MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2020 - IME

Análise Multivariada

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}$$

Já vimos ☺

- Estatísticas Descritivas Multivariadas:
- Distribuição N_p, Distribuições Amostrais, Regiões de Confiança, MANOVA
- Análises Multivariadas Clássicas (n>p, *iid*): CP, AF, CoP, AC, AD, ACC, PLS
- Análises Multivariadas Esparsas (n<<p, *iid*): CP, AD, ACC

Lista 5

- **"Componentes Principais" em Observações Dependentes (dados de famílias)**
- "Componentes Principais" em Dados Heterogêneos
- Aprendizado de Estruturas
  - Modelos de Grafos Probabilísticos
  - Modelos de Equações Estruturais
  - Fatoração da Distribuição Conjunta

Apoio à Lista 5

Estrutura dos Dados: Observações Independentes

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$Y_i \in \Re^p$

- Caso 1: Amostra Aleatória Simples de n-Vetores em $\Re^p$

$$Y_{n\times p} \in \Re^{n\times p}; \quad Y_i \in \Re^p, \quad Y_i = \mu + e_i; \quad e_i \overset{iid}{\sim} (0;\Sigma)$$

Estimar Σ (ou R) e obter os Vetores reducionistas (=**CP**)

$$\hat{\Sigma} = S = \frac{1}{n-1}\sum\left(Y_i - \bar{Y}\right)\left(Y_i - \bar{Y}\right)'$$

$$CP \Rightarrow \max_a \frac{a'\hat{\Sigma}\, a}{a'a}; \quad a \in \mathbb{R}^p, \; a'a = 1$$

Direção com máxima variação Total

Estrutura dos Dados: Agrupados e Observações Independentes

Grupos	Unidade Amostral	Y_1	Y_2	...	Y_p
1	1	Y_{111}	Y_{112}		Y_{11p}
1	2	Y_{121}	Y_{122}		Y_{12p}
...	...				
1	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1p}
...					
G	1	Y_{G11}	Y_{G12}		Y_{G1p}
G	2	Y_{G21}	Y_{G22}		Y_{G2p}
...	...				
G	n_G	$Y_G n_{G1}$	$Y_G n_{G2}$		$Y_G n_{Gp}$

Independência entre e dentro dos grupos

$$n = \sum_{g=1}^{G} n_g$$

$$Y_{ig} \in \Re^p$$

- Caso 2: Amostra Aleatória Simples de G grupos

$$\Rightarrow Y_{ig} = \mu_g + e_g; \quad e_g \sim N_p\left(0; \Sigma_W\right) \quad \Rightarrow Y_{ig} \overset{iid}{\sim} \left(\mu_g; \Sigma_W\right)$$

Fontes de variação: $SS_T = SS_B + SS_W$,
Estimar Σ_W e obter os Vetores reducionistas (=**AD**)

$$CP \Rightarrow \max_a \frac{a' SS_B a}{a' \hat{\Sigma}_W a}; \quad a \in \mathbb{R}^p, \; a' \hat{\Sigma}_W a = 1$$

Direção com máxima discriminação entre os grupos

Tabela MANOVA

$$\Rightarrow Y_{ig} = \mu_g + e_g; \quad e_g \sim N_p(0;\Sigma_W) \quad \Rightarrow Y_{ig} \overset{iid}{\sim} (\mu_g;\Sigma_W)$$

Tabela de MANOVA:

F.V.	g.l.	Matriz de SQPC
Trat	G-1	$\sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)' = SS_B$
Resíduo	n-G	$\sum_{g=1}^{G}\sum_{i=1}^{n_g} \left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)' = SS_W$
TOTAL	n-1	$SS_T = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \left(Y_{gi} - \bar{Y}\right)\left(Y_{gi} - \bar{Y}\right)'$

Sob $H_0 : \mu_g = \mu, \quad g = 1,\ldots,G$

$$\boxed{E\left(\frac{SS_W}{n-G}\right) = \Sigma_W;}$$

$$\hat{\Sigma}_W = \frac{SS_W}{n-G}$$ Estimador MANOVA

Estrutura dos Dados: Agrupados e Dependência Uniforme

Grupos	Unidade Amostral	Y_1	Y_2	...	Y_p
1	1	Y_{111}	Y_{112}		Y_{11p}
1	2	Y_{121}	Y_{122}		Y_{12p}
...	...				
1	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1p}
...					
G	1	Y_{G11}	Y_{G12}		Y_{G1p}
G	2	Y_{G21}	Y_{G22}		Y_{G2p}
...	...				
G	n_G	Y_Gn_{G1}	Y_Gn_{G2}		Y_Gn_{Gp}

Dependência uniforme dentro dos grupos e independência entre grupos

$$n = \sum_{g=1}^{G} n_g$$

$$Y_{ig} \in \Re^p$$

- Caso 3: Amostra Aleatória Simples de G grupos

$$\Rightarrow Y_{ig} = \mu + \tau_g + e_g; \quad \tau_g \sim N_p(0; \Sigma_B); \quad e_g \sim N_p(0; \Sigma_W)$$

Estimar Σ_B e Σ_W e obter os Vetores reducionistas (**CPH**)

$$CPH \Rightarrow \max_a \frac{a'\hat{\Sigma}_B a}{a'\hat{\Sigma}_W a}; \quad a \in \mathbb{R}^p, \; a'\hat{\Sigma}_W a = 1$$

Direção com máxima discriminação entre os grupos e mínima dentro de grupos

Tabela MANOVA

$$\Rightarrow Y_{ig} = \mu + \tau_g + e_g; \quad \tau_g \sim N_p\left(0;\Sigma_B\right); \quad e_g \sim N_p\left(0;\Sigma_W\right)$$

Tabela de MANOVA:

F.V.	g.l.	Matriz de SQPC
Trat	G-1	$\sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)' = SS_B$
Resíduo	n-G	$\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)' = SS_W$
TOTAL	n-1	$SS_T = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi} - \bar{Y}\right)\left(Y_{gi} - \bar{Y}\right)'$

Sob $H_0 : \mu_g = \mu, \quad g = 1,\ldots,G$

$$E\left(\frac{SS_W}{n-G}\right) = \Sigma_W; \quad E\left(\frac{SS_B}{G-1}\right) = \Sigma_W + n_0\Sigma_B$$

$$n_0 = \frac{n - \left(\sum_g n_g^2 / n\right)}{G-1}$$

Componentes de Covariâncias em $\Re^{pxp}$ Sob Correlação Uniforme entre Observações

Estimadores MANOVA dos componentes de covariância

$$\Rightarrow \hat{\Sigma}_W = \frac{SS_W}{n-G}; \qquad \hat{\Sigma}_B = n_0^{-1}\left\{\frac{SS_B}{G-1} - \frac{SS_W}{n-G}\right\}$$

$$\Rightarrow \hat{\Sigma} = \hat{\Sigma}_W + \hat{\Sigma}_B = n_0^{-1}\left\{\frac{SS_B}{G-1} + \frac{(n_0-1)SS_W}{n-G}\right\}$$

$$n_0 = \frac{n - \left(\sum_g n_g^2 / n\right)}{G-1}$$

Dados Agrupados e Dependência Uniforme

$$Y_{ig(p\times 1)} = \mu + \tau_g + e_{ig};$$

$$\tau_g \perp e_{ig}; \quad E(\tau_f) = E(e_{ig}) = 0; \quad E(\tau_g \tau_g') = \Sigma_B \quad E(e_{ig} e_{ig}') = \Sigma_W$$

$$\Rightarrow Cov(Y_{ig}) = \Sigma_{p\times p} = \Sigma_B + \Sigma_W$$

$$\Rightarrow Cov(Y_g)_{(n_g p\times n_g p)} = \Omega_g = \boxed{(1_g 1_g')_{(n_g\times n_g)}} \otimes \Sigma_B + I_{g(n_g\times n_g)} \otimes \Sigma_W$$

$$= \begin{pmatrix} \Sigma_B & \Sigma_B & \cdots & \Sigma_B \\ \Sigma_B & \Sigma_B & \cdots & \Sigma_B \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma_B & \Sigma_B & \cdots & \Sigma_B \end{pmatrix} + \begin{pmatrix} \Sigma_W & 0 & \cdots & 0 \\ 0 & \Sigma_W & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \Sigma_W \end{pmatrix}$$

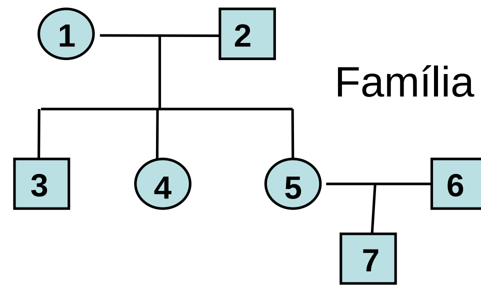
$$\Rightarrow Cov(Y)_{(np\times np)} = \Omega = I_G \otimes \Omega_g$$

$$= \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \Omega_G \end{pmatrix}$$

Estrutura dos Dados: Agrupados e Dependência Familiar

Grupos	Unidade Amostral	Y_1	Y_2	...	Y_p
1	1	Y_{111}	Y_{112}		Y_{11p}
1	2	Y_{121}	Y_{122}		Y_{12p}
...	...				
1	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1p}
...					
G	1	Y_{G11}	Y_{G12}		Y_{G1p}
G	2	Y_{G21}	Y_{G22}		Y_{G2p}
...	...				
G	n_G	$Y_G n_{G1}$	$Y_G n_{G2}$		$Y_G n_{Gp}$

Dependência familiar (grau de parentesco) dentro dos grupos e independência entre grupos



Família g

- Caso 4: Amostra Aleatória Simples de G grupos (Famílias)

$$\Rightarrow Y_{ig} = \mu + \tau_g + e_g; \quad \tau_g \sim N_p\left(0;\Sigma_B\right); \quad e_g \sim N_p\left(0;\Sigma_W\right)$$

Estimar Σ_B e Σ_W e obter os Vetores reducionistas (**CPH**)

$$CPH \Rightarrow \max_a \frac{a'\hat{\Sigma}_B a}{a'\hat{\Sigma}_W a}; \quad a \in \mathbb{R}^p, \; a'\hat{\Sigma}_W a = 1$$

Direção com máxima discriminação entre os grupos e mínima dentro de grupos

Tabela MANOVA

$$\Rightarrow Y_{ig} = \mu + \tau_g + e_g; \quad \tau_g \sim N_p\left(0;\Sigma_B\right); \quad e_g \sim N_p\left(0;\Sigma_W\right)$$

Tabela de MANOVA:

F.V.	g.l.	Matriz de SQPC
Trat	G-1	$\sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)' = SS_B$
Resíduo	n-G	$\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)' = SS_W$
TOTAL	n-1	$SS_T = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi} - \bar{Y}\right)\left(Y_{gi} - \bar{Y}\right)'$

Sob $H_0 : \mu_g = \mu, \quad g = 1,\ldots,G$

$$E\left(\frac{SS_W}{n-G}\right) = \Sigma_W; \quad E\left(\frac{SS_B}{G-1}\right) = \Sigma_W + n_0\Sigma_B$$

$$n_0 = \frac{n - \left(\sum_g n_g^2 / n\right)}{G-1}$$

Dados Agrupados e Dependência Familiar

$$\hat{\Sigma}_B = \frac{SS_B/(G-1) - SS_W/(n-G)}{(\tau_c - \tau_b/n)/(G-1) - (\tau_a - \tau_c)/(n-G)}$$

(Oualkacha et al., 2012)

$$\hat{\Sigma}_W = \frac{1}{(n-G)} SS_W - \frac{(\tau_a - \tau_c)}{(n-G)} \hat{\Sigma}_B$$

$$n = \sum_{g=1}^{G} n_g, \quad \tau_a = \sum_{g=1}^{G} \tau_{a_g}, \quad \tau_b = \sum_{g=1}^{G} \tau_{b_g}, \quad \tau_c = \sum_{g=1}^{G} \frac{1}{n_g} \tau_{b_g}$$

$$\tau_{a_g} = 2Trace\left[\Psi_g\right], \quad \tau_{b_g} = 2\sum_{j=1}^{n_g}\sum_{k=1}^{n_g}\left(\Psi_g\right)_{jk}$$

Dados Agrupados e Dependência Familiar

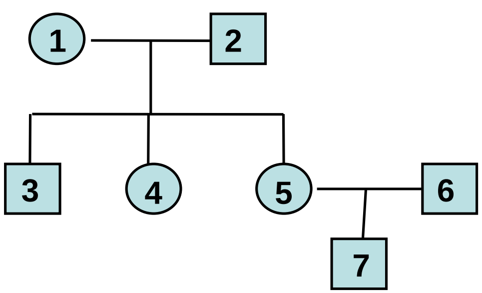
$$Y_{ig(p\times 1)} = \mu + \tau_g + e_{ig};$$

$$\tau_g \perp e_{ig}; \quad E\left(\tau_f\right) = E\left(e_{ig}\right) = 0; \quad E\left(\tau_g \tau_g'\right) = \Sigma_B \quad E\left(e_{ig} e_{ig}'\right) = \Sigma_W$$

$$\Rightarrow Cov\left(Y_{ig}\right) = \Sigma_{p\times p} = \Sigma_B + \Sigma_W$$

$$\Rightarrow Cov\left(Y_g\right)_{n_g p\times n_g p} = \Omega_g = \boxed{\Psi_{g(n_g\times n_g)}} \otimes \Sigma_B + I_{g(n_g\times n_g)} \otimes \Sigma_W$$

Matriz de parentesco



$\boxed{\Psi_g} =$

	1	2	3	4	5	6	7
1	1	0	½	½	½	0	¼
2	0	1	½	½	½	0	¼
3	½	½	1	½	½	0	¼
4	½	½	½	1	½	0	¼
5	½	½	½	½	1	0	½
6	0	0	0	0	0	1	½
7	¼	¼	¼	¼	½	½	1