SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

# Aula 11a: **Análise discriminante linear (LDA)**


Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele



Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Prentice Hall.

Mingoti, S. A. (2007) Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Editora UFMG.

# Análise discriminante

**Objetivos**:

- **Classificação** de elementos de uma amostra ou população,
- **Redução de dimensionalidade** de forma que se obtenha um bom classificador com o número mínimo de dimensões possível.

Difere da análise de agrupamentos pelo fato de que são determinados previamente os grupos aos quais serão direcionados os elementos da amostra.

A análise discriminante pode ser considerada uma técnica de aprendizado supervisionado, enquanto que a análise de agrupamentos é uma técnica de aprendizado não-supervisionado

# Análise discriminante

**Objetivos**:

- **Classificação** de elementos de uma amostra ou população,
- **Redução de dimensionalidade** de forma que se obtenha um bom classificador com o número mínimo de dimensões possível.

Difere da análise de agrupamentos pelo fato de que são determinados previamente os grupos aos quais serão direcionados os elementos da amostra.

A análise discriminante pode ser considerada uma técnica de aprendizado supervisionado, enquanto que a análise de agrupamentos é uma técnica de aprendizado não-supervisionado

# Análise discriminante

A análise discriminante linear (*linear discriminant analysis* ou LDA) foi proposta originalmente por **Sir Ronald Fisher** em **1936**, a princípio para duas classes.

Em 1948, **C. R. Rao** propos uma **generalização para múltiplas classes**.

# Análise discriminante

Suponha que tenhamos $n_1$ e $n_2$ elementos amostrais procedentes das populações A e B, respectivamente, e que em cada um dos $n = n_1 + n_2$ tenham sido observadas $p$ características.

Se observarmos um novo elemento amostral, cuja origem é incerta, como compará-lo ao perfil geral dos grupos A e B e **classificá-lo** como pertencente a um deles?

# Análise discriminante

Suponha que tenhamos $n_1$ e $n_2$ elementos amostrais procedentes das populações A e B, respectivamente, e que em cada um dos $n = n_1 + n_2$ tenham sido observadas $p$ características.

Se observarmos um novo elemento amostral, cuja origem é incerta, como compará-lo ao perfil geral dos grupos A e B e **classificá-lo** como pertencente a um deles?

# Análise discriminante

**Exemplo**:

Em medicina, é comum querer identificar fatores de risco ou distinguir doenças que tenham alguma similaridade sintomática aos apresentados pelo paciente.

Ou seja, deseja-se identificar um perfil de portadores ou não de uma determinada doença para classificar novos pacientes como prováveis ou não prováveis portadores da patologia em questão.

Outros exemplos podem ser encontrados em áreas como educação, finanças, entre outras.

# Análise discriminante

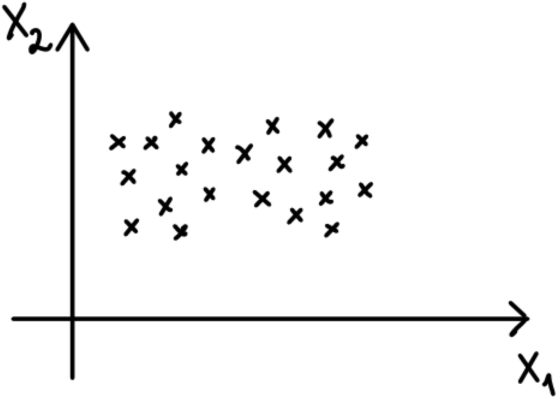
**Exemplo**:

Em medicina, é comum querer identificar fatores de risco ou distinguir doenças que tenham alguma similaridade sintomática aos apresentados pelo paciente.

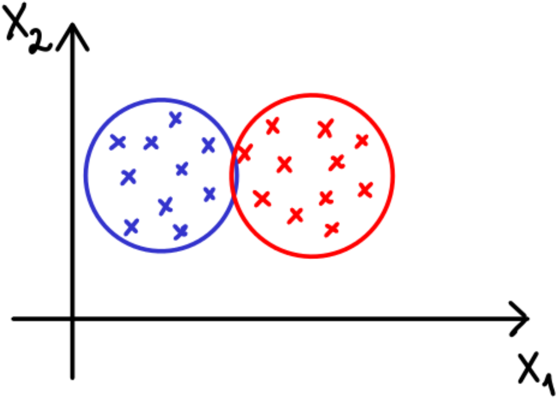
Ou seja, deseja-se identificar um perfil de portadores ou não de uma determinada doença para classificar novos pacientes como prováveis ou não prováveis portadores da patologia em questão.

Outros exemplos podem ser encontrados em áreas como educação, finanças, entre outras.

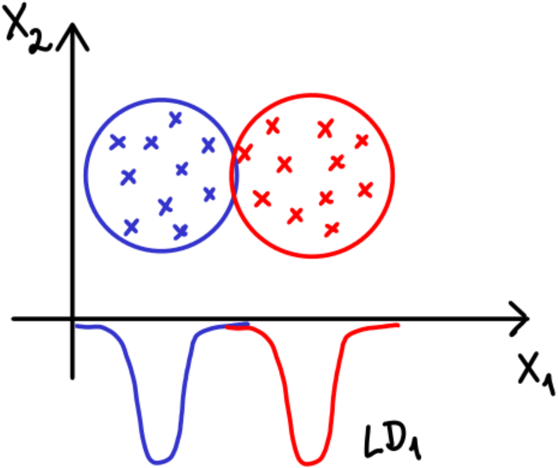
# Análise discriminante - Exemplo



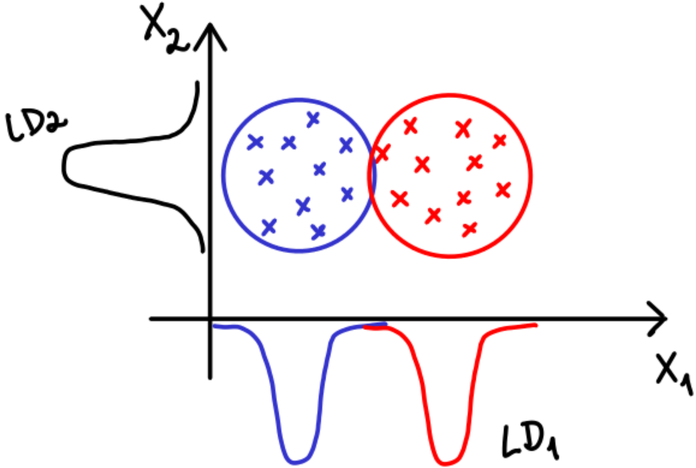
# Análise discriminante - Exemplo



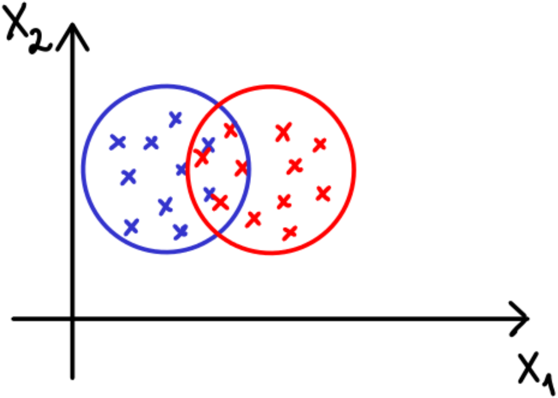
# Análise discriminante - Exemplo



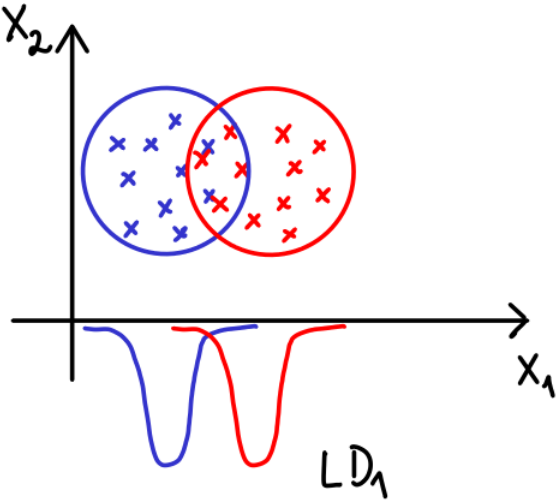
# Análise discriminante - Exemplo



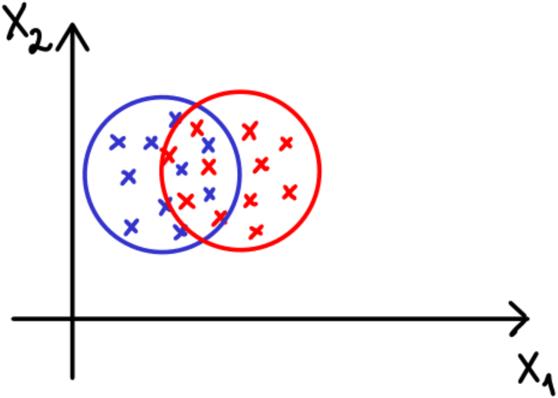
# Análise discriminante - Exemplo



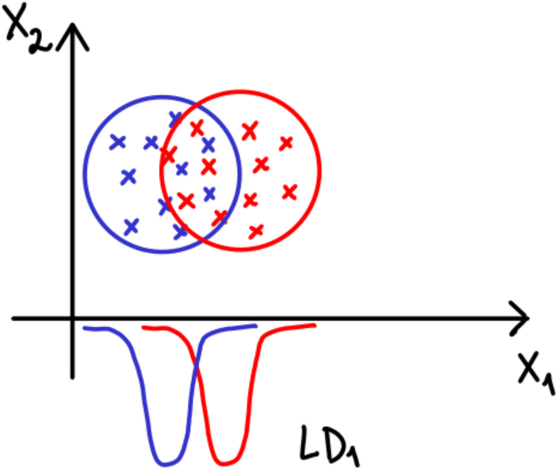
# Análise discriminante - Exemplo



# Análise discriminante - Exemplo



# Análise discriminante - Exemplo



# Análise discriminante

Um bom procedimento de classificação resulta em poucas classificações incorretas.

Se a distribuição de probabilidade das características medidas nos elementos amostrais de cada população for conhecida, é possível utilizar o princípio da máxima verossimilhança (Casella e Berger, 2002) para construir uma regra de classificação que minimize a chance de classificar um elemento amostral incorretamente.

Casella, G.; Berger, R. L., (2002) Statistical inference. Pacific Grove, CA: Duxbury.

# Análise discriminante

Suponha que uma escola adote um processo seletivo de duas fases. Seja $X$ a nota na prova de Matemática de candidatos na fase 1, e considere duas populações de alunos:

**População 1**: Alunos que passaram na $1^a$ fase mas não foram aprovados na $2^a$ fase.

**População 2**: Alunos que passaram em ambas as fases do vestibular.

# Análise discriminante

A partir dos dados, deseja-se criar uma regra de classificação que permita identificar, dentre os aprovados na primeira fase, quais provavelmente serão aprovados na segunda fase. Suponha, no caso univariado, que

**População 1**: $X \sim N(\mu_1, \sigma^2)$
**População 2**: $X \sim N(\mu_2, \sigma^2)$.

Agora, para cada possível nota $x$ de um candidato, pode-se calcular uma razão de probabilidades

$$\lambda(x) = \frac{f_1(x)}{f_2(x)},$$

que indica a razão de densidades de $x$ na população 1 e 2.

# Análise discriminante

A partir dos dados, deseja-se criar uma regra de classificação que permita identificar, dentre os aprovados na primeira fase, quais provavelmente serão aprovados na segunda fase. Suponha, no caso univariado, que

**População 1**: $X \sim N(\mu_1, \sigma^2)$
**População 2**: $X \sim N(\mu_2, \sigma^2)$.

Agora, para cada possível nota $x$ de um candidato, pode-se calcular uma razão de probabilidades

$$\lambda(x) = \frac{f_1(x)}{f_2(x)},$$

que indica a razão de densidades de $x$ na população 1 e 2.

# Análise discriminante

Se $f_1$ e $f_2$ são densidades da distribuição normal como suposto anteriormente, temos

$$\lambda(x) = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\}}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma^2}\right\}},$$

que pode ser simplificado por

$$\lambda(x) = \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma}\right)^2 - \left(\frac{x-\mu_2}{\sigma}\right)^2\right]\right\}$$

# Análise discriminante

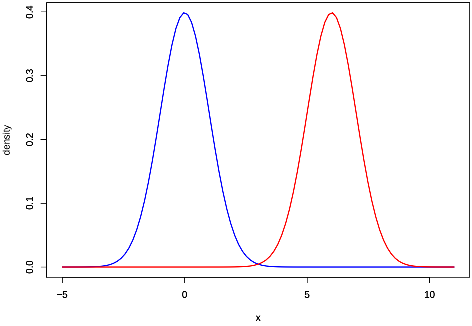
Se $\lambda(x) > 1$, então é razoável classificar o candidato como um provável não aprovado na $2^a$ fase.

Se $\lambda(x) < 1$, ele é um provável aprovado em ambas as fases.

Se $\lambda(x) = 1$, então as probabilidades são as mesmas de estar na população 1 e na população 2, segundo esse critério.

# Análise discriminante

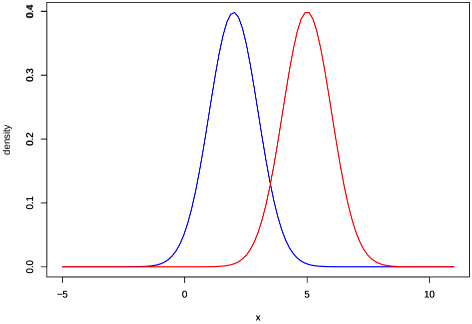
A qualidade da discriminação dependerá do grau de intersecção das duas densidades.




Casos como esse podem ocasionar poucas ou nenhuma classificação incorreta, ou seja, há **forte poder de discriminação**.

# Análise discriminante

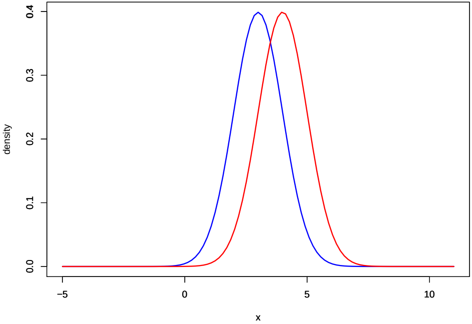
A qualidade da discriminação dependerá do grau de intersecção das duas densidades.




Casos como esse podem ocasionar poucas classificações incorretas, ou seja, há **poder razoável de discriminação**.

# Análise discriminante

A qualidade da discriminação dependerá do grau de intersecção das duas densidades.




Casos como esse podem ocasionar muitas classificações incorretas, ou seja, há **poder fraco de discriminação**.

# Análise discriminante

É comum considerar $-2\log\lambda(x)$ com as seguintes correspondências:

| $\lambda(x)$ | $-2\log\lambda(x)$ | Situação |
|---|---|---|
| $> 1$ | $< 0$ | $x$ mais próximo de $\mu_1$ |
| $< 1$ | $> 0$ | $x$ mais próximo de $\mu_2$ |
| $= 1$ | $= 0$ | $x$ igualmente próximo de $\mu_1$ e $\mu_2$ |

# Análise discriminante

As funções $\lambda(x)$ e $-2\log\lambda(x)$ são chamadas de **funções discriminantes**.

Para o caso multivariado, em que $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}_1, \Sigma_1)$ na população 1 e $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}_2, \Sigma_2)$ na população 2, temos

$$-2\log\lambda(\underset{\sim}{x}) = \log\left[\frac{(2\pi)^{p/2}|\Sigma_1|^{-1/2}\exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)^\top\Sigma_1^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)\right\}}{(2\pi)^{p/2}|\Sigma_2|^{-1/2}\exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)^\top\Sigma_2^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right\}}\right]$$

# Análise discriminante

As funções $\lambda(x)$ e $-2\log\lambda(x)$ são chamadas de **funções discriminantes**.

Para o caso multivariado, em que $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}_1, \Sigma_1)$ na população 1 e $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}_2, \Sigma_2)$ na população 2, temos

$$-2\log\lambda(\underset{\sim}{x}) = \log\left[\frac{(2\pi)^{p/2}|\Sigma_1|^{-1/2}\exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)^\top\Sigma_1^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)\right\}}{(2\pi)^{p/2}|\Sigma_2|^{-1/2}\exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)^\top\Sigma_2^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right\}}\right]$$

# Análise discriminante

ou seja,

$$\begin{aligned} -2\log\lambda(\underset{\sim}{x}) &= (\underset{\sim}{x}-\underset{\sim}{\mu}_1)^\top\Sigma_1^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)-(\underset{\sim}{x}-\underset{\sim}{\mu}_2)^\top\Sigma_2^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2) \\ &\quad +\log|\Sigma_1|-\log|\Sigma_2| \end{aligned}$$

e então classificamos $\underset{\sim}{x}$ na população 1 se $-2\log\lambda(\underset{\sim}{x})<0$ e na população 2 se $-2\log\lambda(\underset{\sim}{x})>0$ .

# Análise discriminante

Quando $\Sigma_1 = \Sigma_2 = \Sigma$, temos a **função discriminante de Fisher**:

$$fd(\underset{\sim}{x}) = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} \underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2).$$

Um elemento amostral com vetor de observações $\underset{\sim}{x}$ seria classificado na população 1 se $fd(\underset{\sim}{x}) > 0$, ou seja, se

$$(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} \underset{\sim}{x} > \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)$$

e seria classificado na população 2 se

$$(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} \underset{\sim}{x} < \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2).$$

# Análise discriminante

Quando $\Sigma_1 = \Sigma_2 = \Sigma$, temos a **função discriminante de Fisher**:

$$fd(\underset{\sim}{x}) = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} \underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2).$$

Um elemento amostral com vetor de observações $\underset{\sim}{x}$ seria classificado na população 1 se $fd(\underset{\sim}{x}) > 0$, ou seja, se

$$(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} \underset{\sim}{x} > \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)$$

e seria classificado na população 2 se

$$(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1} \underset{\sim}{x} < \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)^\top \Sigma^{-1}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2).$$

# Estimação da função discriminante

Se as matrizes de variâncias e covariâncias populacionais $\Sigma_1$ e $\Sigma_2$ são desconhecidas, como também as médias populacionais $\underset{\sim}{\mu}_1$ e $\underset{\sim}{\mu}_2$, mas é possível calcular as médias amostrais $\overline{\underset{\sim}{x}}_1$ e $\overline{\underset{\sim}{x}}_2$ e as matrizes de variâncias e covariâncias $S_1$ e $S_2$, para as populações 1 e 2, respectivamente, estima-se a função discriminante por

$$\begin{aligned} -2\log\widehat{\lambda}(\underset{\sim}{x}) &= (\underset{\sim}{x} - \overline{\underset{\sim}{x}}_1)^\top S_1^{-1}(\underset{\sim}{x} - \overline{\underset{\sim}{x}}_1) - (\underset{\sim}{x} - \overline{\underset{\sim}{x}}_2)^\top S_2^{-1}(\underset{\sim}{x} - \overline{\underset{\sim}{x}}_2) \\ &\quad + \log|S_1| - \log|S_2| \end{aligned}$$

# Funções discriminantes lineares de Fisher

Considere que $p$ variáveis $\underset{\sim}{X}$ foram observadas em elementos amostrais de $g$ populações distintas, e que não seja possível supor normalidade dos dados, mas que seja razoável assumir que

$$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g = \Sigma.$$

Sejam

- $\underset{\sim}{\mu}_i$ o vetor de médias populacionais no grupo $i$
- $\overline{\underset{\sim}{\mu}}$ o vetor de médias das médias populacionais

# Funções discriminantes lineares de Fisher

Considere a soma de produtos cruzados entre grupos

$$B_\mu = \sum_{i=1}^{g} (\underset{\sim}{\mu}_i - \overline{\underset{\sim}{\mu}})(\underset{\sim}{\mu}_i - \overline{\underset{\sim}{\mu}})^\top$$

e a combinação linear

$$Y = \underset{\sim}{a}^\top \underset{\sim}{X}.$$

# Funções discriminantes lineares de Fisher

Sejam $\pi_1, \ldots, \pi_g$ variáveis indicadoras da população $i$ à qual pertence uma determinada observação. Logo,

$$\mu_{i,Y} = E(Y|\pi_i) = \underset{\sim}{a}^\top \mathsf{E}(\underset{\sim}{X}|\pi_i) = \underset{\sim}{a}^\top \underset{\sim}{\mu}_i \text{ e}$$

$$\mathsf{Var}(Y|\pi_i) = \underset{\sim}{a}^\top \mathsf{Cov}(\underset{\sim}{X})\underset{\sim}{a} = \underset{\sim}{a}^\top \Sigma \underset{\sim}{a}$$

para todas as populações.

# Funções discriminantes lineares de Fisher

Ou seja,

$\mu_{i,Y} = \underset{\sim}{a}^\top \underset{\sim}{\mu}_i$ muda conforme a população à qual a observação pertence.

A média geral de $Y$ é dada por

$$\overline{\mu}_Y = \mathsf{E}(Y) = \frac{1}{g}\sum_{i=1}^{g} \mu_{i,Y} = \frac{1}{g}\sum_{i=1}^{g} \underset{\sim}{a}^\top \underset{\sim}{\mu}_i = \underset{\sim}{a}^\top \frac{1}{g}\sum_{i=1}^{g} \underset{\sim}{\mu}_i$$

# Funções discriminantes lineares de Fisher

Considere a razão

$$\frac{\text{(soma dos quadrados das distâncias das populações à média geral)}}{\text{(variância de Y)}} =$$

$$\frac{\sum_{i=1}^{g}(\mu_{i,Y} - \overline{\mu}_Y)^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{g}(\underset{\sim}{a}^\top \underset{\sim}{\mu}_i - \underset{\sim}{a}^\top \overline{\mu}_i)^2}{\underset{\sim}{a}^\top \Sigma \underset{\sim}{a}} = \frac{\underset{\sim}{a}^\top B_\mu \underset{\sim}{a}}{\underset{\sim}{a}^\top \Sigma \underset{\sim}{a}}$$

A razão acima mede a variabilidade entre grupos sobre a variabilidade intra grupos. O objetivo é buscar $\underset{\sim}{a}$ que maximize essa razão para obter a maior discriminação possível.

# Funções discriminantes lineares de Fisher

Em geral, como $\mu_i$ e $\Sigma$ são desconhecidos, utilizamos as estimativas $\overline{x}_i$ e $S_i$'s, para $i = 1, \ldots, g$.

Assim

$$B = \sum_{i=1}^{g} (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^\top$$

e

$$W = \sum_{i=1}^{g} (n_i - 1) S_i.$$

# Funções discriminantes lineares de Fisher

Note que $W = (n_1 + n_2 + \ldots + n_g - g)S_{pooled}$ , então a mesma constante $\widehat{\underset{\sim}{a}}$ que maximiza $\dfrac{\widehat{\underset{\sim}{a}}^\top B \widehat{\underset{\sim}{a}}}{\widehat{\underset{\sim}{a}}^\top S_{pooled} \widehat{\underset{\sim}{a}}}$ maximiza $\dfrac{\widehat{\underset{\sim}{a}}^\top B \widehat{\underset{\sim}{a}}}{\widehat{\underset{\sim}{a}}^\top W \widehat{\underset{\sim}{a}}}$.

É possível mostrar que o autovetor $\widehat{\underset{\sim}{e}}_1$ referente ao maior autovalor $\lambda_1$ de $W^{-1}B$ levam ao máximo de $\dfrac{\widehat{\underset{\sim}{a}}^\top B \widehat{\underset{\sim}{a}}}{\widehat{\underset{\sim}{a}}^\top W \widehat{\underset{\sim}{a}}}$.

# Funções discriminantes lineares de Fisher

## Discriminantes lineares de Fisher

Sejam $\widehat{\lambda_1} > \ldots > \widehat{\lambda_s} > 0$ os autovalores de $W^{-1}B$ e $\widehat{\underset{\sim}{e}}_1, \ldots, \widehat{\underset{\sim}{e}}_s$ os autovetores ortonormais correspondentes. Então o vetor de coeficientes $\widehat{\underset{\sim}{a}}$ que maximiza a razão $\dfrac{\widehat{\underset{\sim}{a}}^\top B \widehat{\underset{\sim}{a}}}{\widehat{\underset{\sim}{a}}^\top W \widehat{\underset{\sim}{a}}}$ é $\widehat{\underset{\sim}{a}}_1 = \widehat{\underset{\sim}{e}}$.

A combinação linear $\widehat{\underset{\sim}{a}}_1^\top \underset{\sim}{X}$ é chamada de **primeiro discriminante linear**.

A combinação linear $\widehat{\underset{\sim}{a}}_2^\top \underset{\sim}{X}$ com $\widehat{\underset{\sim}{a}}_2 = \widehat{\underset{\sim}{e}}_2$ é chamada de **segundo discriminante linear** e assim por diante.

# Como usar os discriminantes para classificar observações?

Seja o $k$-ésimo discriminante linear amostral,

$$\widehat{Y}_k = \widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{X}.$$

Uma possibilidade para classificar uma observação $\underset{\sim}{x}$ na $l$-ésima população é utilizar o $k$-ésimo discriminante linear amostral fazendo a verificação de que

$$(\widehat{y}_k - \overline{y}_l)^2 \leq (\widehat{y}_k - \overline{y}_i)^2 \text{ para todo } i \neq l \text{ ou}$$

$$(\widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{x} - \widehat{\underset{\sim}{a}}_k^\top \overline{\underset{\sim}{x}}_l)^2 \leq (\widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{x} - \widehat{\underset{\sim}{a}}_k^\top \overline{\underset{\sim}{x}}_i)^2 \text{ para todo } i \neq l.$$

# Como usar os discriminantes para classificar observações?

Seja o $k$-ésimo discriminante linear amostral,

$$\widehat{Y}_k = \widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{X}.$$

Uma possibilidade para classificar uma observação $\underset{\sim}{x}$ na $l$-ésima população é utilizar o $k$-ésimo discriminante linear amostral fazendo a verificação de que

$$(\widehat{y}_k - \overline{y}_l)^2 \leq (\widehat{y}_k - \overline{y}_i)^2 \text{ para todo } i \neq l \text{ ou}$$

$$(\widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{x} - \widehat{\underset{\sim}{a}}_k^\top \overline{\underset{\sim}{x}}_l)^2 \leq (\widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{x} - \widehat{\underset{\sim}{a}}_k^\top \overline{\underset{\sim}{x}}_i)^2 \text{ para todo } i \neq l.$$

# Como usar os discriminantes para classificar observações?

Seja o $k$-ésimo discriminante linear amostral,

$$\widehat{Y}_k = \widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{X}.$$

Uma possibilidade para classificar uma observação $\underset{\sim}{x}$ na $l$-ésima população é utilizar o $k$-ésimo discriminante linear amostral fazendo a verificação de que

$$(\widehat{y}_k - \overline{y}_l)^2 \leq (\widehat{y}_k - \overline{y}_i)^2 \text{ para todo } i \neq l \text{ ou}$$

$$(\widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{x} - \widehat{\underset{\sim}{a}}_k^\top \overline{\underset{\sim}{x}}_l)^2 \leq (\widehat{\underset{\sim}{a}}_k^\top \underset{\sim}{x} - \widehat{\underset{\sim}{a}}_k^\top \overline{\underset{\sim}{x}}_i)^2 \text{ para todo } i \neq l.$$