

Análise multivariada

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

28 de Agosto de 2018

Conteúdo programático

- Introdução à análise multivariada
- Álgebra matricial
- Análise de componentes principais
- Análise de agrupamento (*cluster analysis*)
- Análise fatorial*
- Análise discriminante*

Bibliografia

FERREIRA, D. F. **Estatística multivariada**. Lavras, MG: Ed. UFLA, 2008.

HAIR, J. F. et al. **Análise multivariada de dados**, 6.ed. Bookman, 2009.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: UFMG, 2007.

PYTHON. **The Python programming language**. Disponível em: github.com/python/cpython Acesso em: 26 fev. 2018.

Avaliação

- Prova 1: 10/05/2018 - peso 30%
- Prova 2: 05/07/2018 - peso 30%
- Trabalho final - peso 30%
- Exercícios práticos - peso 10%
- Prova especial: 19/07/2018

Dados multivariados e análise multivariada

Introdução

- Dados multivariados: pesquisador armazena os valores de várias v.a.s de várias unidades (sujeitos, indivíduos, objetos). Cada unidade é uma observação multidimensional
- A representação dos dados se dá como em planilhas
- p variáveis (colunas) medidas em n elementos (linhas)

Introdução

- Dados multivariados: pesquisador armazena os valores de várias v.a.s de várias unidades (sujeitos, indivíduos, objetos). Cada unidade é uma observação multidimensional
- A representação dos dados se dá como em planilhas
- p variáveis (colunas) medidas em n elementos (linhas)

unidade	variável 1	variável 2	...	variável p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
\vdots	\vdots	\vdots	\vdots	\vdots
n	x_{n1}	x_{n2}	...	x_{np}

Exemplos de dados multivariados

Ex.1: Notas de provas de alunos em diferentes disciplinas

aluno	matemática	inglês	história	geografia	química	física
1	60	70	75	58	53	42
2	80	65	66	75	70	76
3	53	60	50	48	45	43
4	85	79	71	77	68	79
5	45	80	80	84	44	46

Ex.1: Notas de provas de alunos em diferentes disciplinas

aluno	matemática	inglês	história	geografia	química	física
1	60	70	75	58	53	42
2	80	65	66	75	70	76
3	53	60	50	48	45	43
4	85	79	71	77	68	79
5	45	80	80	84	44	46

- Neste caso, $n = 5$ e $p = 6$

Ex.2: Variáveis armazenadas por psicólogos sobre seus pacientes

indivíduo	sexo	idade	QI	depressão	saúde
1	M	21	120	S	MB
2	M	60	92	S	B
3	M	22	135	N	M
4	M	86	150	N	MR
5	F	16	130	S	B
6	F	22	84	N	M
7	F	80	70	N	B

Ex.2: Variáveis armazenadas por psicólogos sobre seus pacientes

indivíduo	sexo	idade	QI	depressão	saúde
1	M	21	120	S	MB
2	M	60	92	S	B
3	M	22	135	N	M
4	M	86	150	N	MR
5	F	16	130	S	B
6	F	22	84	N	M
7	F	80	70	N	B

- Neste caso, $n = 7$ e $p = 5$

Ex.3: Medidas de tórax, cintura e quadril (pol)

tórax	cintura	quadril	gênero	tórax	cintura	quadril	gênero
34	30	32	M	36	24	35	F
37	32	37	M	34	24	37	F
38	30	36	M	34	24	37	F
36	33	39	M	33	22	34	F
38	29	33	M	36	26	38	F
43	32	38	M	37	26	37	F
40	33	42	M	34	25	38	F
38	30	40	M	36	26	37	F
40	30	37	M	38	28	40	F
41	32	39	M	35	23	35	F

Ex.3: Medidas de tórax, cintura e quadril (pol)

tórax	cintura	quadril	gênero	tórax	cintura	quadril	gênero
34	30	32	M	36	24	35	F
37	32	37	M	34	24	37	F
38	30	36	M	34	24	37	F
36	33	39	M	33	22	34	F
38	29	33	M	36	26	38	F
43	32	38	M	37	26	37	F
40	33	42	M	34	25	38	F
38	30	40	M	36	26	37	F
40	30	37	M	38	28	40	F
41	32	39	M	35	23	35	F

- Obs.: 1 polegada = 2,54 cm
- Neste caso, $n = 20$ e $p = 4$

Análise multivariada

- Análise simultânea de um conjunto de variáveis

Análise multivariada

- Análise simultânea de um conjunto de variáveis
- As variáveis são, geralmente, correlacionadas entre si
 - temos todas as medidas em cada unidade, indivíduo

Análise multivariada

- Análise simultânea de um conjunto de variáveis
- As variáveis são, geralmente, correlacionadas entre si
 - temos todas as medidas em cada unidade, indivíduo
- Se cada variável for analisada isoladamente, a estrutura dos dados pode não ser percebida
 - padrões podem não aparecer

Tipos de técnicas multivariadas

Exploratórias (foco aqui)

- apelo prático
- independe do conhecimento da distribuição de probabilidade dos dados
- detecção de padrões nos dados
- uso de gráficos para visualização

Inferência

- estimação de parâmetros
- testes de hipóteses
- foco: além dos dados, usar a amostra para realizar inferência sobre a população
- distribuição normal multivariada

Tipos de técnicas multivariadas

Exploratórias (foco aqui)

- apelo prático
- independe do conhecimento da distribuição de probabilidade dos dados
- detecção de padrões nos dados
- uso de gráficos para visualização

Ex.: An. de componentes principais, an. fatorial exploratória, an. de correlação canônica etc.

Inferência

- estimação de parâmetros
- testes de hipóteses
- foco: além dos dados, usar a amostra para realizar inferência sobre a população
- distribuição normal multivariada

Ex.: regressão multivariada, testes de hipóteses sobre médias e correlação, MANOVA etc.

Níveis de mensuração

Níveis de mensuração (tipos de variáveis)

- Qualitativo (não métrico):
 - nominal
 - ordinal
- Quantitativo (métrico):
 - intervalar
 - razão

Níveis de mensuração (qualitativos)

- Nominal: variáveis categóricas não numeradas.
Ex.: sexo, cor do cabelo, S/N.

Níveis de mensuração (qualitativos)

- Nominal: variáveis categóricas não numeradas.
Ex.: sexo, cor do cabelo, S/N.
- Ordinal: há ordem mas não implica igual distância entre pontos na escala. Ex.: classe social, nível de saúde (péssimo a ótimo), nível educacional (não escolarizado, fundamental, médio, superior).

Níveis de mensuração (quantitativos)

- Intervalar: há diferenças iguais entre pontos na escala, mas a posição do 0 é arbitrária. Ex.: temperatura medida em $^{\circ}\text{C}$ ou F , QI.

Níveis de mensuração (quantitativos)

- Intervalar: há diferenças iguais entre pontos na escala, mas a posição do 0 é arbitrária. Ex.: temperatura medida em $^{\circ}C$ ou F , QI.
- Razão: mais alto nível, em que é possível investigar as magnitudes relativas e as diferenças entre os pontos. O 0 é fixo. Ex.: temperatura em K , idade (ou qualquer outra contagem de tempo), peso, altura, dinheiro.

Técnicas e exemplos de aplicação

1 - Análise de componentes principais (ACP) e análise fatorial (AF)

- Analisam inter-relações entre um grande número de variáveis
- Objetivam encontrar um meio de condensar a informação contida em várias variáveis em um conjunto menor delas com perda mínima de informação

1 - Análise de componentes principais (ACP) e análise fatorial (AF)

- Analisam inter-relações entre um grande número de variáveis
- Objetivam encontrar um meio de condensar a informação contida em várias variáveis em um conjunto menor delas com perda mínima de informação

Ex.: Entender relações entre avaliações de clientes de um restaurante.

- Clientes fazem avaliação sobre 6 variáveis: sabor da comida, temperatura da comida, se a comida é fresca, tempo de espera, limpeza do estabelecimento, atendimento
- O analista quer combinar as variáveis em um conjunto menor
- Descobre-se que as 3 primeiras formam fator de qualidade da comida e as 3 últimas formam fator de qualidade do serviço

2 - Análise de agrupamento - *cluster analysis* - AA

- Técnica analítica para obter grupos de indivíduos ou objetos
- Objetiva classificar uma amostra de indivíduos/objetos em um número menor de grupos com base em suas similaridades
- Grupos não são pré-definidos, usa-se AA para identificar grupos

Ex.: *Pioneer Petroleum* - identificação dos clientes

Tabela: Agrupamento dos clientes de posto de gasolina.

Grupo	Características
I (16%)	Homens de meia idade, alta renda, compram na loja de conveniências, lavam o carro no posto, usam gasolina <i>premium</i>
II (16%)	Homens e mulheres, renda média-alta, leais a marca e posto, usam gasolina <i>premium</i>
III (27%)	Homens e mulheres em ascensão, metade com menos de 25 anos, dirigem muito e comem na loja do posto
IV (21%)	Donas de casa que transportam seus filhos, usam qualquer posto
V (20%)	Não são leais a marca ou posto e raramente compram gasolina <i>premium</i> , orçamento apertado

3 - Análise discriminante (AD)

- Variável dependente é dicotômica (M/F, por ex.) ou policotômica (B/M/A) e não métrica
- Variáveis independentes métricas
- Amostra total pode ser dividida em grupos (baseados na v.d.)
- Objetiva prever a probabilidade que um indivíduo/objeto pertencerá a um grupo baseando-se em variáveis independentes

3 - Análise discriminante (AD)

- Variável dependente é dicotômica (M/F, por ex.) ou policotômica (B/M/A) e não métrica
- Variáveis independentes métricas
- Amostra total pode ser dividida em grupos (baseados na v.d.)
- Objetiva prever a probabilidade que um indivíduo/objeto pertencerá a um grupo baseando-se em variáveis independentes

Ex. (Inadimplência):

- Banco deseja saber se um candidato a empréstimo tem chances de vir a ser inadimplente
- Dados históricos de clientes são usados para diferenciar o perfil dos que foram inadimplentes e dos que não foram
- Verifica-se se o perfil de um novo cliente se encaixa no grupo dos inadimplentes ou não

Agrupamento × discriminante

- Análise de agrupamento (*cluster analysis*): métodos exploratórios usados para dividir a população que não é conhecida *a priori*
- Análise discriminante: os grupos em que um elemento amostral pode vir a ser classificado devem ser conhecidos antes em relação às características

Artigos

Componentes principais:

- ▶ variáveis socioeconômicas e dengue
- ▶ padrão locacional de bancos

Fatorial:

- ▶ índice de desenvolvimento rural
- ▶ avaliação do desempenho financeiro - empresas de seguros
- ▶ fatores de risco cardiovasculares

Agrupamento (*cluster analysis*):

- ▶ municípios baianos
- ▶ gestão de riscos - previdência
- ▶ serviços de saúde - espacial
- ▶ análise de risco - seguradoras agrícolas

Discriminante:

- ▶ situação financeira - estados brasileiros
- ▶ capital humano no Ceará