Análise Discriminante: parte 2

Prof. Caio Azevedo

- A classificação entre duas populações sob heterocedasticidade ($\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$) pode ser encontrada em Johnson & Wichern.
- A regra de classificação apresentada no supramencionado livro pode gerar resultados indesejados (altas taxas de erro), principalmente quando temos mais de duas variáveis e quando a suposição de normalidade multivariada não é válida.
- Não discutiremos a AD sob heterocedasticidade.

Classificação com várias populações baseado no CECE

- Temos g populações e cada unidade amostral pertence à uma e somente uma população. Defina:
  - Suporte da distribuição: $A = \{\boldsymbol{x} \in \mathcal{R}^p, f(\boldsymbol{x}) > 0\}$ e a respectiva partição $A = \dot{\bigcup}_{i=1}^{g} A_i$.
  - p_i : probabilidade (à priori) de um determinado indivíduo pertencer à população i.
  - $c(k|i)$: custo de classificar o indivíduo na população k dado que ele pertence à população i, naturalmente $c(i|i) = 0, \forall i$.
  - $P(k|i) = \int_{A_k} f_i(\boldsymbol{x}) d\boldsymbol{x}$: probabilidade de classificar um indivíduo na população k dado que ele pertence à população i. Além disso $P(i|i) = 1 - \sum_{k=1}^{g} P(k|i)$.

Classificação com várias populações baseado no CECE

- A esperança condicional de classificar equivocadamente uma unidade pertencente à população 1 em uma outra população (2,3,..,g) é dado por:

$$\begin{aligned} ECE(1) &= P(2|1)c(2|1) + P(3|1)c(3|1) + ... + P(g|1)c(g|1) \\ &= \sum_{k=1}^{g} P(k|1)c(k|1) \end{aligned} \quad (1)$$

- Este valor esperado condicional ocorre com probabilidade p_1. Analogamente podemos definir $ECE(2), ..., ECE(g)$ e, além disso, cada uma delas ocorre com probabilidade $p_k, k = 2, ..., g$

Classificação com várias populações baseado no CECE

- Assim o custo esperado de classificação errada é dado por:

$$\begin{aligned} CECE &= p_1 ECE(1) + p_2 ECE(2) + ... + p_g ECE(g) \\ &= \sum_{i=1}^{g} p_i \sum_{k=1, k\neq i}^{g} P(k|i)c(k|i) \end{aligned} \tag{2}$$

- A regra de classificação que minimiza o CECE (equação (2)) consiste em classificar uma dada observação ($\boldsymbol{x}_0$), associada à uma determinada unidade, na população $k, k = 1, 2, ..., g$, para o qual

$$\sum_{i=1}^{g} p_i f_i(\boldsymbol{x}_0) c(k|i)$$

é mínimo. Para a demonstração veja: Anderson. T. W. An introduction to multivariate statistical analysis. New York: John Wiley, 2003.

Classificação com várias populações baseado no CECE com custos iguais

- Nesse caso a regra passa a ser: aloca-se a unidade amostral na população k, com base em um vetor de observações $\boldsymbol{x}_0$ se

$$p_k f_k(\boldsymbol{x}_0) > p_i f_i(\boldsymbol{x}_0), \forall\, i \neq k \tag{3}$$

ou de modo equivalente, se

$$\ln(p_k f_k(\boldsymbol{x}_0)) > \ln(p_i f_i(\boldsymbol{x}_0)), \forall\, i \neq k$$

Classificação com várias populações baseado no CECE com custos iguais

- Um aspecto interessante é que a regra de clasificação (3) equivale àquela que maximiza a probabilidade à posteriori $P(k|\boldsymbol{x}_0)$ = ($\boldsymbol{x}_0$ vem da população k dado que $\boldsymbol{x}_0$ foi observado), a qual é dada por:

$$P(k|\boldsymbol{x}_0) = \frac{p_k f_k(\boldsymbol{x}_0)}{\sum_{i=1}^{g} p_i f_i(\boldsymbol{x}_0)} = \frac{\text{priori} \times \text{verossimilhança}}{\sum[\text{priori} \times \text{verossimilhança}]}$$

(Teorema de Bayes)

Método de Fisher para AD com várias populações

- Fisher também desenvolveu uma metodologia de AD considerando várias populações.
- A idéia é semelhante ao caso de duas populações, no sentido que ele buscou definir funções univariadas (função discriminante) com base nas observações multivariadas.
- Neste caso, também, a metodologia de Fisher é equivalente à regra do mínimo CECE, sob normalidade multivariada, homocedasticidade, custos de classificação errada e probabilidades à priori iguais (veja Johnson & Wichern).

Método de Fisher para AD com várias populações

- Características importantes:
  - Trabalhar com representações univariadas de dados multivariados.
  - Análises gráficas que permitem analisar de modo simples o comportamento das populações de interesse.
  - Não assume normalidade dos dados.
  - Considera a homocedasticidade.

Método de Fisher para AD com várias populações

- A idéia é trabalhar com combinações lineares das observações $\boldsymbol{x}$ ($\boldsymbol{x}_0$), ou seja, $Y = \boldsymbol{a}'\boldsymbol{x}$. Assuma que $\mathcal{E}(\boldsymbol{X}|i) = \boldsymbol{\mu}_i$ e $Cov(\boldsymbol{X}|i) = \boldsymbol{\Sigma}$.
- Assim $\mathcal{E}(Y) = \boldsymbol{a}'\boldsymbol{\mu}_i$ (população i) e $Cov(Y) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$ (para todas as populações).
- Defina $(\overline{\boldsymbol{\mu}} = \frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\mu}_i)$

$$V = \frac{\boldsymbol{a}'\left(\sum_{i=1}^{g}(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})'\right)\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} = \frac{\text{variabilidade entre grupos}}{\text{variabilidade dentro de cada grupo}}.$$

- A idéia de Fisher foi tentar separar ao máximo as populações em relação à medida V, ou seja, ele buscou maximizá-la.

Método de Fisher para AD com várias populações

- Resultado: defina $\boldsymbol{w} = \sum_{i=1}^{g}(n_i - 1)\boldsymbol{s}_i^2$ (variabilidade dentro de cada população), $\boldsymbol{b} = \sum_{i=1}^{g}(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})'$, em que $\overline{\boldsymbol{x}}_i$ é a média amostral, $\boldsymbol{s}_i^2$ é a matriz de covariâncias amostral, ambas relativas a população i e $\overline{\boldsymbol{x}} = \frac{1}{g}\sum_{i=1}^{g}\overline{\boldsymbol{x}}_i$.
- Sejam $\widetilde{\lambda}_1, \widetilde{\lambda}_2, ..., \widetilde{\lambda}_s > 0$, em que $s \leq min(g-1, p)$ os autovalores diferentes de zero de $\boldsymbol{w}^{-1}\boldsymbol{b}$ e $\widetilde{\boldsymbol{e}}_1, \widetilde{\boldsymbol{e}}_2, ..., \widetilde{\boldsymbol{e}}_s$ os respectivos autovetores, devidamente reescalonados (veja Johnson & Wichern).

Método de Fisher para AD com várias populações

- Então o vetor de coeficientes $\widetilde{\boldsymbol{a}}$ que maximiza a razão

$$\frac{\widetilde{\boldsymbol{a}}' \boldsymbol{b} \widetilde{\boldsymbol{a}}}{\widetilde{\boldsymbol{a}}' \boldsymbol{w} \widetilde{\boldsymbol{a}}}$$

é dado por $\widetilde{\boldsymbol{a}}_1 = \widetilde{\boldsymbol{e}}_1$. A combinação linear $\widetilde{y}_1 = \widetilde{\boldsymbol{a}}_1 \boldsymbol{x}$ é denominada de primeira função discriminante (amostral). Assim, a combinação linear $\widetilde{y}_k = \widetilde{\boldsymbol{a}}_k \boldsymbol{x}$ é denominada k-ésima função discriminante (amostral).

Método de Fisher para AD com várias populações

- Regra de classificação: aloca-se a observação $\boldsymbol{x}_0$ (ou seja, a unidade amostral), à população k, com base em s funções discriminantes, se:

$$\sum_{j=1}^{s}(\widetilde{y}_j - \overline{y}_{kj})^2 = \sum_{j=1}^{s}[\widetilde{\boldsymbol{a}}_j'(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_k)]^2 \leq \sum_{j=1}^{s}[\widetilde{\boldsymbol{a}}_j'(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_i)]^2, \forall\, i \neq k$$

- Nota: assim como no caso de duas populações, a regra de classificação de Fisher coincide com a regra do CECE (sob normalidade e homocedasticidade), sob custos iguais e probabilidades de classificação iguais.

Voltando ao Exemplo 1 (considerando os três grupos)

- Resultados da classificação:

	S	VE	VI
S	25	0	0
VE	0	24	1
VI	0	0	25

- TEA (%) : 1,33.

Medidas resumo

Função discriminante 1

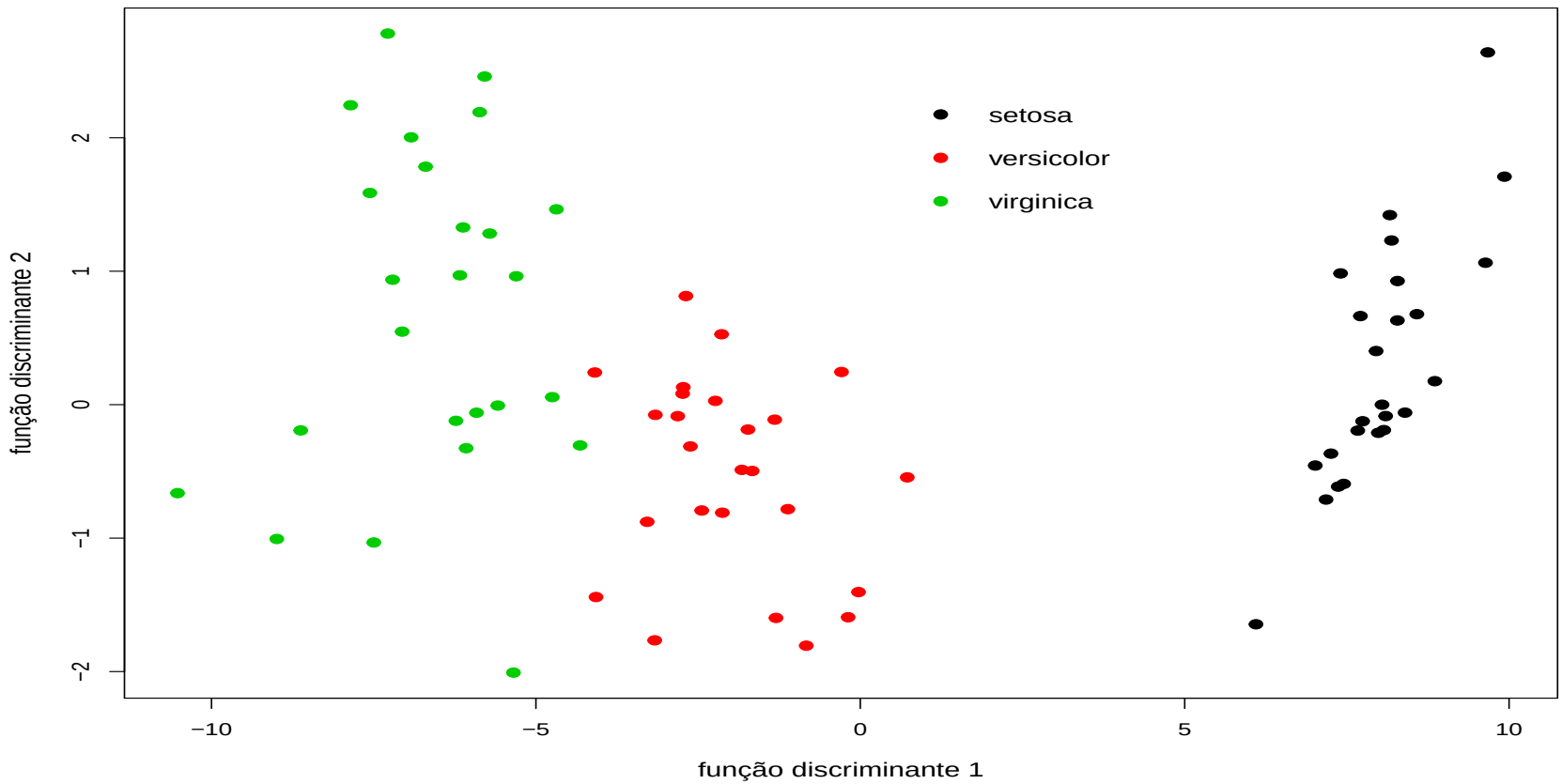
Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Setosa	8,04	0,86	0,74	6,10	8,04	9,93	25
Versicolor	-2,00	1,24	1,53	-4,09	-2,14	0,72	25
Virginica	-6,56	1,44	2,06	-10,52	-6,17	-4,32	25

Medidas resumo

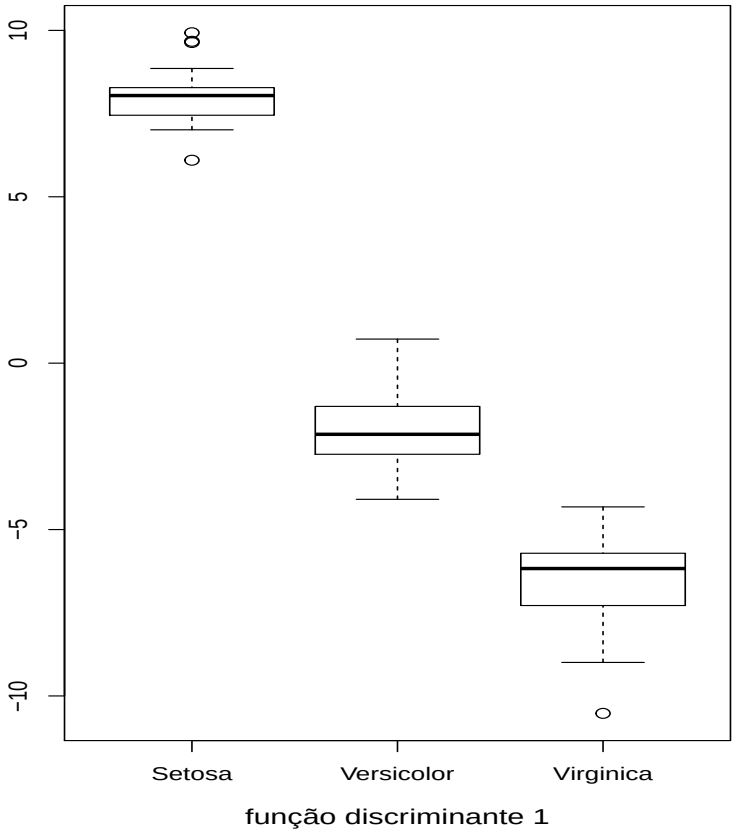
Função discriminante 2

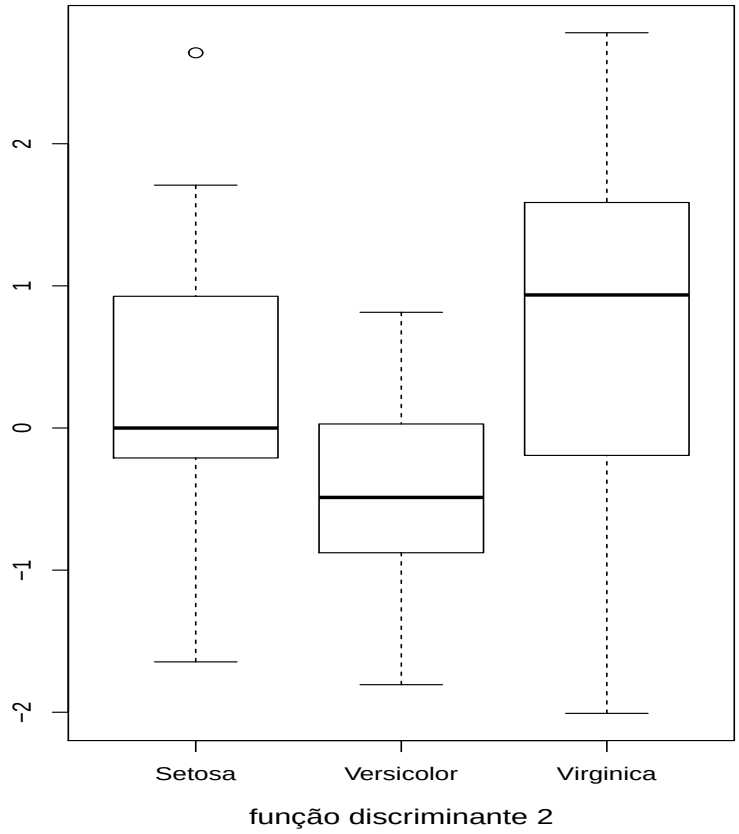
Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Setosa	0,29	0,92	0,84	-1,65	-0,00	2,64	25
Versicolor	-0,52	0,75	0,56	-1,81	-0,49	0,81	25
Virginica	0,67	1,25	1,56	-2,01	0,94	2,78	25

Dispersões entre as funções discriminantes



Ex. 1: boxplots da função discriminante





Ex. 1: densidade estimada da função discriminante