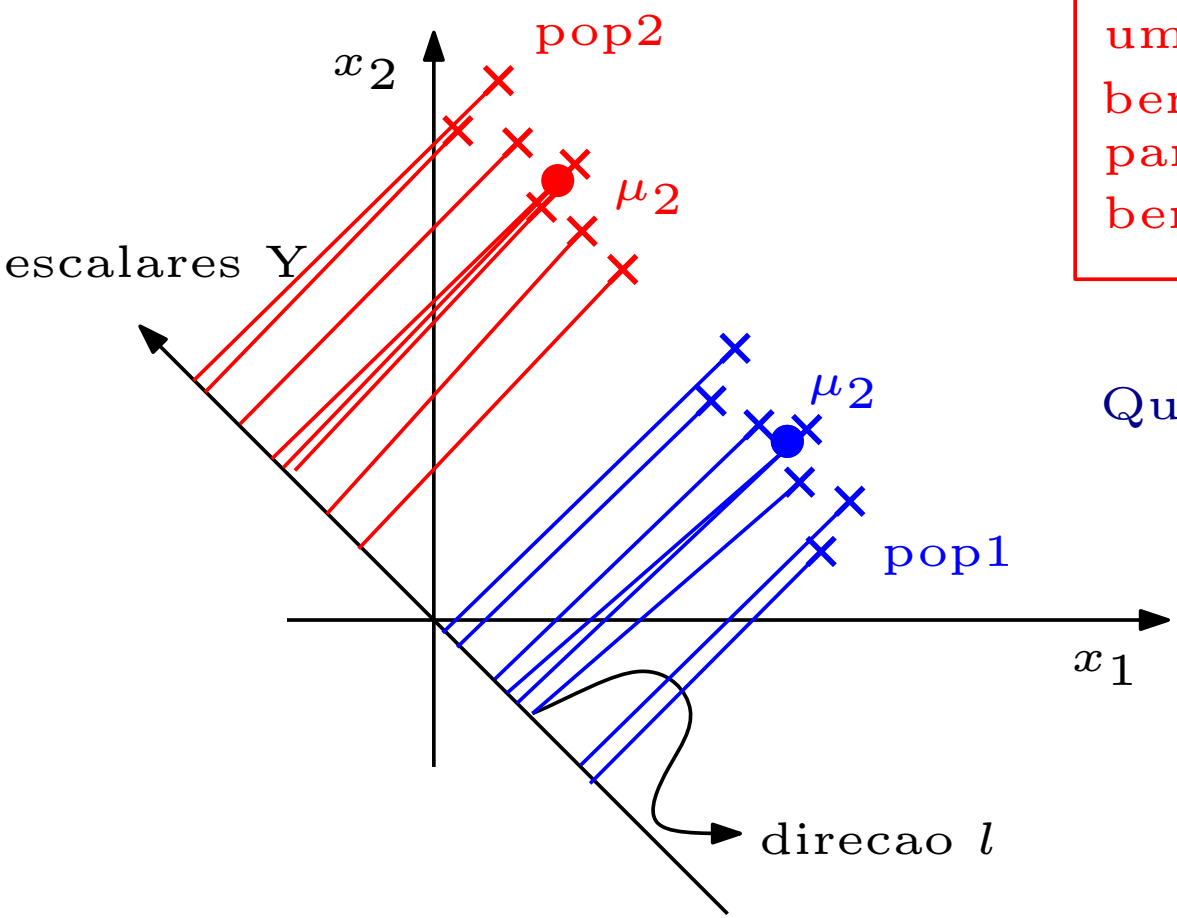
- Suponha que $\|\ell\|^2 = 1$ (vetor de comprimento 1)
- Estamos buscando direção ℓ em que $\ell^t \mathbf{X} = Y$ dos dois grupos sejam maximalmente separados
- O vetor ℓ acima é uma <u>má escolha!</u>



- Projeção $Y = \ell^t \mathbf{X}$ de dois vetores $\mathbf{X}$: um de pop1, outro de pop2.

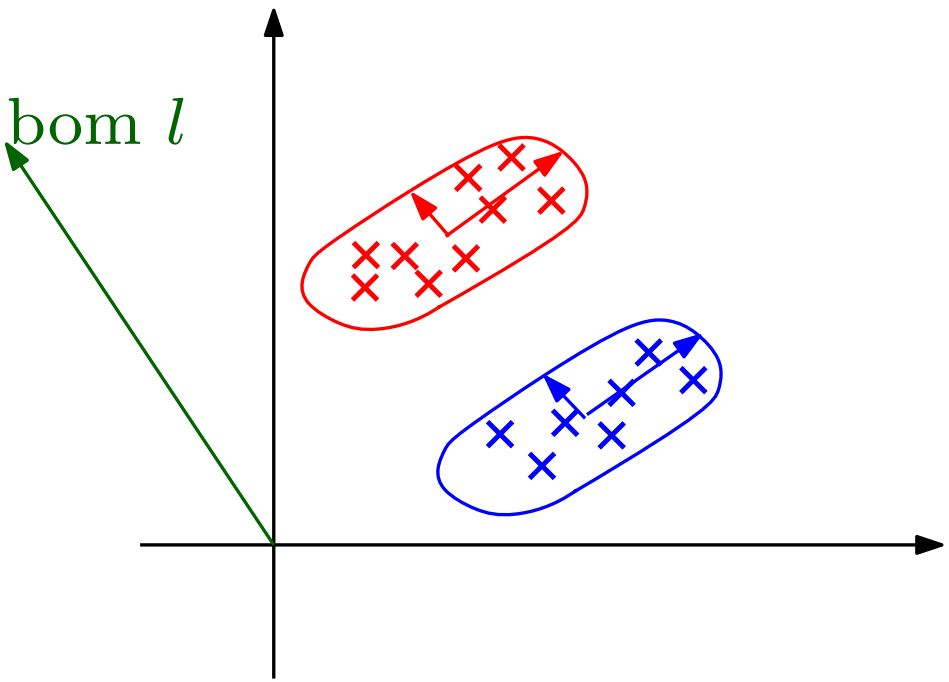


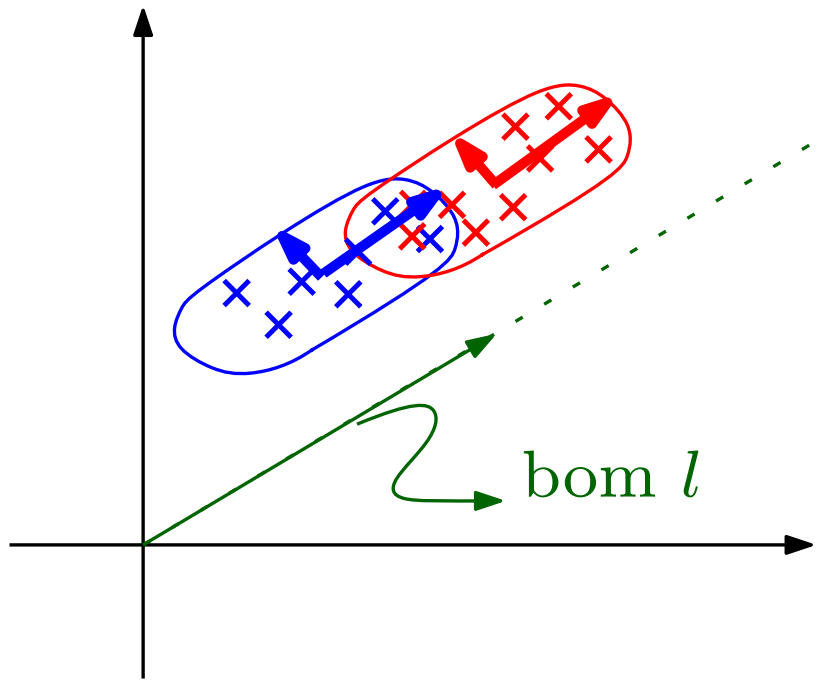
- Projeção ao longo de ℓ de todos os vetores $\mathbf{X}$ gerando os escalares Y_{11} Y_{12} ... Y_{1m_1} (pop1) e Y_{21} Y_{22} ... Y_{2m_2} (pop2).
- Pop1 e pop2 não ficam separadas.



uma direção l
bem melhor para
para gerar grupos
bem separados

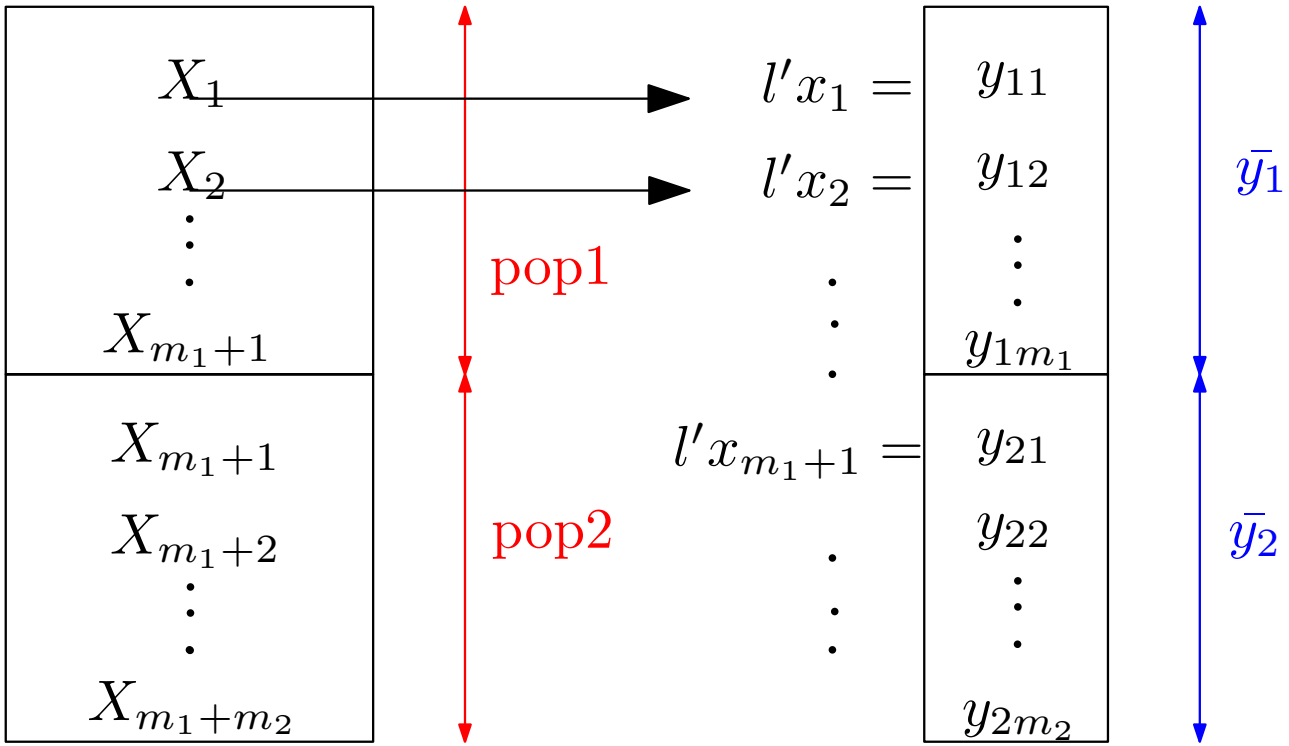
Qual a melhor direção de l?





- Melhor direção não tem associação simples com os autovetores de Σ
- À esquerda, bom $l \approx 2^o$ (menor) autovetor.
- À direita, bom $l \approx 1^o$ (maior) autovetor.

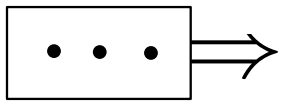
- Na maioria das vezes, não é nenhum dos autovetores de Σ
- Para encontrar a solução Fisher raciocinou assim inicialmente:
  - Projete cada ponto $\mathbf{X}$ ao longo de ℓ gerando o escalar $Y = \ell^t \mathbf{X}$.
  - Calcule a média de: pop1 $= \bar{y}_1$ e a média de pop2 $= \bar{y}_2$
- Procure a direção ℓ em que $\|\bar{y}_1 - \bar{y}_2\|$ seja máxima.



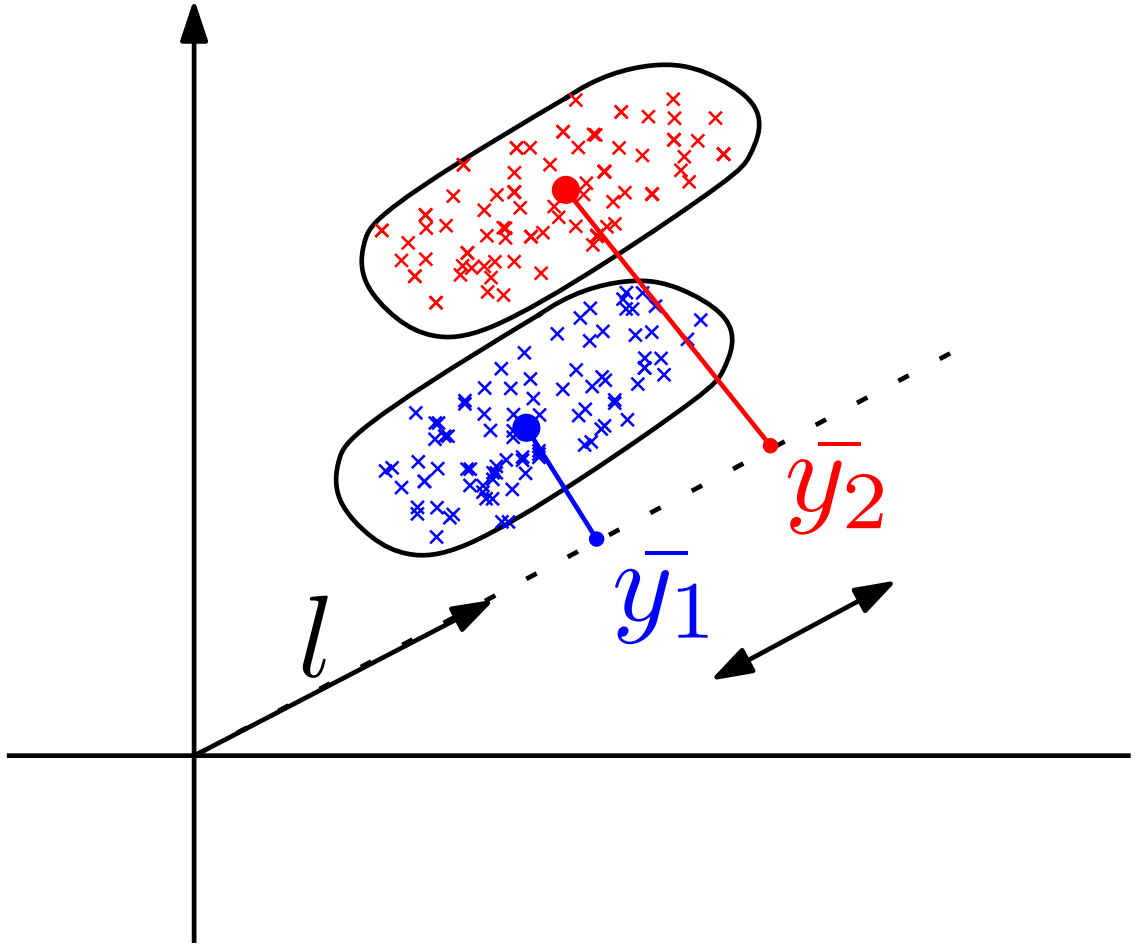
Matriz dos dados $(m_1 + m_2)$xp

n^{os} reais escalares

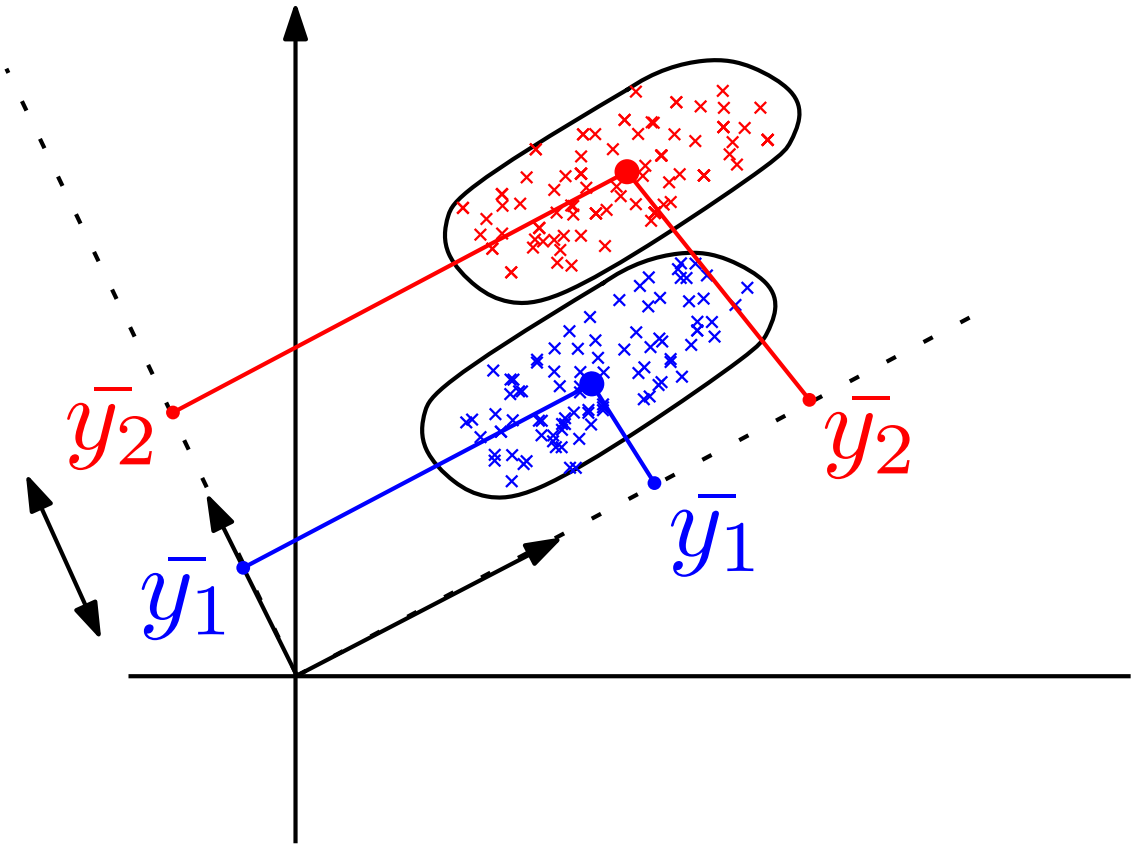
Separação entre os grupos das duas populações $= \|\bar{y}_1 - \bar{y}_2\|$



- Mas isso tem um problema:



- $\|\bar{y}_1 - \bar{y}_2\|$ pode ser grande mas as projeções não estão bem separadas.



- Esta é uma outra direção em que $\|\bar{y}_1 - \bar{y}_2\|$ é menor do que a anterior, mas que separa muito melhor as duas populações.

Conclusão fundamental

- A projeção dos dados aleatórios $\mathbf{X}$ em um direção fixa ℓ produz a variável aleatória $Y = \ell^t \mathbf{X}$, uni-dimensional.
- Para medir a separação entre as populações projetadas não devemos simplesmente olhar $\|\bar{y_1} - \bar{y_2}\|$.
- A razão é que a projeção impacta não apenas as médias $\bar{y_1}$ e $\bar{y_2}$ *mas impacta também a variabilidade* dos dados.
- Assim, devemos considerar $\|\bar{y_1} - \bar{y_2}\|$ *relativamente* ao desvio padrão s_y dos dados projetados.
- Isto é, vamos procurar ℓ para maximizar $\frac{\|\bar{y_1} - \bar{y_2}\|}{s_y}$.

Formulação do problema

- Vamos supor que temos dados de duas populações ou duas classes: π_1 e π_2.
- Dados são vetores aleatórios $\mathbf{X}$ de dimensão $p \times 1$ com densidades $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$.
- Dados *não precisam* ser gaussianos.
- As classes possuem médias diferentes mas a *mesma* matriz de covariância:
- $\mathbb{E}(\mathbf{X}| \in \pi_1) = \boldsymbol{\mu}_1, \quad p \times 1$
- $\mathbb{E}(\mathbf{X}| \in \pi_2) = \boldsymbol{\mu}_2, \quad p \times 1$
- $\mathbb{V}(\mathbf{X}| \in \pi_1) = \mathbb{V}(\mathbf{X}| \in \pi_2) = \boldsymbol{\Sigma}, \quad p \times p$

Função objetivo

- Com um vetor $\ell \in \mathbb{R}^p$, reduzimos o vetor p-dimensional $\mathbf{X}$ a um escalar uni-dimensional: $Y = \ell^t \mathbf{X}$.
- Para π_1, o valor esperado de Y será $\mathbb{E}(Y| \in \pi_1) = m_1 = \ell^t \boldsymbol{\mu}_1$.
- Para π_2, temos $\mathbb{E}(Y| \in \pi_2) = m_2 = \ell^t \boldsymbol{\mu}_2$.
- Isto é, a média das projeções é a projeção da média.
- A variância da projeção Y em torno de suas duas médias é a mesma: $\mathbb{V}(Y| \in \pi_1) = \mathbb{V}(Y| \in \pi_2) = \ell^t \boldsymbol{\Sigma} \ell \quad 1 \times 1$
- Veja que a variância de Y muda com ℓ.
- Queremos encontrar ℓ que maximize a separação das duas populações.
- Como medir a separação?

Função objetivo

- Como discutimos, não queremos apenas maximizar $\|m_1 - m_2\|$.
- Devemos considerar também a dispersão (variância, desvio-padrão) de Y em torno de suas duas médias m_1 e m_2.
- Função objetivo: Queremos ℓ que maximize

$$
\begin{aligned}
U &= \frac{\|m_1 - m_2\|^2}{\mathbb{V}(Y)} \\
&= \frac{\|\mathbb{E}(Y| \in \pi_1) - \mathbb{E}(Y| \in \pi_2)\|^2}{\mathbb{V}(Y)} \\
&= \frac{\left\|\ell^t \mu_1 - \ell^t \mu_2\right\|^2}{\ell^t \mathbf{\Sigma} \ell} = \frac{\left\|\ell^t (\mu_1 - \mu_2)\right\|^2}{\ell^t \mathbf{\Sigma} \ell} \\
&= \frac{\ell^t (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t \ell}{\ell^t \mathbf{\Sigma} \ell}
\end{aligned}
$$

- Precisamos da última representação para usar um teorema de álgebra linear

Teorema de álgebra linear - OPCIONAL - LER SOZINHO

- Seja $\mathbf{v}$ um vetor $p \times 1$ e $\mathbf{\Sigma}$ uma matriz $p \times p$ positiva definida e simétrica.
- Seja $\mathbf{x}$ um vetor $p \times 1$ não-nulo.
- Vamos considerar o comprimento (ao quadrado) de $\mathbf{x}$ usando a distância estatística (baseada em $\mathbf{\Sigma}$) até a origem $\mathbf{0}$ e definida por $d^2 = \mathbf{x}^t \mathbf{\Sigma} \mathbf{x}$.
- Assim, $\frac{\mathbf{x}}{d} = \frac{\mathbf{x}}{\sqrt{\mathbf{x}^t \mathbf{\Sigma} \mathbf{x}}}$ é um vetor de comprimento 1 (ou norma-$\mathbf{\Sigma}$ igual a 1).
- Seja $\mathcal{B}$ o conjunto de vetores de norma-$\mathbf{\Sigma}$ igual a 1.
- O conjunto $\mathcal{B}$ é um p-dim elipsóide com eixos nas direções dos auto-vetores de $\mathbf{\Sigma}$.

Teorema de álgebra linear - OPCIONAL - LER SOZINHO

- Dentre todos os vetores $\mathbf{w}$ com norma-$\mathbf{\Sigma}$ igual a 1
  - (isto é, dentre todos os vetores $\mathbf{w} = \mathcal{B}$),
- aquele vetor $\mathbf{w}$ que maximiza

$$\max_{\mathbf{w} \in \mathcal{B}} (\mathbf{w}^t \mathbf{v})^2 = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{x}^t \mathbf{v}\|^2}{\mathbf{x}^t \mathbf{\Sigma} \mathbf{x}}$$

- é igual a

$$\mathbf{w} = c \mathbf{\Sigma}^{-1} \mathbf{v} \, .$$

Aplicando o Teorema no problema de LDA

- No problema do LDA precisamos encontrar um vetor ℓ que maximize

$$U = \frac{\left\| \ell^t(\mu_1 - \mu_2) \right\|^2}{\ell^t \mathbf{\Sigma} \ell}$$

- Isto cai perfeitamente no caso do teorema anterior e a solução é

$$\ell = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$$

- Na prática, temos de estimar os parâmetros das duas distribuições com os dados da amostra.
- Por exemplo, μ_1 vira o vetor de médias aritméticas dos dados da amostra.

Solução de Fisher com os dados amostrais

- Queremos maximizar a separação $= \frac{\|\bar{y}_1 - \bar{y}_2\|^2}{S_y^2}$
  - S_y = DP amostral com as observações no eixo da projeção

$$S_y^2 = \frac{1}{m_1 + m_2 - 2} \left(\sum_{j=1}^{m_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{m_2} (y_{2j} - \bar{y}_2)^2 \right)$$

- A combinação linear $Y = \ell^t X$ que maximiza a separação $\frac{\|\bar{y}_1 - \bar{y}_2\|}{S_y}$ é dada por

$$\ell^t = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t \left(S^2_{pooled} \right)^{-1} .$$

- $S^2_{pooled} = \frac{m_1}{m} S^2_{pop1} + \frac{m_2}{m} S^2_{pop2}$;
  - $S^2_{pop1} \rightarrow$ Matriz de variância e covariância amostral da pop1.
- Prova: Johnson & Wichern, Applied Multivariate Statistical Analysis.

Como usar para classificar? Algoritmo

- *INPUT:* Dados $\mathbf{X}$ (vetores p-dim) seperados em duas classes.
- Obtenha $\bar{\mathbf{x}}_1$ (vetor p-dim) com as médias das p variáveis dos dados da classe 1.
- Obtenha S^2_{pop1} (matriz $p \times p$) com as variâncias e covariâncias das p variáveis dos dados da classe 1.
- Classe 2: $\bar{\mathbf{x}}_2$ e S^2_{pop2}.
- Misture as duas matrizes: $S^2_{pooled} = \frac{m_1}{m} S^2_{pop1} + \frac{m_2}{m} S^2_{pop2}$.
- Calcule $\ell = \left(S^2_{pooled}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, um vetor $p \times 1$.

Como usar para classificar? Algoritmo

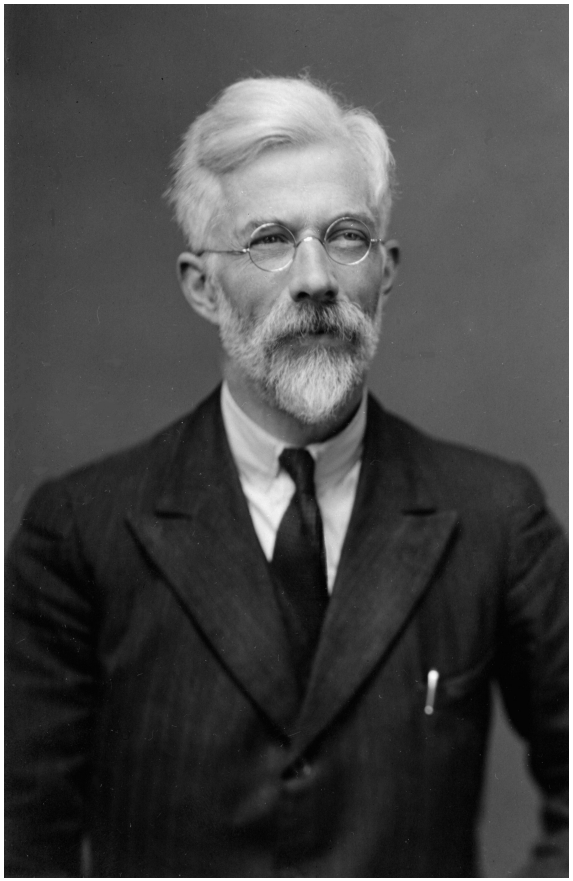
- Projete cada exemplo do conjunto de treinamento obtendo o escalar (número) $Y = \ell^t \mathbf{X}$ para cada um deles.
- Calcule a média dos Y's da classe 1 : $\bar{y}_1$
- Calcule a média dos Y's da classe 1 : $\bar{y}_2$
- Calcule a média simples das duas médias: $m = (\bar{y}_1 + \bar{y}_2)/2$

Como usar para classificar? Algoritmo

- Recebe *um novo exemplo* $\mathbf{x}$ (um vetor $p \times 1$) do *qual não conhecemos a classe*.
- Projete este $\mathbf{x}$ ao longo do vetor ℓ obtendo o escalar $y = \ell^t \mathbf{x}$.
- Regra de classificação:
  - Aloque à classe 1 se $y \geq m$
  - Caso contrário, aloque à classe 2.
- Se a classes forem bem separáveis em uma direção, o LDA vai enontrá-la.
- Como saber se funciona? Particione os dados em treino-teste e meça precision, recall, etc.

- LDA de Fisher assume que as matrizes de covariância Σ_1 e Σ_2 das duas classes sejam iguais.
- Assim, Fisher LDA fornece uma direção ℓ em que, se projetarmos os dados, teremos o máximo de separação entre as populações.
- Podemos usar LDA para classificação, mas...
- Este procedimento resulta na mesma regra de classificação ótima vista antes no caso gaussiano com $\Sigma_1 = \Sigma_2$ e iguais custos e prioris.
- Assim, ao fazer LDA estamos fazendo a regra de classificação ótima de Bayes! (sob estas hipóteses).
- O caso LDA com $\Sigma_1 \neq \Sigma_2$ foi estudado por Rao e é também a a regra de classificação ótima de Bayes no caso gaussiano com custos e prioris iguais.

Sir Ronald A. Fisher



THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

By R. A. FISHER, Sc.D., F.R.S.

I. Discriminant functions

When two or more populations have been measured in several characters, $x_1, \ldots, x_s$, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated. At the author's suggestion use has already been made of this fact in craniometry (a) by Mr E. S. Martin, who has applied the principle to the sex differences in measurements of the mandible, and (b) by Miss Mildred Barnard, who showed how to obtain from a series of dated series the particular compound of cranial measurements showing most distinctly a progressive or secular trend. In the present paper the application of the same principle will be illustrated on a taxonomic problem; some questions connected with the precision of the processes employed will also be discussed.

Calyampudi Radhakrishna Rao (C. R. Rao)

Journal of the Royal Statistical Society

SERIES B (METHODOLOGICAL)

Vol. X, No. 2, 1948

THE UTILIZATION OF MULTIPLE MEASUREMENTS IN PROBLEMS OF BIOLOGICAL CLASSIFICATION*

By C. RADHAKRISHNA RAO

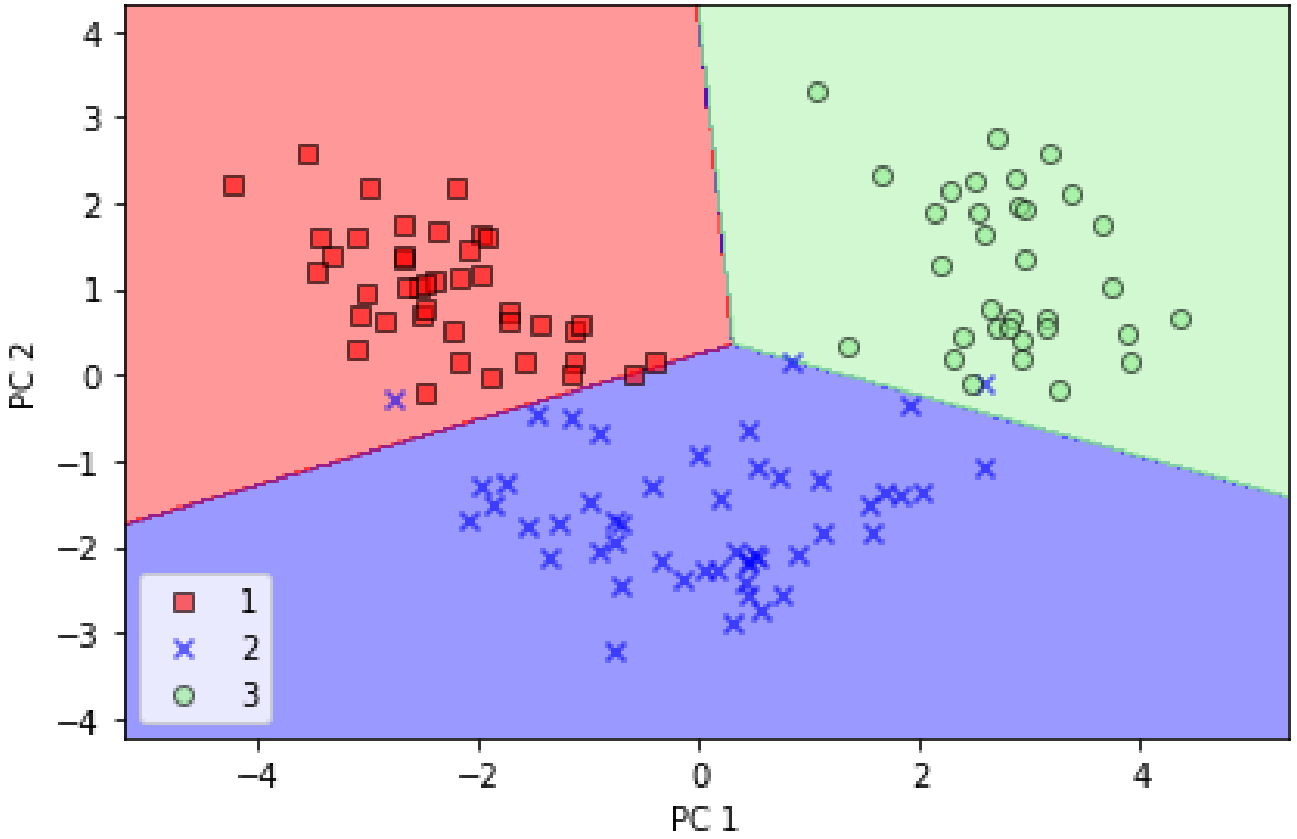
From the Duckworth Laboratory, University Museum of Archaeology and Ethnology, Cambridge

[Read before the RESEARCH SECTION of the ROYAL STATISTICAL SOCIETY, April 6, 1948, Dr. J. O. IRWIN in the Chair]

LDA como redução de dimensionalidade

- Vimos como obter UMA ÚNICA direção ℓ para projetar os dados onde as classes sejam o mais separável possível.
- Se os dados $\mathbf{X}$ são vetores com, por exemplo, 13-dim, porque ficar em apenas UMA ÚNICA direção ℓ?
- Podemos procurar DUAS direções ℓ_1 e ℓ_2 para projetar os dados.
- Assim, cada vetor $\mathbf{X}$ de dimensão 13 fica reduzido a um vetor de DUAS dimensões (y_1, y_2) onde:
  - $y_1 = \ell_1^t \mathbf{x}$
  - $y_2 = \ell_2^t \mathbf{x}$
- Podemos visualizar os dados como pontos no plano.
- Talvez LDA com uma única direção não funcione bem mas com duas, sim!

LDA com dois componentes



LDA como redução de dimensionalidade

- Vamos apresentar apenas a solução do LDA com mais de um componente.
- Não vamos deduzir esta solução.
- Interessados na prova, consultar Murphy(2012) Machine Learning, a Probabilistic Perspective.
- As s primeiras componentes do LDA são os s autovetores principais $\ell_1, \ell_2, \ldots, \ell_s$ da matriz $\mathbf{S}_w^{-1}\mathbf{S}_b$ onde
  - $\mathbf{S}_w$ é a matriz de covariância das variáveis *dentro (within)* das classes.
  - $\mathbf{S}_b$ é a matriz de covariância das *médias* das variáveis *entre (between)* as classes.