Introdução à Análise e Modelagem de Dados Multivariados com o R

Lupércio França Bessegato
Dep. de Estatística/UFJF

Roteiro Geral

1. Fundamentos gráficos em R
2. Visualização de dados multivariados
3. Técnicas de interdependência
   - Componentes principais, análise fatorial, escalonamento multidimensional; análise de agrupamentos
4. Análise de dependência
   - Discriminação e classificação; análise discriminante linear; modelos logit.
5. Referências

Análise de Dependência

Roteiro do Módulo

4. Análise de dependência
   - Discriminação e classificação
   - Análise discriminante linear
   - Modelo de regressão logístico

Discriminação e Classificação

Agrupamento e Classificação

- Agrupar:
  - √ Processo de alocar item em grupo
  - √ Não há suposições sobre o número de grupos ou sobre a estrutura dos grupos
    - Técnica mais primitiva
- Classificar:
  - √ Predição de pertinência a grupo
  - √ Número de grupos é conhecido e o objetivo é alocar novas observações a um desses grupos
  - √ Usa status conhecido para encontrar preditores, aplicando-os a uma nova observação

Conjunto de Dados

- Partição do conjunto de dados:
  - √ Conjunto de treinamento
    - Usado para desenvolver modelo de classificação
  - √ Conjunto de teste
    - Usado para determinar desempenho do modelo
  - √ Importante não avaliar desempenho com as mesmas observações usadas para desenvolver o modelo

Passos para Classificação

1. Conjunto de dados coletado, com alocações de item em grupo já conhecidas (ou atribuídas)
   - √ Observação, julgamento de especialista, procedimentos de agrupamento
2. Dados são divididos em conjunto de treinamento e teste
   - √ Treinamento: de 50% a 80% (comum: 67%)
   - √ Restante atribuído ao conjunto de teste

3. Construção do modelo de predição
   - √ Predizer alocação dos dados de treinamento tão bem quanto possível
4. Avaliação do desempenho do modelo usando os dados do conjunto de teste

Métodos de Classificação

- Há inúmero métodos de classificação:
  - √ Análise discriminante
  - √ Regressão logística
  - √ Naive Bayes Classification
  - √ Random Forest Classifiers
  - √ Método do vizinho mais próximo
  - √ Classification and Regression Trees – CART
  - √ Support Vector Machine – SVM
  - √ Método dos núcleos estimadores
  - √ Redes neurais artificiais

Análise de Agrupamento e Análise Discriminante

- Análise de Agrupamentos
  - √ Dividir os elementos da amostra (ou população) em grupos, de maneira que:
    - Elementos de um grupo são similares entre si
    - Elementos de grupos diferentes sejam heterogêneos em relação a essas características

- Análise discriminante:
  - √ Classificação de elementos de amostra (população)
    - Grupos são pré-definidos
  - √ Procedimento:
    - Regra de classificação

Análise Discriminante

Análise Discriminante

- Caso especial de correlações canônicas
  - √ Variáveis dependentes são categóricas por natureza
- Objetivo:
  - √ Usar informações das ariáveis independentes para a separação (discriminação) mais clara possível entre os grupos

- Abordagens:
  - √ Fischer
  - √ Mahalanobis

Aplicações Potenciais

- Perfil:
  - √ Compreender como cada variável independente (X) influencia a variável dependente (Y: grupo)
  - √ Descrição, em análise de regressão
  - √ Quando os objetivos do estudo são principalmente exploratórios

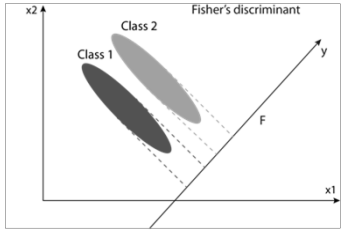
- √ Como os grupos são discriminados pelas variáveis subjacentes?
  - Exame dos perfis de segmentos do mercado para entender como consumidores diferem com relação a variáveis demográficas e psicológicas
  - Diferenças entre usuários de categoria de produto em relação ao tamanho da família, renda, educação, etc.
- √ Como potenciais consumidores de marca diferem da população em geral em relação ao seu envolvimento com a mídia?

- Diferenciação:
  - √ Capacidade de afirmar, com certo nível de confiança, se a relação entre X e Y se deve ao acaso
  - √ Inferência, em análise de regressão
  - √ Traçados os perfis dos grupo, pode ser importante verificar se as diferenças aparentes entre eles dão de fato significativas
  - √ Exemplo:
    - Entender e controlar as variações associadas a certos processos de produção

- Classificação:
  - √ Usar o modelo para avaliar o valor da variável dependente, com observações fora da amostra de treinamento
    - Predizer a pertinência a grupo
  - √ Predição, em análise de regressão
  - √ Exemplos:
    - *Credit scoring*
      - Traçar o perfil dos clientes de empréstimo e julgar se novos candidatos oferecem risco ao crédito
    - Marketing direto
      - Que perfil de clientes devem receber oferta de mala direta?

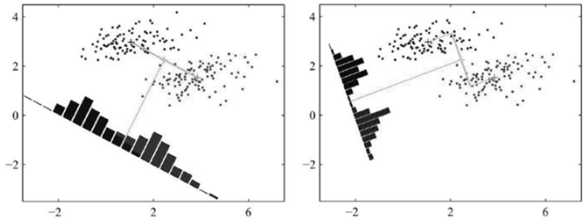
Fisher – Intuição

- Baseia-se na noção de pontuação discriminante



- √ Encontrar combinação linear das variáveis independente que produza pontuações discriminantes maximamente diferentes

- Função objetivo:
  - √ Quantifica a noção de "maximamente diferente"

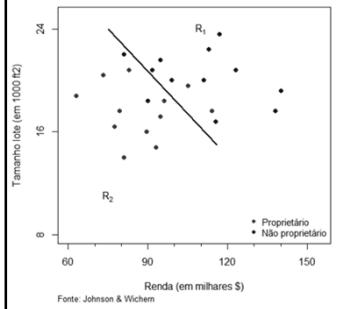


  - √ Função linear que melhor aloca as observações
    - Eixo que descreve diferença entre centróides
    - Ajusta de acordo com o padrão de covariância

21

Mahalanobis – Intuição

- Encontrar o 'locus' dos pontos equidistantes das médias dos 2 grupos



Fonte: Johnson & Wichern

  - √ 2 variáveis explicativas:
    - 'locus' dos pontos é uma linha
  - √ 3 variáveis explicativas:
    - 'locus' dos pontos é um plano ou hiperplano
  - √ 'locus' serve para discriminar os dois grupos

22

- Medida de distância ajustada

$$D_i^2 = (\mathbf{x} - \bar{\mathbf{x}}_{(i)})' \mathbf{C}_W^{-1} (\mathbf{x} - \bar{\mathbf{x}}_{(i)}),\ i = 1, 2.$$

  - √ Distância ao quadrado da covariância ajustada de qualquer ponto ***x*** à média do grupo i
  - √ Dados seguem normal multivariada:
    - Distância ajustada reflete com mais precisão a probabilidade de pertinência ao grupo do que a distância euclidiana

23

- Por definição, 'locus' dos pontos descritos por Malahanobis é ortogonal ao eixo da função discriminante proposta por Fisher

24

Análise Discriminante – Abordagens

- São complementares:
  - √ Fisher:
    - Reduz os dados em uma única dimensão de modo a maximizar a separação entre grupos
  - √ Mahalanobis:
    - Determina linha divisória (ou plano) que separa mais precisamente os dois grupos
    - Ortogonal à dimensão discriminante

Regras de Alocação e Classificação

- Em geral, são desenvolvidas a partir de amostras de treinamento:
  - √ Examinadas diferenças das medidas características de objetos selecionados
  - √ Todos os resultados amostrais possíveis são dividido em duas regiões (R_1 e R_2)
    - Se uma nova observação pertencer à região R_1 ela é alocada à população π_1.
    - Se uma nova observação pertencer à região R_2 ela é alocada à população π_2.

Problema da Classificação

- Como saber se algumas observações pertencem a uma particular população?
  - √ Incerteza na classificação

Paradoxos da Classificação

- Informação incompleta sobre desempenho futuro:
  - √ Classificação de candidato como capaz de concluir ou não um mestrado
- Informação perfeita exige destruição objeto:
  - √ Classificação de itens como bons ou defeituosos
- Informação cara ou indisponível:
  - √ Problemas médicos que podem ser identificados conclusivamente apenas com procedimentos caros

Erros de Classificação

- Caso médico:
  - √ Em geral, deseja-se diagnosticar um mal a partir de sintomas externos facilmente observáveis
- Erro de classificação:
  - √ A distinção entre as características medidas das duas populações pode não ser clara.

29

Critérios para Classificação

- Bom procedimento de classificação:
  - √ Poucos erros de classificação
- Probabilidades a priori deveriam integrar regra ótima:
  - √ Classe (ou população) com verossimilhança de ocorrência maior que outra
  - √ Classe é relativamente maior que outra
  - √ Ex.:
    - Há muito mais empresas solventes que insolventes

30

- Outro aspecto a considerar:
  - √ Custo associado ao erro de classificação
  - √ Ex.:
    - Classificar um objeto π_1 como π_2 é mais sério que classificar um objeto π_2 como π_1.

31

Critérios para Classificação

- Bom procedimento de classificação:
  - √ Poucos erros de classificação
- Probabilidades a priori deveriam integrar regra ótima:
  - √ Classe (ou população) com verossimilhança de ocorrência maior que outra
  - √ Classe é relativamente maior que outra
  - √ Ex.:
    - Há muito mais empresas solventes que insolventes

32

- Outro aspecto a considerar:
  - √ Custo associado ao erro de classificação
  - √ Ex.:
    - Classificar um objeto π_1 como π_2 é mais sério que classificar um objeto π_2 como π_1.

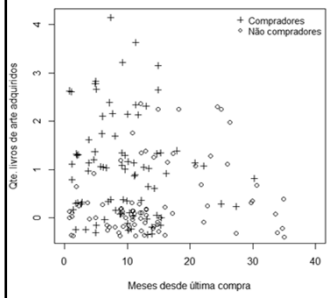
Exemplo

- Clube de livro '*Books by Mail*':
  - √ Oferta de livro de arte
    - Correspondência de teste enviada para 1.000 clientes escolhidos aleatoriamente
    - 83, responderam à oferta
  - √ Informações de compras passadas:
    - X_1: tempo desde a última compra, meses
    - X_2: quantidade de livros de artes adquiridos
  - √ Objetivo:
    - Discriminar compradores e não compradores
  - √ Dados: *BOOKS_1*.txt e *BOOKS_2.txt*

- Importação e tratamento dos dados:

```
> # Carregamento e tratamento conjunto de dados
>
> books <- read.table("BOOKS_1.txt")
> books <- books[-1]
> colnames(books) <- c("tempo", "livros", "compra")
> books$compra <- factor(books$compra, labels=c("N","Y"))
> levels(books$compra) <- c("N", "Y")
> books$tmpc <- cut(books$tempo, breaks = c(0, seq(2.5, 37.5, 5)),
+   labels = c(1, seq(5, 35, 5)))
> medias.livros <- aggregate(livros ~ tmpc, data = books, mean)
> head(books)
  tempo livros compra tmpc
1    24      0      N   25
2    16      0      N   15
3    15      0      N   15
4    22      0      N   20
5    15      0      Y   15
6     6      2      N    5
```


- Análise descritiva do conjunto de dados:
  - √ 83 compradores e 83 não compradores (ao acaso)



- √ Pontos plotados com perturbação
- √ Distribuição dos compradores deslocada em relação aos não compradores
  - A sobreposição é substancial

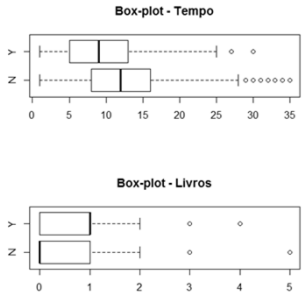
Média das variáveis por grupo:

```
> # Carregamento e tratamento conjunto de dados
> setNames(aggregate(. ~ compra, data = books, mean),
+   c("Compra", "Tempo", "Qte. Livros"))
  Compra     Tempo Qte. Livros
1      N 12.731734   0.3336968
2      Y  9.409639   1.0000000
```

√ Compradores tendem a apresentar;

- Intervalo médio mais curto desde a última compra
- Número médio mais alto de livros de arte adquiridos
  - Maior interesse pela categoria

Diferenças univariadas entre os grupos:



√ Há diferenças entre as médias individuais
√ As diferenças entre as médias conjuntas (centróides) são significativas?
√ Como visualizar

Scatter plot matrix das variáveis:

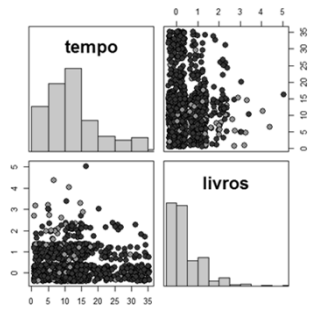
√ Função customizada

```
> # Matrix plot variáveis livros e tempo
> # Customização plot
> panel.hist = function(x, ...) {
+   usr <- par("usr"); on.exit(par(usr))
+   par(usr = c(usr[1:2], 0, 1.5) )
+   h <- hist(x, plot = FALSE)
+   breaks <- h$breaks; nB <- length(breaks)
+   y <- h$counts; y <- y/max(y)
+   rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
+ }
```

√ Gráfico de todos os pontos:

```
> # Gráfico tempo e livros - todos os pontos
> with(books, pairs(jitter(cbind(tempo, livros), ruido), cex = 1.5,
pch = 21,
+       bg = c("red","green3")[unclass(compra)], diag.panel =
panel.hist,
+       cex.labels = 2, font.labels = 2)
+ )
```


Matrix plot das variáveis:

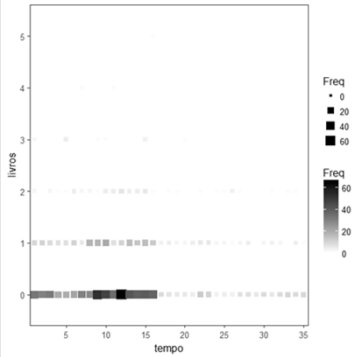


√ Difícil visualização dos grupos
  - Não compradores dominam
  - Há muito empates
√ Variáveis são discretas
√ Distribuições marginais

- Alternativa – *Fluctuation plot*:
  - √ Tamanho das células por frequência:

```
> # Alternativa 1 - Fluctiation plot
> theme_nogrid <- function (base_size = 12, base_family = "") {
+   theme_bw(base_size = base_size, base_family = base_family)
%+replace%
+     theme(panel.grid = element_blank())
+ }
> contagens.df <- with(books, as.data.frame(table(tempo, livros)))
> ggplot(contagens.df, aes(tempo, livros)) +
+   geom_point(aes(size = Freq, color = Freq, stat = "mean",
+ position = "identity"), shape = 15) +
+   scale_size_continuous(range = c(1,5)) +
+   scale_color_gradient(low = "white", high = "black") +
+   scale_x_discrete(breaks = seq(0, 35, 5)) +
+   theme_nogrid()
```

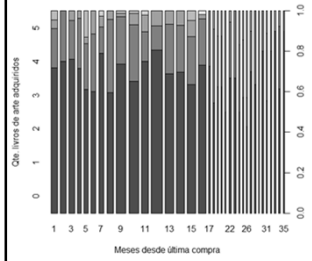

- Alternativa – *Fluctuation plot*:



  - √ Muitos valores zero para tempos abaixo de 16 meses

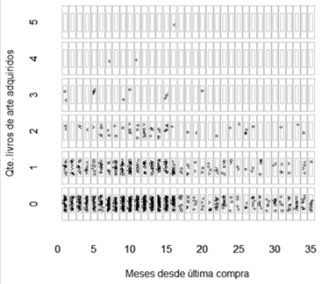
- Alternativa – *Spine plot*:
  - √ Larguras e alturas ponderadas por frequência

```
> # Alternativa 2 - spine plot
> with(books, spineplot(factor(livros)~ factor(tempo),
+ xlab = "Meses desde última compra",
+ ylab = "Qte. livros de arte adquiridos"))
> with(books, spineplot(factor(livros)~ factor(tempo),
+ xlab = "Meses desde última compra",
+ ylab = "Qte. livros de arte adquiridos"))
```



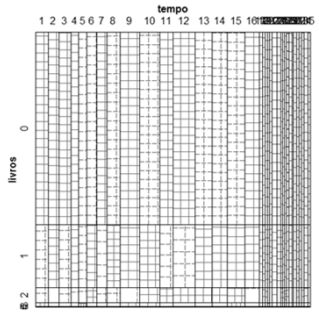
  - √ Valores concentrados abaixo de 16 meses

- Alternativa – *Scatter plot* modificado:
  - √ Visualização dos valores observados em cada par de valores da variáveis



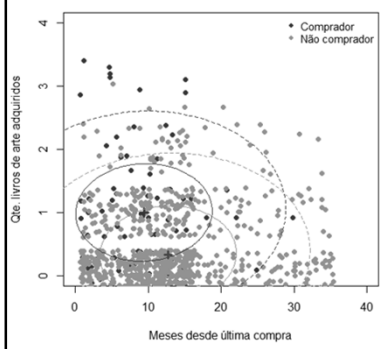
  - √ Valores concentrados abaixo de 16 meses e 1 livro

- Alternativa – *Sive plot*:
  - √ Visualização com contagem por célula



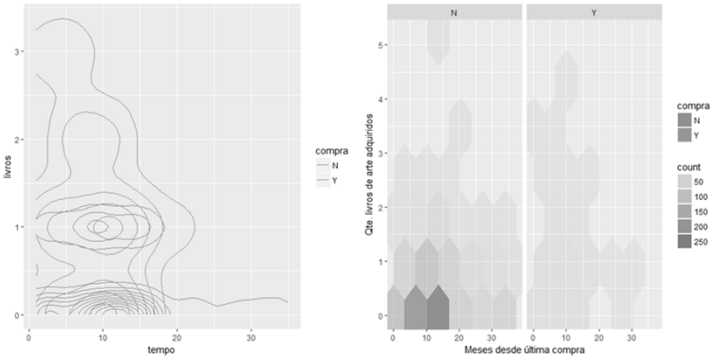
- √ Concentração nos valores de zero livros comprados

- Scatter plot por grupos:
  - √ Há diferenças entre os centróides?

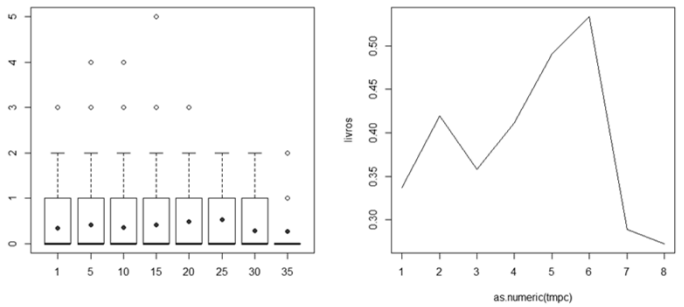



```
> # centróides
  compra     Tempo Qte. Livros
1      N 12.731734   0.3336968
2      Y  9.409639   1.0000000

> # Matriz covariâncias - grupo
> cov.lista
[[1]]
              tempo    livros
tempo  65.7270814 0.2391806
livros  0.2391806 0.3688742

[[2]]
              tempo     livros
tempo  35.4155157 -0.6707317
livros -0.6707317  1.1219512

> # Matriz covariâncias combinada
> sigma.pol
              tempo    livros
tempo  63.2365519 0.1644183
livros  0.1644183 0.4307503
```

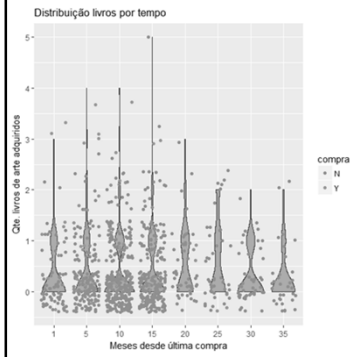

- *Scatter plot* por grupos – Estimado:



- Box-plot: livros por intervalo entre compras:
  - √ Tempo categorizado



- *Violin plot*:
  - √ Visualização da distribuição dos dados e de sua densidade.



  - √ Semelhante box plot
  - √ Apresenta densidade condicional
  - √ Cuidado com o uso:
    - No caso, as variáveis são discretas

- Centróides:

```
> centroide <- aggregate(cbind(tempo, livros) ~ compra, data = books, mean)
> centroide
  compra     tempo    livros
1      N 12.731734 0.3336968
2      Y  9.409639 1.0000000
```

- Distância entre os centróides:

$$\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}}_{(1)} = \begin{bmatrix} 9,40 \\ 1,00 \end{bmatrix} - \begin{bmatrix} 12,70 \\ 0,33 \end{bmatrix} = \begin{bmatrix} -3,30 \\ 0,67 \end{bmatrix}.$$

- √ Matrizes das somas de quadrados – *within*:

```
> cov(books[books$compra == "N", 1:2])*(1000 - 83 -1) # SS within N
           tempo   livros
tempo  60206.0065 219.0894
livros   219.0894 337.8888

> cov(books[books$compra == "Y", 1:2])*(83-1)# SS within Y
           tempo livros
tempo  2904.072    -55
livros  -55.000     92
```

- √ Matriz de covariâncias combinada – *between*:

```
> books.aov <- manova(cbind(tempo, livros) ~ compra, data = books)
> estVar(books.aov)# Matriz de covariâncias combinada (entre grupos)
           tempo     livros
tempo  63.2365519 0.1644183
livros  0.1644183 0.4307503
```

- √ Inversa da matriz de covariâncias combinada:

```
> solve(estVar(books.aov))     # Inversa da matrz de covariâncias combinada
             tempo       livros
tempo    0.015829349 -0.006042095
livros -0.006042095  2.323837045
```


- Função discriminante:

```
> # Função Discriminante de Fisher
> library(MASS) # comando lda
> ajuste.df <- lda(books[, 1:2], books$compra, data = books)
> ajuste.df
Call:
lda(books[, 1:2], books$compra, data = books)

Prior probabilities of groups:
    N     Y
0.917 0.083

Group means:
      tempo    livros
N 12.731734 0.3336968
Y  9.409639 1.0000000

Coefficients of linear discriminants:
                LD1
tempo  -0.05098078
livros  1.41242601
```

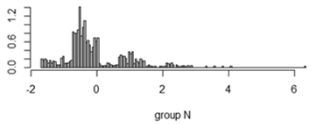
Centróides

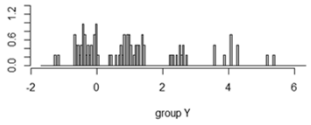
Proporcional a [–0,056; 1,577]

  - √ R padroniza variável discriminante:
    - Média zero e desvio-padrão 1

Escores discriminantes dos grupos:

```
> predicao <- predict(ajuste.df, books[, 1:2])
> GrupoPrevisto <- predicao$class
> ldahist(data = predicao$x, g = books$compra, h = 0.05)
```





- Compradores:
  - √ Em média mais positivos
- Não compradores:
  - √ Em média mais negativos

56

Modelo de Regressão Logístico

Referências

Bibliografia Recomendada

- ALBERT, J.; RIZZO, M. *R by Example*. Springer, 2012.
- CHAPMAN, C.; FEIT, E. M. *R for marketing research and analytics*. Springer, 2015.
- KLEIBER, C.; ZEILEIS, A. *Applied econometrics with R*. Springer, 2008.
- DALGAARD, P. *Introductory statistics with R*. Springer, 2008.

59