SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

Aula 6a: **Análise de Variância Multivariada (MANOVA)**

Prof. Cibele Russo

cibele@icmc.usp.br

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

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

Comparação de médias em várias populações independentes

Sejam

- $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n_1}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{1j}) = \underset{\sim}{\mu}_1$ para $j = 1, \ldots, n$ e $\text{Var}(\underset{\sim}{X}_{1j}) = \Sigma$,
- $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n_2}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{2j}) = \underset{\sim}{\mu}_2$ para $j = 1, \ldots, n$ e $\text{Var}(\underset{\sim}{X}_{2j}) = \Sigma$

  ...

- $\underset{\sim}{X}_{g1}, \ldots, \underset{\sim}{X}_{gn_g}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{gj}) = \underset{\sim}{\mu}_g$ para $j = 1, \ldots, n_g$ e $\text{Var}(\underset{\sim}{X}_{gj}) = \Sigma$

supondo que a todas as populações são independentes entre si.

Comparação de médias em várias populações independentes

Sejam

- $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n_1}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{1j}) = \underset{\sim}{\mu}_1$ para $j = 1, \ldots, n$ e $Var(\underset{\sim}{X}_{1j}) = \Sigma$,
- $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n_2}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{2j}) = \underset{\sim}{\mu}_2$ para $j = 1, \ldots, n$ e $Var(\underset{\sim}{X}_{2j}) = \Sigma$

...

- $\underset{\sim}{X}_{g1}, \ldots, \underset{\sim}{X}_{gn_g}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{gj}) = \underset{\sim}{\mu}_g$ para $j = 1, \ldots, n_g$ e $Var(\underset{\sim}{X}_{gj}) = \Sigma$

supondo que a todas as populações são independentes entre si.

Comparação de médias em várias populações independentes

Sejam

- $\underset{\sim}{X}_{11}, \ldots, \underset{\sim}{X}_{1n_1}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{1j}) = \underset{\sim}{\mu}_1$ para $j = 1, \ldots, n$ e $Var(\underset{\sim}{X}_{1j}) = \Sigma$,
- $\underset{\sim}{X}_{21}, \ldots, \underset{\sim}{X}_{2n_2}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{2j}) = \underset{\sim}{\mu}_2$ para $j = 1, \ldots, n$ e $Var(\underset{\sim}{X}_{2j}) = \Sigma$

  ...

- $\underset{\sim}{X}_{g1}, \ldots, \underset{\sim}{X}_{gn_g}$ a.a. de uma população com distribuição normal com $E(\underset{\sim}{X}_{gj}) = \underset{\sim}{\mu}_g$ para $j = 1, \ldots, n_g$ e $Var(\underset{\sim}{X}_{gj}) = \Sigma$

supondo que a todas as populações são independentes entre si.

Comparação de médias em várias populações independentes

Deseja-se avaliar as hipóteses

$$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \ldots \underset{\sim}{\mu}_g = \mu \text{ contra}$$
$$H_1 : \text{ pelo menos um } \underset{\sim}{\mu}_i \text{ diferente}$$

Para isso, vamos considerar a reparametrização

$$\underset{\sim}{\mu}_k = \underset{\sim}{\mu} + \underset{\sim}{\tau}_k, \; k = 1, \ldots, g,$$

e então avaliar se

$$H_0 : \underset{\sim}{\tau}_1 = \underset{\sim}{\tau}_2 = \ldots \underset{\sim}{\tau}_g = \underset{\sim}{0} \text{ contra}$$
$$H_1 : \text{ pelo menos um } \underset{\sim}{\tau} \text{ diferente de } \underset{\sim}{0}$$

Comparação de médias em várias populações independentes

Deseja-se avaliar as hipóteses

$$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \ldots \underset{\sim}{\mu}_g = \mu \text{ contra}$$
$$H_1 : \text{ pelo menos um } \underset{\sim}{\mu}_i \text{ diferente}$$

Para isso, vamos considerar a reparametrização

$$\underset{\sim}{\mu}_k = \underset{\sim}{\mu} + \underset{\sim}{\tau}_k, \; k = 1, \ldots, g,$$

e então avaliar se

$$H_0 : \underset{\sim}{\tau}_1 = \underset{\sim}{\tau}_2 = \ldots \underset{\sim}{\tau}_g = \underset{\sim}{0} \text{ contra}$$
$$H_1 : \text{ pelo menos um } \underset{\sim}{\tau} \text{ diferente de } \underset{\sim}{0}$$

Modelo de ANOVA multivariada - MANOVA

Seja o modelo de ANOVA multivariada (MANOVA)

$$\underset{\sim}{X}_{kj} = \underset{\sim}{\mu} + \underset{\sim}{\tau}_k + \underset{\sim}{\epsilon}_{kj},$$

para $j = 1, \ldots, n_k, k = 1, \ldots, g$.

Suposições:

- $\underset{\sim}{\epsilon}_{kj} \overset{ind}{\sim} N_p(\underset{\sim}{0}, \Sigma)$,
- $\underset{\sim}{\mu}$ é a média geral,
- $\underset{\sim}{\tau}_k$ é o efeito do $k-$ ésimo tratamento,
- $\sum_{k=1}^{g} n_k \underset{\sim}{\tau}_k = \underset{\sim}{0}$.

Modelo de ANOVA multivariada - MANOVA

Seja o modelo de ANOVA multivariada (MANOVA)

$$\underset{\sim}{X}_{kj} = \underset{\sim}{\mu} + \underset{\sim}{\tau}_k + \underset{\sim}{\epsilon}_{kj},$$

para $j = 1, \ldots, n_k, k = 1, \ldots, g$.

Suposições:

- $\underset{\sim}{\epsilon}_{kj} \overset{ind}{\sim} N_p(\underset{\sim}{0}, \Sigma)$,
- $\underset{\sim}{\mu}$ é a média geral,
- $\underset{\sim}{\tau}_k$ é o efeito do $k-$ ésimo tratamento,
- $\sum_{k=1}^{g} n_k \underset{\sim}{\tau}_k = \underset{\sim}{0}$.

Modelo de ANOVA multivariada - MANOVA

Considerando os vetores observados, podemos escrever

$$\underbrace{\underset{\sim}{X}_{kj}}_{\text{observação}} = \underbrace{\overline{\underset{\sim}{X}}}_{\text{média geral}} + \underbrace{(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})}_{\substack{\text{efeito estimado} \\ \text{do tratamento } k}} + \underbrace{(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)}_{\text{resíduo}},$$

para $j = 1, \ldots, n_k$ e $k = 1, \ldots, g$.

Modelo de ANOVA multivariada - MANOVA

Como no caso univariado, queremos decompor a variabilidade dos dados em torno da média em variabilidade intra e entre tratamentos. Temos que

$$
\begin{aligned}
&(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}})^\top = \\
&= \left[(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k) + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})\right]\left[(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k) + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})\right]^\top = \\
&= (\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)^\top + (\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})^\top + \\
&\quad + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)^\top + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})^\top
\end{aligned}
$$

Modelo de ANOVA multivariada - MANOVA

Como no caso univariado, queremos decompor a variabilidade dos dados em torno da média em variabilidade intra e entre tratamentos. Temos que

$$
\begin{aligned}
&(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}})^\top = \\
&= \left[(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k) + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})\right]\left[(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k) + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})\right]^\top = \\
&= (\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)^\top + (\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})^\top + \\
&\quad + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)^\top + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})^\top
\end{aligned}
$$

Modelo de ANOVA multivariada - MANOVA

Como no caso univariado, queremos decompor a variabilidade dos dados em torno da média em variabilidade intra e entre tratamentos. Temos que

$$
\begin{aligned}
&(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}})^\top = \\
&= \left[(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k) + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})\right]\left[(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k) + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})\right]^\top = \\
&= (\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)^\top + (\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})^\top + \\
&\quad + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)^\top + (\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})^\top
\end{aligned}
$$

Modelo de ANOVA multivariada - MANOVA

Vamos fazer a soma em j, mas

$$
\begin{aligned}
&\sum_{j=1}^{n_k}\left[(\underset{\sim}{X}_{kj}-\overline{\underset{\sim}{X}}_k)(\overline{\underset{\sim}{X}}_k-\overline{\underset{\sim}{X}})^\top\right]=\\
&\sum_{j=1}^{n_k}\left[(\underset{\sim}{X}_{kj}-\overline{\underset{\sim}{X}}_k)\right](\overline{\underset{\sim}{X}}_k-\overline{\underset{\sim}{X}})^\top=\\
&(\sum_{j=1}^{n_k}\underset{\sim}{X}_{kj}-n_k\overline{\underset{\sim}{X}}_k)(\overline{\underset{\sim}{X}}_k-\overline{\underset{\sim}{X}})^\top=\\
&(\sum_{j=1}^{n_k}\underset{\sim}{X}_{kj}-\sum_{j=1}^{n_k}\underset{\sim}{X}_{kj})(\overline{\underset{\sim}{X}}_k-\overline{\underset{\sim}{X}})^\top=0
\end{aligned}
$$

Idem para

$$\sum_{j=1}^{n_k}(\overline{\underset{\sim}{X}}_k-\overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj}-\overline{\underset{\sim}{X}}_k)^\top=0$$

Modelo de ANOVA multivariada - MANOVA

Então, fazendo a soma em j e em k, ficamos com

$$
\begin{aligned}
&\sum_{k=1}^{g}\sum_{j=1}^{n_k}(\underset{\sim}{X}_{kj}-\overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj}-\overline{\underset{\sim}{X}})^\top = \\
&\quad = \sum_{k=1}^{g} n_k(\overline{\underset{\sim}{X}}_k-\overline{\underset{\sim}{X}})(\overline{\underset{\sim}{X}}_k-\overline{\underset{\sim}{X}})^\top + \sum_{k=1}^{g}\sum_{j=1}^{n_k}(\underset{\sim}{X}_{kj}-\overline{\underset{\sim}{X}}_k)(\underset{\sim}{X}_{kj}-\overline{\underset{\sim}{X}}_k)^\top
\end{aligned}
$$

Ou seja,

$T = B + W$, com

- T: soma de quadrados e produtos cruzados **total**
- B: soma de quadrados e produtos cruzados **entre** tratamentos
- W: soma de quadrados e produtos cruzados **intra** tratamento (**dentro**)

Modelo de ANOVA multivariada - MANOVA

Consideramos as somas de quadrados e produtos cruzados

$$T = \sum_{k=1}^{g}\sum_{j=1}^{n_k}(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}})^\top (\textbf{total})$$

$$B = \sum_{k=1}^{g} n_k(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})(\overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}})^\top (\textbf{entre})$$

$$W = \sum_{k=1}^{g}\sum_{j=1}^{n_k}(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)^\top (\textbf{dentro})$$

Modelo de ANOVA multivariada - MANOVA

Temos

$$T = B + W$$

$$SQT = SQTrat + SQRes$$

Como supomos que todos os grupos têm a mesma variância, podemos considerar como estimativa de Σ:

$$S_{pooled} = W = \sum_{k=1}^{g} \sum_{j=1}^{n_k} (\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)(\underset{\sim}{X}_{kj} - \overline{\underset{\sim}{X}}_k)^\top$$

$$W = (n_1 - 1)S_1 + (n_2 - 1)S_2 + \ldots + (n_g - 1)S_g.$$

Tabela MANOVA

Tabela ANOVA multivariada

Fonte de variação	Somas de quadrados e produtos cruzados	graus de liberdade
Tratamento	B	$g-1$
Resíduo	W	$N-g$
Total	T	$N-1$

em que $N = \sum_{k=1}^{g} n_k$. Assim, para avaliar

$$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \ldots \underset{\sim}{\mu}_g = \mu \text{ contra } H_1 : \text{ pelo menos um } \underset{\sim}{\mu}_i \text{ diferente}$$

ou seja

$$H_0 : \underset{\sim}{\tau}_1 = \underset{\sim}{\tau}_2 = \ldots \underset{\sim}{\tau}_g = \underset{\sim}{0} \text{ contra } H_1 : \text{ pelo menos um } \underset{\sim}{\tau} \text{ diferente de } \underset{\sim}{0}.$$

Tabela MANOVA

Queremos verificar se W é pequeno em relação a T.

Mas como avaliar, se são matrizes?

Como são matrizes, podemos considerar

Lambda de Wilks (variância generalizada):

$$\Lambda^{\star} = \frac{|W|}{|W + B|}.$$

A distribuição de $\Lambda^{\star}$ foi estudada em diferentes casos e existem tabelas com os resultados (ver Tabela 6.3 de Johnson & Wichern).

Tabela MANOVA

Queremos verificar se W é pequeno em relação a T.

Mas como avaliar, se são matrizes?

Como são matrizes, podemos considerar

Lambda de Wilks (variância generalizada):

$$\Lambda^{\star} = \frac{|W|}{|W + B|}.$$

A distribuição de $\Lambda^{\star}$ foi estudada em diferentes casos e existem tabelas com os resultados (ver Tabela 6.3 de Johnson & Wichern).

Lambda de Wilks

Exemplo: Lambda de Wilks

Se $p = 2$ e $g \geq 2$, então

$$\left(\frac{\sum_{k=1}^{g} n_k - g - 1}{g - 1}\right)\left(\frac{1 - \sqrt{\Lambda^\star}}{\sqrt{\Lambda^\star}}\right) \sim F_{2(g-1), 2\sum n_k - g - 1}$$

Bartlett mostrou que se H_0 é verdadeira e $\sum_{k=1}^{g} = N$ é grande, então

$$-\left(n - 1 - \frac{p+g}{2}\right)\log(\Lambda^\star) \sim \chi^2_{p(g-1)}.$$

MANOVA

Exemplo: Uma base de dados com 150 observações e 5 variáveis referentes a crânios egípcios de cinco épocas:

- época: a época à qual o crânio foi atribuído, um fator ordenado com níveis c4000BC, c3300BC, c1850BC, c200BC e cAD150, onde os anos são dados apenas aproximadamente, é claro.
- MB: largura máxima do crânio.
- bh: altura basibregmática do crânio.
- bl: comprimento basialveolar do crânio.
- nh: altura nasal do crânio.

Mais informações em:

http://friendly.github.io/heplots/reference/Skulls.html