MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2020 - IME

Análise Multivariada

Já vimos ☺

- Análise Descritiva Multivariada
- Elipsóides de Dispersão e de Confiança, MANOVA
- **Metodologias Clássicas** de redução de dimensionalidade: $Y_{n\times p}$; $\Re^p \rightarrow \Re^m$
  - ✓ ACP *(m≤min(n,p))*, ACoP *(m≤min(n,p))*, AC *(m≤min(I-1,J-1))*, AF *(m≤min(n,p))*
  - ✓ AG
  - ✓ AD *(m≤min(n,p,G-1))*, ACC *(m≤min(n,p,q))*
  - ✓ Soluções Duais ($\Re^{nxp}$, $\Re^{pxp}$, $\Re^{nxn}$), Representações Biplot
  - ✓ PCR (Regressão via CP), PLS (Regressão via MQ Parciais)
- Integração de Bancos de Dados - Diagrama de Caminhos (Grafos)

Var. quantitativas
n > p
Obs iid

Casos mais Gerais:

- ✓ n << p (*big p*): ACP, AD e ACC (Soluções Regularizadas e Penalizadas via Fatoração de Matrizes e via Modelos de Regressão)
- ✓ Observações "dependentes": CP em dados agrupados em famílias (soluções via modelos MANOVA de efeitos fixos e aleatórios)

⇒ **Dados Multivariados Heterogêneos (variáveis em diferentes escalas)**

Matriz de Dados Multivariados

	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}

Variáveis em diferentes escalas: quantitativas (contínuas e discretas) e qualitativas (nominais e ordinais)

Como realizar a redução de dimensionalidade em dados heterogêneos?

Motivação

ID	Idade	Gênero	Renda	Escolaridade	Número de Filhos	
1	49	F	8000,00	Superior	NA	....
2	23	F	NA	Médio	0	....
3	38	M	1200,50	NA	2	....
⋮	⋮	⋮	⋮	⋮	⋮	

$n>p$
$n<<p$

Discreta

Binária

Contínua

Ordinal

Contínua

Componentes Principais - n>p

$Y_{i\ p\times 1} \overset{iid}{\sim} (\mu\ ;\Sigma\ )$

Suposições: AASn de uma única população com matriz de cov Σ, obs independentes, n>p

Redução de dimensionalidade:

$\Re^p \rightarrow \Re^m; \quad m \leq \min(n, p)$

$$Y_{n\times p} V_{p\times r} = U_{n\times n} \Lambda_r^{1/2}$$

Equivalência entre os Componentes Principais e as Cordenadas Principais

(Gabriel, 2008)

$$Y_{n\times p} = \underbrace{U_{n\times n}\Lambda_r^{1/2}}_{A}\underbrace{V_{p\times r}}_{B} \Rightarrow \min_{a,b} \sum_{i=1}^{n}\sum_{j=1}^{p} (Y_{ij} - a_i b_j)^2$$

Componentes Principais Regularizados – n<<p

$Y_{i\ p\times 1} \overset{iid}{\sim} (\mu\ ;\Sigma\ )$

Suposições: AASn de uma única população com matriz de cov Σ, obs independentes

Redução de dimensionalidade:

$\Re^p \rightarrow \Re^m$; $m \leq \min(n, p)$

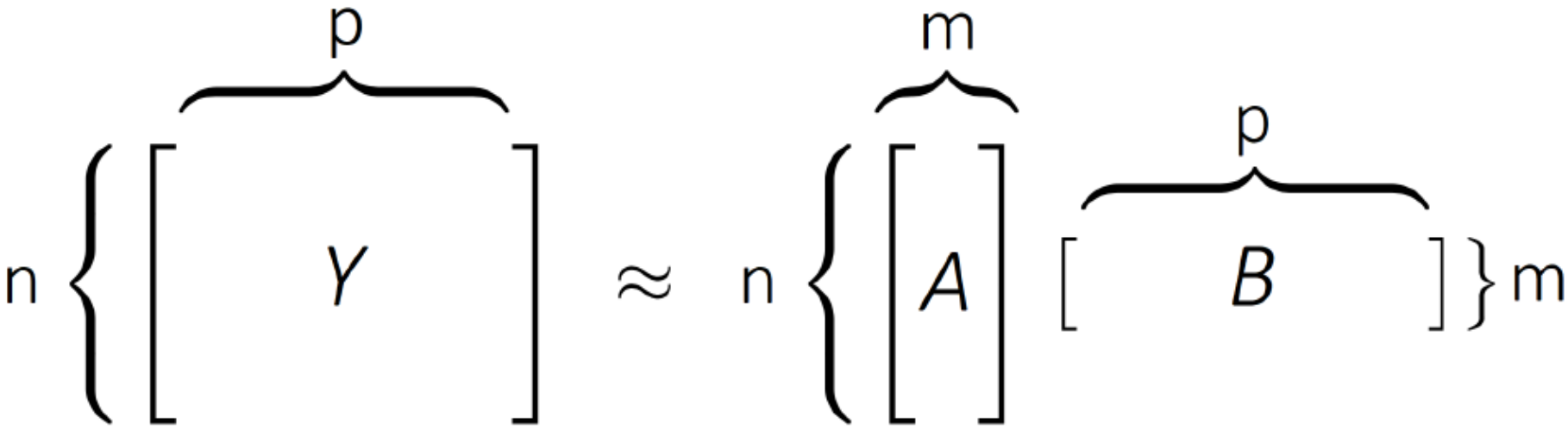
$$\min_{a,b} \sum_{i=1}^{n}\sum_{j=1}^{p}\left(Y_{ij} - a_i b_j\right)^2 + \gamma \sum_{j=1}^{p} b_j^2$$

$$\Rightarrow \quad \min_{a,b} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{ij} - a_i b_j\right)^2 + \gamma \sum_{j=1}^{m} \tilde{r}(b_j)$$

Diferentes funções de regularização dependendo do objetivo

Parâmetro de regularização

Generalized Low Rank Models (GLRM)



- $Y = (Y_{ij}), Y_{ij} \in \mathcal{F}_j$: valor da j-ésima variável do i-ésimo indivíduo
- $A \in \Re^{n \times m}$
- $B \in \Re^{m \times p}$
- $Y_{ij} \approx a_i b_j$

Generalized Low Rank Models (GLRM)

(Udell et al. 2016)

$Y_{ij} \in \mathcal{F}_j$

- $\mathcal{F}_j = \{V, F\}$
- $\mathcal{F}_j = 1,2,3,4, \ldots$
- $\mathcal{F}_j = \{um\ pouco, médio, muito\}$
- $\mathcal{F}_j = \{(a, b) : a \in \Re\}$

Minimização Alternada

Função Perda L_{ij}: $\Re \times \mathcal{F}_j \to \Re$

$$\min_{a,b} \sum_{i=1}^{n} \sum_{j=1}^{p} L_{ij}\left(Y_{ij}, a_i b_j\right) + \gamma_a \sum_{i=1}^{n} r(a_i) + \gamma_b \sum_{j=1}^{p} \tilde{r}(b_j)$$

$A \in \Re^{n \times m}$

$B \in \Re^{m \times p}$

Parâmetro de regularização da matriz A

Parâmetro de regularização da matriz B

Ajusta o problema de otimização apenas nas observações observadas !

(Lida com dados faltantes)

$$\min_{a,b} \sum_{(i,j) \in \Omega} L_{ij}\left(Y_{ij}, a_i b_j\right) + \gamma_a \sum_{i=1}^{n} r(a_i) + \gamma_b \sum_{j=1}^{p} \tilde{r}(b_j)$$

$$\Rightarrow \quad \hat{Y}_{ij} = \underset{u}{argmin}\, L_{ij}(a_i b_j, u)$$

Generalized Low Rank Models (GLRM)

(Udell et al. 2016)

Regularizadores

Estrutura imposta	$r(a)$	$\tilde{r}(b)$
Pequena	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
Esparsa	$\|\|a\|\|_1$	$\|\|b\|\|_1$
Não-negativa	$\mathbb{I}(a \geq 0)$	$\mathbb{I}(b \geq 0)$
Cluster	$\mathbb{I}(card(a) = 1)$	0

Funções perda para cada tipo de variáveis

Tipo de dados	Perda	$L(u, v)$
Real	Quadrática	$(u - v)^2$
	Absoluta	$\|u - v\|$
	Huber	$huber(u - v)$
Binário	Hinge	$(1 - uv)_+$
	Logística	$log(1 + exp(-uv))$
Inteiro	Poisson	$exp(u) - uv + vlog(v) - v$
Ordinal	Hinge Ordinal	$\sum_{v'=1}^{v-1}(1 - u + v')_+ + \sum_{v'=v+1}^{d}(1 + u - v')_+$
Categórico	Um-vs-Todos	$(1 - u_v)_+ + \sum_{v' \neq v}(1 + u_{v'})_+$

Generalized Low Rank Models (GLRM)

(Udell et al. 2016)

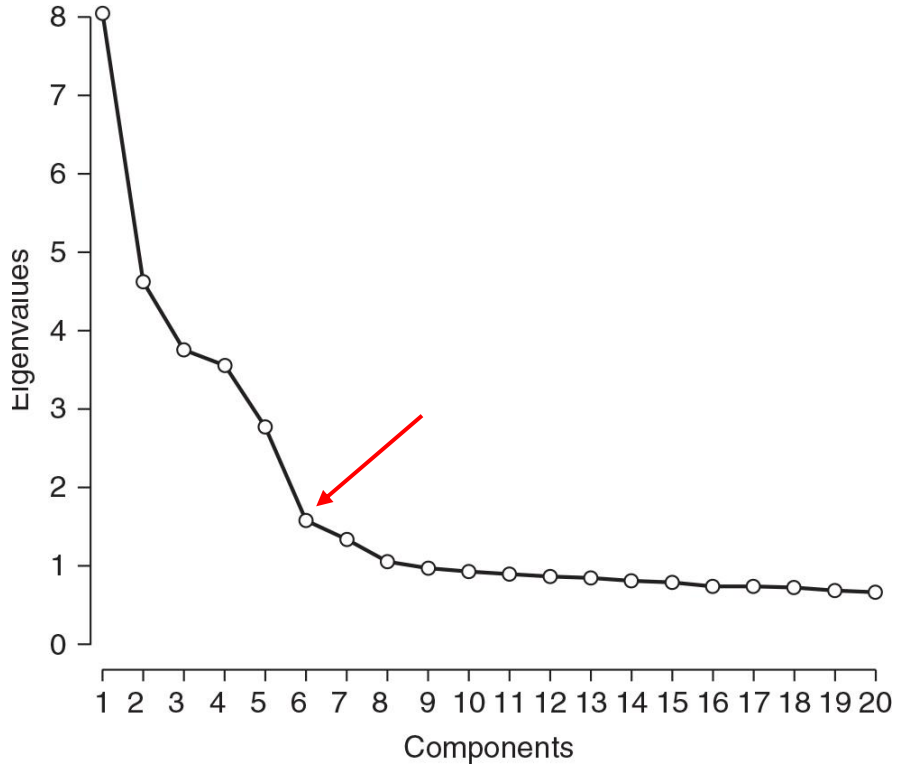
- Ao combinar diferentes Perdas e regularizadores é possível recuperar modelos conhecidos

Modelo	$L_j(u,v)$	$r(a)$	$\tilde{r}(b)$
CP	$(u-v)^2$	0	0
CPs Regularizados	$(u-v)^2$	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
NNMF	$(u-v)^2$	$\mathbb{I}(a \geq 0)$	$\mathbb{I}(b \geq 0)$
CP esparso	$(u-v)^2$	$\|\|a\|\|_1$	$\|\|b\|\|_1$
CP Robusto	$\|u-v\|$	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
CP Logístico	$log(1+esp(-vu))$	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
CP Binário	$(1-vu)_+$	$\|\|a\|\|_2^2$	$\|\|b\|\|_2^2$
K-means	$(u-v)^2$	$\mathbb{I}(card(a)=1)$	0

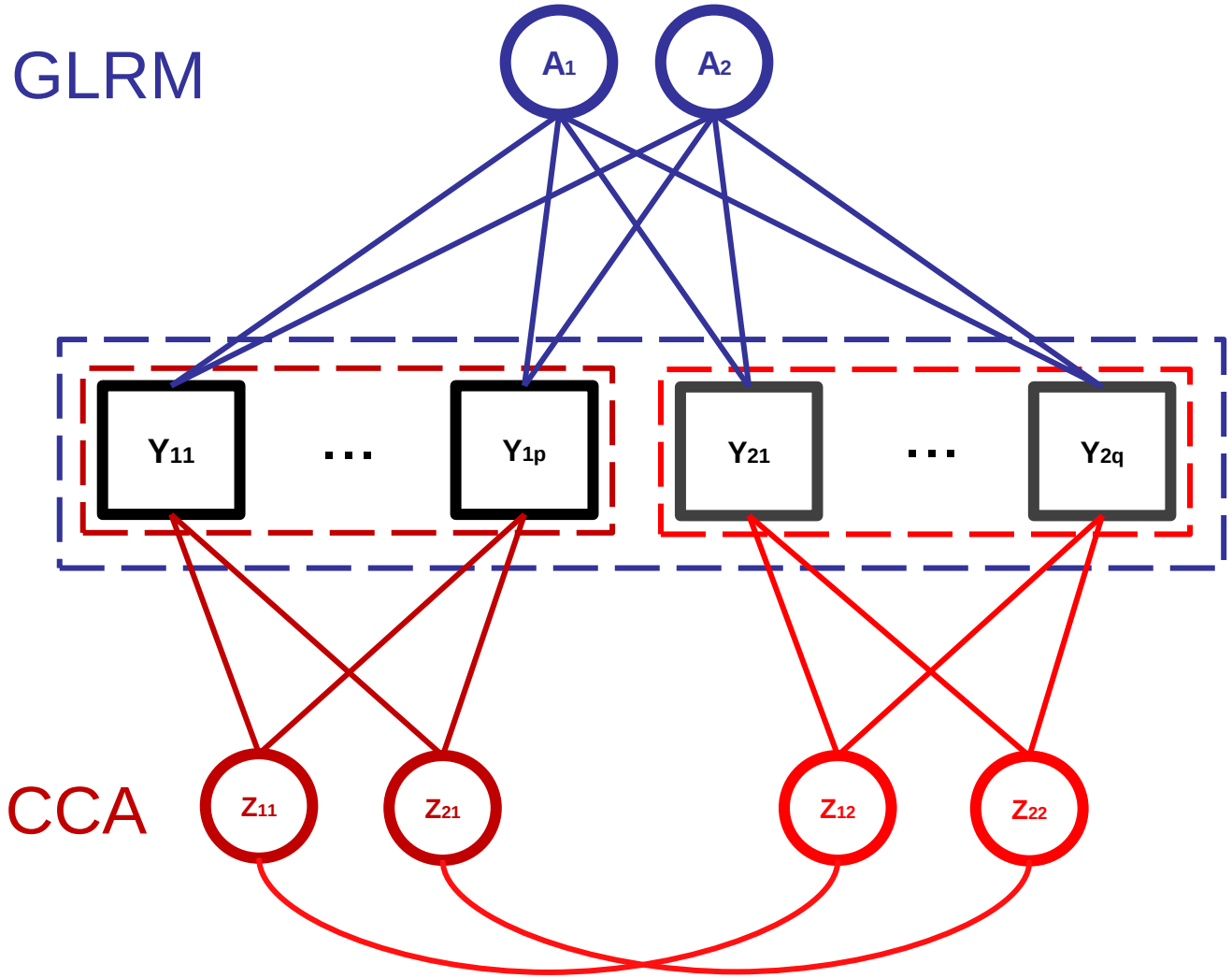
Generalized Low Rank Models (GLRM)- Escolha de parâmetros

(Udell et al. 2016)

- Os parâmetros devem ser escolhidos inicialmente;
- GLRM com mais componentes se ajustam melhor a dados com ruído, mas podem superestimar os dados;
- Modelos com regularizadores apresentam métricas de ajuste menores mas lida melhor com dados não observados.



Redução de Dimensionalidade



Exemplo – Dados de Marketing

ID	Age	Gender	OwnHome	Married	Location	Salary	Children	History	AmountSpent
1	Old	Female	Own	Single	Far	47500	0	High	755
2	Middle	Male	Rent	Single	Close	63600	0	High	1318
3	Young	Female	Rent	Single	Close	13500	0	Low	296
4	Middle	Male	Own	Married	Close	85600	1	High	2436
5	Middle	Female	Own	Single	Close	68400	0	High	1304
6	Young	Male	Own	Married	Close	30400	0	Low	495
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
995	Young	Male	Rent	Single	Close	17600	0	<NA>	273
996	Young	Female	Rent	Single	Close	19400	1	<NA>	384
997	Middle	Male	Rent	Single	Far	40500	1	<NA>	1073
998	Old	Male	Own	Single	Close	44800	0	Medium	1417
999	Middle	Male	Own	Married	Close	79000	2	Medium	671
1000	Young	Male	Rent	Married	Close	53600	1	Medium	973