Discriminação e Classificação

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

1. Introdução
2. Classificação e Discriminação
3. Análise Discriminante
4. Regressão Logística Multinomial
5. Outras Abordagens
6. Referências

Introdução

Análise Multivariada

- Considera várias variáveis relacionadas simultaneamente
- Variáveis de interesse não-independentes uma das outras
- Associação entre conjuntos de medidas

Objetos

- Entidades das quais são tomadas medidas
  - √ Itens, pessoas, organizações, etc.
  - √ São portadores de medidas
  - √ São medidos somente com respeito a certas variáveis de interesse

Variáveis

- Características ou propriedades
  - √ São os aspectos dos objetos que são medidos

Organização de Dados Multivariados

$$\mathbf{X}_{n\times p} = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ \mathrm{n} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & \mathrm{p} \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{array}$$

p: número de variáveis
n: número de objetos

- √ Matriz:
  - – Organização em arranjo estrutural de dados multivariados (métricos ou não métricos)
- √ Colunas: variáveis
  (características medidas de objetos)
- √ Linhas: objetos
  (lista de características medidas de um objeto)

Classificação e Agrupamento

- Classificar:
  - √ Número de grupos é conhecido e o objetivo é alocar novas observações a um desses grupos
- Agrupar:
  - √ Não há suposições sobre o número de grupos ou sobre a estrutura dos grupos
    - – Técnica mais primitiva

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

Discriminação e Classificação para Duas Populações

Objetivo

1. Separar duas classes de objetos
2. Atribuir um novo objeto a uma das duas classes

- Classes:
  - √ π_1 e $\pi 2$
- Objetos:
  - √ São separados ordinariamente ou classificados com base em medidas de p variáveis aleatórias associadas

Conceitos

- Classes:
  - √ π_1 e $\pi 2$
- Objetos:
  - √ São separados ordinariamente ou classificados com base em medidas de p variáveis aleatórias associadas

$$\mathbf{X}' = [X_1, X_2, \ldots, X_p].$$

- Hipótese:
  - √ Os valores observados de **X** diferem em alguma quantidade de uma classe a outra

- Populações das duas classes:
  - √ Podem ser descritas por suas funções de densidade [$f_1(\mathbf{x})$ e $f_2(\mathbf{x})$]
  - √ Pode-se falar em atribuir:
    - Observações a populações ou
    - Objetos a classes

Exemplos

Populações π_1 e π_2	**Variáveis X**
Risco de crédito alto/baixo	Renda, Idade, nº de cartões de crédito, tamanho família
Duas espécies de flor	Comprimento e largura de pétalas e sépalas, diâmetro do pólen, etc.
Masculino e feminino	Medidas antropológicas tais como: circunferência e volume de crânios antigos
Seleção a curso de pós-graduação	Histórico escolar, curriculum vitae, cartas de referência, experiência profissional

- Outros Exemplos:
  - √ Agricultura:
    - Identificar áreas de maior potencial para plantação de determinadas sementes
  - √ Marketing:
    - Identificar mercados potenciais e não potenciais para determinados produtos e serviços
  - √ Esporte:
    - Identificação de atletas promissores para cada modalidade
  - √ Estudos de criminalidade:
    - Identificação de regiões que necessitam de política de segurança diferenciada

Regras de Alocação e Classificação

- Em geral, são desenvolvidas de amostras de treinamento:
  - √ Examinadas diferenças das medidas características de objetos selecionados
  - √ O conjunto de 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.

Semana da Estatística/UFJF - 2016 19
Discriminação e Classificação

Problema da Classificação

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

Semana da Estatística/UFJF - 2016 20
Discriminação e Classificação

Paradoxo da Classificação

- Conhecimento incompleto de 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 cirurgia cara

Semana da Estatística/UFJF - 2016 21
Discriminação e Classificação

Erros de Classificação

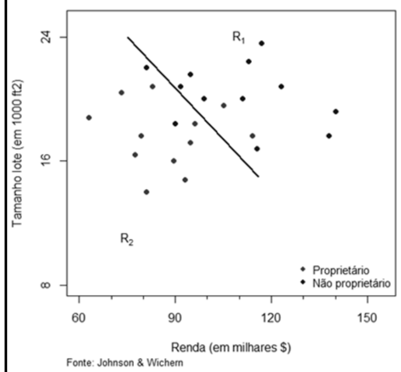
- Caso médico:
  - √ Em geral, deseja-se diagnosticar um mal a partir de sintomas externos facilmente observáveis
- Erro de classificação:
  - √ Pode não ser clara a distinção entre as características medidas das duas populações.

Semana da Estatística/UFJF - 2016 22
Discriminação e Classificação

Exemplo

- Discriminação de proprietários e não-proprietários de cortador de grama
  - √ X_1: renda
  - √ X_2= tamanho do lote

- Proprietários e não proprietários: Renda e tamanho lote



Fonte: Johnson & Wichern

- Proprietários tendem a ter rendas e lotes maiores
- Renda aparenta discriminar melhor que tamanho do lote
- Há uma certa sobreposição entre os dois grupos
  - √ Erros de classificação

Ideia:
Criar regra que minimize a chance de erros de classificação

Critérios para Classificação

- Bom procedimento de classificação:
  - √ Poucos erros de classificação
- Regra ótima deveria considerar estas probabilidades a priori
  - √ Pode ser que uma classe (ou população) tenha uma verossimilhança de ocorrência maior que outra
  - √ Uma das classes é relativamente maior que a outra
  - √ Ex.:
    - Há muito mais empresas solventes que insolventes

- 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.

Caso de Classificação de Duas Populações Normais Multivariadas

Classificação em Duas Populações

- Supondo disponíveis:
  - √ Conjunto de observações independentes de duas populações π_1 e π_2
  - √ Distribuições de probabilidades do vetor X, associadas às populações π_1 e π_2.
- Regra de classificação que minimize a chance de se classificar incorretamente elemento amostral:
  - √ Princípio da máxima verossimilhança

Exemplo

- Processo de seleção de alunos:
  - √ Fase 1: todos fazem várias provas
  - √ Fase 2: apenas aprovados na fase 1
- Populações:
  - √ População 1:
    - Alunos que passaram na 1ª. Fase, mas reprovados na 2ª
  - √ População 2:
    - Alunos aprovados em ambas as fases

Exemplo

- Objetivo:
  - √ A partir dos dados, construir uma rega de classificação que permita identificar, dentre os aprovados na 1ª. Fase, quais provavelmente serão aprovados na 2ª Fase
- Considere apenas a variável aleatória nota na prova de Matemática dos candidatos na fase 1

Exemplo

- Considere apenas a variável aleatória nota na prova de Matemática dos candidatos na fase 1
- Suponha que X tenha uma distribuição normal
  - √ População 1: média μ_1.
  - √ População 2: média μ_2
  - √ Ambas populações como o mesmo desvio padrão σ.

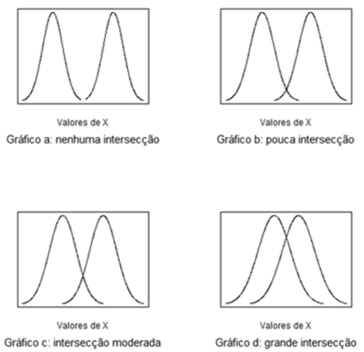
Exemplo

- Razão de verossimilhança entre as 2 populações

$$\lambda(x) = \frac{f_1(x)}{f_2(x)} = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right\}}{\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2\right\}}$$
$$= \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma}\right)^2 - \left(\frac{x-\mu_2}{\sigma}\right)^2\right]\right\}$$

- Para uma nota fixa x:
  - √ $\lambda(x) > 1$: razoável classificar candidato em π_1 (não aprovado da Fase 2)
  - √ $\lambda(x) < 1$: provável aprovado Fase 2 (π_2)
  - √ $\lambda(x) = 1$: candidato poderia ser classificado em π_2 ou π_2)
    - – Obter informações adicionais sobre o candidato

- Classificação de 2 populações normais, com mesma variabilidade



Fonte: Mingoti

Qualidade da discriminação depende do grau de intersecção

- √(a): número de classificações incorretas é 0
- √(b): pequeno número de erros de classificação
- √(c) e (d): número de erros de classificação tende a aumentar
- √Intersecção pode chegar a valores que inviabiliza o uso da função discriminante como regra de classificação

Função Discriminante

$$\lambda(x) = \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma}\right)^2 - \left(\frac{x-\mu_2}{\sigma}\right)^2\right]\right\}$$
$$-2\ln(\lambda(x)) = \left(\frac{x-\mu_1}{\sigma}\right)^2 - \left(\frac{x-\mu_2}{\sigma}\right)^2$$
$$= \frac{1}{\sigma^2}\left[(x-\mu_1)^2 - (x-\mu_2)^2\right].$$

- √ Relacionada com a diferença das distância euclidiana ponderadas ao quadrado
- √ $\lambda(x) > 1$ ➔ $-2\ln(\lambda(x)) < 0$: x está mais próximo de μ_1.
- √ $\lambda(x) < 1$ ➔ $-2\ln(\lambda(x)) > 0$: x está mais próximo de μ_2.

- Regra de classificação:
  - √ Se $-2\ln(\lambda(x))<0$, classifique o elemento amostral em π_1
  - √ Se $-2\ln(\lambda(x))>0$, classifique o elemento amostral em π_2.
  - √ Se $-2\ln(\lambda(x))=0$, o elemento amostral poderá ser classificado tanto em π_1 como π_2.

Populações com Variâncias Diferentes

- Função discriminante:

$$\lambda(x) = \frac{f_1(x)}{f_2(x)} = \frac{\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\}}{\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right\}}$$

$$= \frac{\sigma_2}{\sigma_1}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_{\ 2}}\right)^2\right]\right\}$$

  - √ e, considerando $-2\ln(\lambda(x))$:

$$-2\ln(\lambda(x)) = -2\ln\left(\frac{\sigma_2}{\sigma_1}\right) + \left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right].$$

- Com mesma regra de classificação

Populações Multivariadas – Caso $\Sigma_1 \neq \Sigma_2$

- Populações normais multivariadas, com vetor de médios μ_i e matriz de covariâncias $\mathbf{\Sigma}_i$, i = 1, 2.
- Função discriminante:

$$-2\ln(\lambda(\mathbf{x})) = -2\ln\left\{\frac{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}_1|^{-\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}_2|^{-\frac{1}{2}}}\left[\frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}}\right]\right\}$$
$$= \left[(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] + \left[\ln|\mathbf{\Sigma}_1| - \ln|\mathbf{\Sigma}_2|\right]$$

  - √ para um vetor de observações: $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$.
- Com mesma regra de classificação

- Função discriminante quadrática

$$\left[(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] + \left[\ln|\mathbf{\Sigma}_1| - \ln|\mathbf{\Sigma}_2|\right]$$

  - √ Depende das distâncias de Mahalanobis do vetor **x** aos vetores de médias $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$
  - √ Fator de correção relacionando as variâncias generalizadas das duas populações

- Quando $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

$$-2\ln(\lambda(\mathbf{x})) = \left[(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]$$

√ Pode ser reescrita como

$$-2\ln(\lambda(\mathbf{x})) = -2\left[(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right].$$

Função Discriminante de Fisher

$$f_d(\mathbf{x}) = \ln(\lambda(\mathbf{x})) = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2).$$

- Regra de classificação:

√ **x** é classificado à π_1 se f_d(x) > 0 ou seja, se

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} > \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2).$$

√ **x** é classificado à π_2 se f_d(x) < 0 ou seja, se

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} < \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2).$$

- A função discriminante de Fisher tem a forma

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \mathbf{b}'\mathbf{x} = b_1X_1 + b_2X_2 + \cdots + b_pX_p.$$

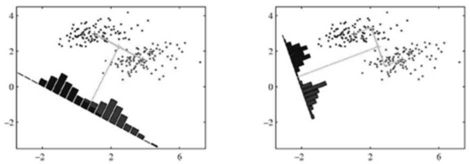
√ Dependendo do valor numérico desta combinação, o elemento amostral é classificado em uma ou outra população

- Constante de delimitação da região de classificação

$$\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) = \mathbf{b}'\frac{(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)}{2}.$$

√ Combinação linear dos vetores de médias das populações

- Discriminação de duas populações normais:



√ Os valores observados das combinações lineares **b'X** na população π_1 são os mais separados possíveis daqueles observados da população π_2.

Padronização

- O vetor **b** é único $\mathbf{b}' = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}$.
  - √ Exceto para multiplicações de todos seus componentes pela mesma constante c.
- É recomendável que os componentes do vetor **b** sejam padronizados ou normalizados, como em:

$$\mathbf{b}^* = \frac{\mathbf{b}}{\sqrt{\mathbf{b}'\mathbf{b}}}.$$

  - √ Componentes de $\mathbf{b}^*$ estarão no intervalo [–1, 1]
  - √ É possível comparar os loadings de f_{dn}

- Mesclar com Scaling – Johnson, pág. 589/590

Custos

- Regra de discriminação pelo princípio da máxima verossimilhança minimiza as probabilidades de erros de classificações
  - √ Máxima separação entre as combinações lineares
- Não leva em consideração possíveis diferenças entre os custos associados aos erros de classificação

Mistura de Duas Normais Multivariadas

- Suponha:
  - √ 2 populações normais p-variadas com vetor de médias $\boldsymbol{\mu}_i$, i = 1,2
  - √ Probabilidades de mistura p_j, j=1,2
    - – Probabilidade de que uma observação escolhida ao acaso pertença à π_j, j=1,2.
  - √ Função de densidade de π_j. $f(\mathbf{x}|\pi_j) = f(\mathbf{x}|(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}))$.

√ Dada uma observação **x**, qual a melhor maneira de distinguir de qual das duas populações ela foi amostrada?

$$f(\mathbf{x}\text{ amostrada de }\pi_j) = p_j\, f(\mathbf{x}|(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})\,.$$

√ Densidade de **x** oriunda de uma população não especificada:

$$f(\mathbf{x}) = p_1\, f(\mathbf{x}|(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + p_2\, f(\mathbf{x}|(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\,.$$

√ Densidade posterior de **x** de amostras da população j para uma dada observação **x**

$$f(\pi_j|\mathbf{x}) = \frac{p_j\, f(\mathbf{x}|(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})}{p_1\, f(\mathbf{x}|(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + p_2\, f(\mathbf{x}|(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})}\,.$$

Expressão básica para estimar a população da qual a observação **x** foi amostrada

- Função discriminante de Fisher:

$$f_d(\mathbf{x}) = \underbrace{\ln\left(\frac{p_1}{p_2}\right) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}_{\text{constante}} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}\,.$$

- Limite entre as populações π_1 e π_2.

$$f_d(\mathbf{x}) = \text{constante} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} = 0\,.$$

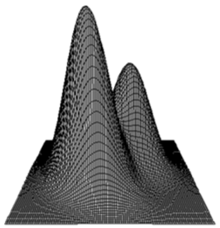
Exemplo

- Mistura de normais bivariadas com mesma estrutura de variabilidade:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0,4 \\ -0,4 & 1,2 \end{bmatrix}.$$

√ π_1: $p_1 = 0{,}6$ e $\boldsymbol{\mu}_1 = (-1, -1)'$.

√ π_2: $p_2 = 0{,}4$ e $\boldsymbol{\mu}_2 = (1, 1)'$.



- Função discriminante:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \begin{bmatrix} -1-1 & -1-1 \end{bmatrix} \begin{bmatrix} 1 & -0,4 \\ -0,4 & 1,2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$
$$= \begin{bmatrix} -3,0769 & -2,6923 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$
$$= -3,0769x_1 - 2,6923x_2\,.$$

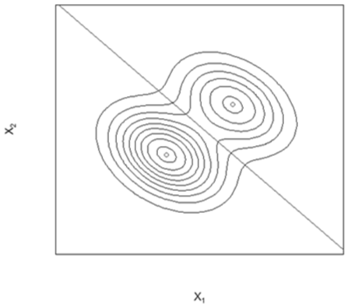
- Constante de delimitação:

$$\text{cte} = \ln\left(\frac{0,6}{0,4}\right) - \frac{1}{2}\left(\begin{bmatrix} -1-1 & -1-1 \end{bmatrix} \begin{bmatrix} 1 & -0,4 \\ -0,4 & 1,2 \end{bmatrix}^{-1} \begin{bmatrix} -1+1 \\ -1+1 \end{bmatrix}\right)$$
$$= 0,40546\,.$$

- Fronteira

$$-3,0769x_1 - 2,6923x_2 = 0,40546$$
$$x_2 = 0,1506 - 1,1428x - 1\,.$$

- Fronteira entre as duas populações



$$x_2 = 0,1506 - 1,1428x - 1\,.$$

√ Linha divisória discriminando as duas populações

Estimação da Regra de Classificação

- Na prática, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ e $\boldsymbol{\Sigma}_2$ não são conhecidos
- Caso 1: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

√ $\boldsymbol{\Sigma}$ é estimada por **S**: $S_{\text{pol}} = \dfrac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}\,.$

√ Função discriminante de Fisher estimada por:

$$\hat{f}_d(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pol}}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pol}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\,.$$

- Caso 2: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

√ Função discriminante quadrática estimada por:

$$\begin{aligned}-2\ln(\hat{\lambda}(\mathbf{x})) &= \left[(\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}_1^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) - (\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}_2^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2)\right] \\ &\quad + \left[\ln|\mathbf{S}_1| - \ln|\mathbf{S}_2|\right]\end{aligned}$$

√ Considera os sistemas de variabilidade das duas populações separadamente.

Estrutura de Variabilidades das Populações

- Testes de hipóteses para decidir se as matrizes $\boldsymbol{\Sigma}_1$ e $\boldsymbol{\Sigma}_2$ são iguais ou diferentes
- Alternativa prática:

√ Ajuste aos dados dos dois modelos: linear de Fisher e quadrático

√ Escolhe-se o modelo que resultar em menores proporções de erros de classificação

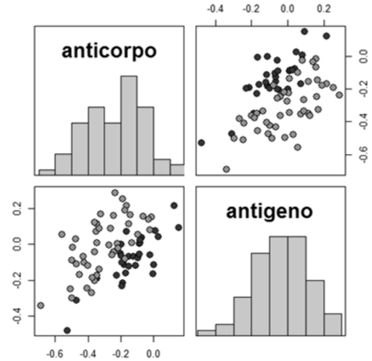
√ Caso os resultados sejam semelhantes, opta-se pelo modelo linear

– Matriz de covariâncias estimada com mais observações

Exemplo

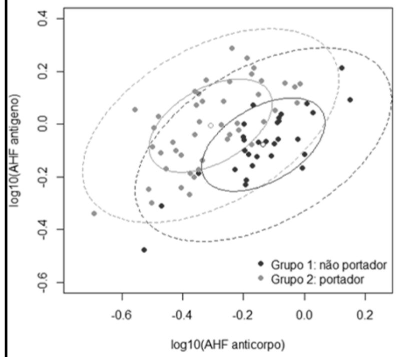
- Detecção de portadoras de hemofilia A
  - √ Grupo 1: mulheres que não têm o gene da hemofilia
    - Grupo normal ($n_1 = 30$)
  - √ Grupo 2: mulheres portadoras do gene da hemofilia
    - Filhas de hemofílicos, mães com mais de um filho hemofílico e outros parentes hemofílicos ($n_2 = 45$)
  - √ Variáveis:
    - V_1: grupos
    - V_2: log(atividade AHF)
    - V_3: log(antígeno AHF)

- Análise descritiva do conjunto de dados:




```
> t(centroide)
          [,1]         [,2]
V2 -0.13487000 -0.307946667
V3 -0.07785667 -0.005991111

> sigma.pol
           V2        V3
V2 0.02263709 0.0154313
V3 0.01543130 0.0216058
```


- *Contour plot* por grupo



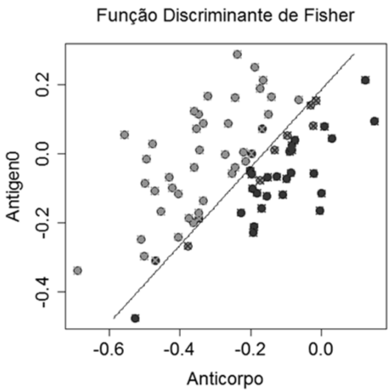
√ Curvas contendo 50% e 95% de probabilidade para normais bivariadas centradas em $\bar{x}_1$ e $\bar{x}_2$.

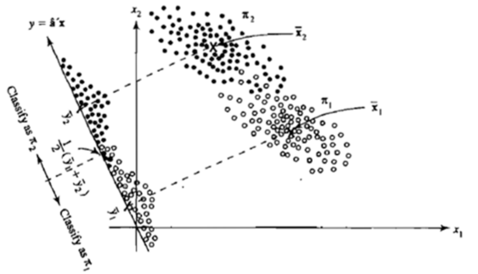
√ Normal bivariada aparenta se ajustar bem aos dados

- Função discriminante de Fisher

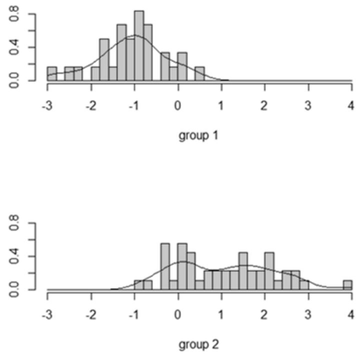
```
>> ajuste.ad = lda(dados, grupo)
> ajuste.ad
Call:
lda(dados, grupo)

Prior probabilities of groups:
  1   2
0.4 0.6

Group means:
          V2           V3
1 -0.1348700 -0.077856667
2 -0.3079467 -0.005991111

Coefficients of linear discriminants:
         LD1
V2 -9.032787
V3  8.006605
```

- Gráfico da função discriminante





- Grupos ajustados



- Predição:
  - √ Mulher com V1 = –0,210 e V2 = –0,044

```
> predict(ajuste.ad, newdata = c(-0.210, -0.044))$class
[1] 2
Levels: 1 2
```

  - √ Mulher pode ser portadora de hemofilia

Avaliação das Funções de Classificação

Erros de Classificação

- Erros a serem avaliados:
  - √ Erro 1: Elemento amostral pertence a π_1, mas é classificado em π_2.
  - √ Erro 2: Elemento amostral pertence a π_2, mas é classificado em π_1.
- Notação:
  - √ P(Erro 1) = p(2|1)
  - √ P(Erro 2) = p(1|2)
- Quanto menores essas probabilidades, melhor será a função de discriminação

Procedimentos de Estimação dos Erros de Classificação

1. Método da Ressubstituição:
   - √ Calculado o escore de cada elemento amostral
   - √ Calculada a frequência das classificações corretas e incorretas
   - √ Estimação da regra de classificação e dos erros de classificação com os mesmos elementos

- Tabela de frequência de classificação:

		Classificação		
		π_1	π_2	Total
Origem	π_1	n_{11}	n_{12}	n_1
	π_2	n_{21}	n_{22}	n_2

  - √ Estimativa das probabilidades de erros de classificação:

$$\hat{p}(2|1) = \frac{n_{12}}{n_1}$$
$$\hat{p}(1|2) = \frac{n_{21}}{n_2}\,.$$

Comentários

- Também denominado de estimação do erro aparente de classificação (APER)
- Procedimento consistente, mas viciado
  - √ Vício tende a zero para n_1 e n_2 grandes
  - √ Tende a subestimar os verdadeiros valores de p(1|2) e p(2|1) para elementos que não pertencem à amostra conjunta $n = n_1 + n_2$.
  - √ Pode servir como etapa inicial de avaliação
    - Valores elevados indica a necessidade de reformulação da regra de discriminação

2. Método de colocação de elementos à parte para classificação (Hold-out validation)
   - √ Amostra conjunta é repartida em duas partes
     - Amostra de treinamento: construção da regra de discriminação
     - Amostra de validação: para estimação dos erros de classificação
   - √ São selecionados aleatoriamente os elementos amostrais que constituirão cada amostra
   - √ Estimação dos erros de classificação da maneira descrita no método de resubstituição

Comentários

- Procedimento não é enviesado
- Recomendável:
  - √ Separar de 25% a 50% dos elementos originais para a amostra de validação
- Desvantagem:
  - √ Redução do tamanho da amostra original para estimação da regra de discriminação
  - √ Não pode ser empregado em amostras pequenas
- Para amostras grandes, é melhor que o método 1

3. Método de validação cruzada (Método de Lachenbruch)
   - √ Retira-se um elemento amostral da amostra conjunta e constrói-se a função de discriminação
   - √ Utiliza-se a regra de discriminação para classificar o elemento que ficou à parte
   - √ Elemento amostral é retornado à amostra e retira-se elemento amostral diferente do anterior, repetindo-se o procedimento.

- Estimação dos erros de classificação:

$$\hat{p}(2|1) = \frac{n_{12}}{n_1}$$
$$\hat{p}(1|2) = \frac{n_{21}}{n_2}\,.$$

Comentários

- Estimativas são aproximadamente não viciadas
  - √ Melhores que o método da ressubstituição para populações normais e não-normais

Estimação da Probabilidade Global de Acerto

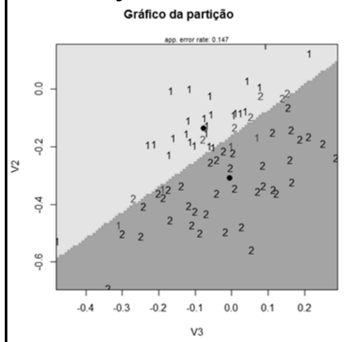
- Estimação da probabilidade global de acerto da função discriminante:

$$\hat{p}(\text{acerto}) = \frac{n_{11} + n_{22}}{n_1 + n_2}.$$

  - √ Recomendável estimar as probabilidades de ocorrência dos erros de classificação tipo 1 e 2
    - – Possível função discriminante com alta probabilidade de acerto global, mas apresentando alta probabilidade de algum dos erros parciais.

Exemplo

- Conjunto de dados: Hemofilia




```
> ajuste.ad <- lda(dados, grupo)
> predicao <- predict(ajuste.ad, dados)
> GrupoPrevisto <- predicao$class
> contagem <- table(grupo, GrupoPrevisto)
> diag(prop.table(contagem, 1))
        1         2
0.8666667 0.8444444
> sum(diag(prop.table(contagem)))
[1] 0.85333336058
```


Construção da Regra de Discriminação: Caso de Várias Populações

Regra de Discriminação para Várias Populações

- Classificar unidades amostrais em g>2 populações
  - √ $f_i(x)$: função de densidade π_i, i = 1, 2, ... g.
- Objetivo:
  - √ Construir regra de classificação que minimize as probabilidades de erros de classificação

Procedimento

- Para um vetor de observações $\boldsymbol{x}$:
  - √ Calcula-se o valor de $f_i(x)$, para cada i.
  - √ Classifica-se o elemento amostral na população k correspondente ao maior valor $f_i(x)$.
- No caso de população normal multivariada, corresponde a classificar na população k, tal que:

$$d_k^Q(x) = \max\{d_1^Q(x), d_2^Q(x), \ldots, d_k^g(x)\}$$

  - √ Sendo d_i^Q: escore quadrático de discriminação

$$d_i^Q = -\frac{1}{2}\ln(|\boldsymbol{\Sigma}_i|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

- Na prática, os escores quadrático de discriminação são estimados por:

$$\hat{d}_i^Q = -\frac{1}{2}\ln(|\mathbf{S}_i|) - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)$$

- Se $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\ldots=\boldsymbol{\Sigma}_g=\boldsymbol{\Sigma}$, usa-se $\mathbf{S}_{pol}$ para estimar $\hat{d}_i^Q$:

$$S_{\text{pol}} = \frac{1}{\sum_{i=1}^{g} n_i - g}\sum_{i=1}^{g}(n_i-1)S_i$$

Erro de Classificação

- Elemento amostral pertence a π_j, mas a regra de discriminação o classifica em π_k, j, k = 1,2,...,g, $j \neq k$.
- Erros estimados por:

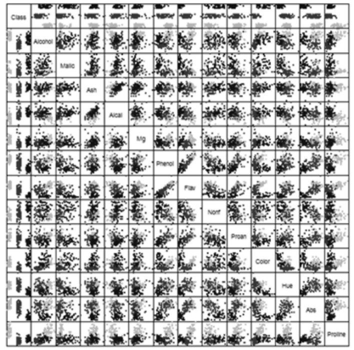
$$\hat{p}(k|j) = \frac{n_{jk}}{n_j}$$

  - √ n_{jk}: número de elementos de π_j classificados em π_k.

Exemplo

- Cultivares de Vinho: 178 tipos de vinhos
  - √ Class: cultivares de vinho
  - √ Variáveis:
    - Alcohol
    - Malic:
    - Ash
    - Alcal
    - Mg
    - Phenol
    - Flav:
    - Nonf
    - Proan
    - Color
    - Hue
    - Abs
    - Proline:

- Matrix scatter plot:



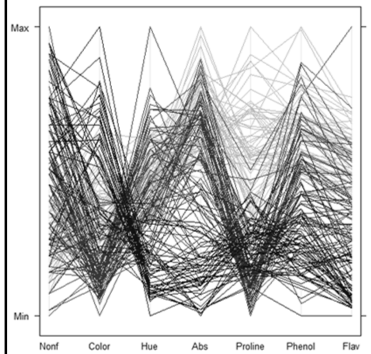
- Há variáveis que diferem entre os grupos

- Médias das variáveis por grupo

```
> round(t(aggregate(vinho, list(vinho$Class),
    mean)[,2:14]),2)
          [,1]  [,2]  [,3]
Class     1.00  2.00  3.00
Alcohol  13.74 12.28 13.15
Malic     2.01  1.93  3.33
Ash       2.46  2.24  2.44
Alcal    17.04 20.24 21.42
Mg      106.34 94.55 99.31
Phenol    2.84  2.26  1.68
Flav      2.98  2.08  0.78
Nonf      0.29  0.36  0.45
Proan     1.90  1.63  1.15
Color     5.53  3.09  7.40
Hue       1.06  1.06  0.68
Abs       3.16  2.79  1.68
```

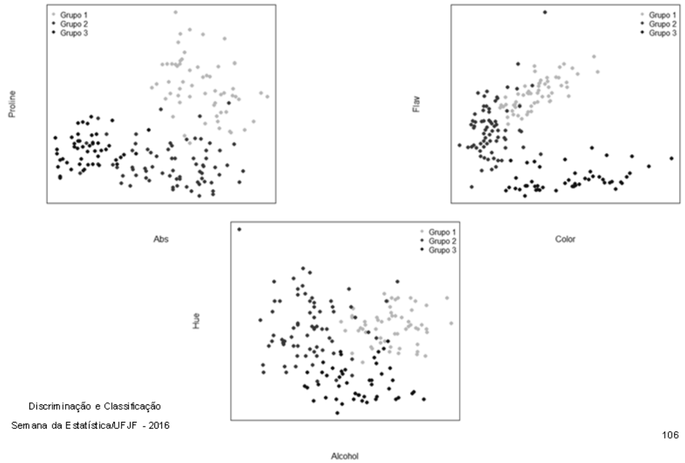
- Abordagem marginal é intuitiva para identificar médias dos grupos
  - √ Falha em identificar características para discriminar os indivíduos
  - √ Abordagem univariada não considera correlação entre variáveis
- Pode haver combinação de variáveis que fornecem nível mais alto de discriminação

- Parallel coordinate plot



- Há alguma sobreposição entre grupos

√ Prováveis pares de variáveis úteis para discriminação

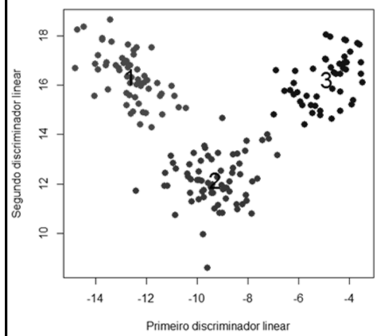


- Análise discriminante

```
> ajuste.ld <- lda(Class~., data = vinho)
> ajuste.ld
Call:
lda(Class ~ ., data = vinho)

Prior probabilities of groups:
        1         2         3
0.3314607 0.3988764 0.2696629

Group means:
   Alcohol    Malic      Ash    Alcal       Mg   Phenol
Flav     Nonf
1 13.74475 2.010678 2.455593 17.03729 106.3390 2.840169
2.9823729 0.290000
2 12.27873 1.932676 2.244789 20.23803  94.5493 2.258873
2.0808451 0.363662
3 13.15375 3.333750 2.437083 21.41667  99.3125 1.678750
0.7814583 0.447500
     Proan    Color       Hue      Abs   Proline
1 1.899322 5.528305 1.0620339 3.157797 1115.7119
2 1.630282 3.086620 1.0562817 2.785352  519.5070
3 1.153542 7.396250 0.6827083 1.683542  629.8958
```

```
Coefficients of linear discriminants:
                 LD1           LD2
Alcohol -0.403399781  0.8717930699
Malic    0.165254596  0.3053797325
Ash     -0.369075256  2.3458497486
Alcal    0.154797889 -0.1463807654
Mg      -0.002163496 -0.0004627565
Phenol   0.618052068 -0.0322128171
Flav    -1.661191235 -0.4919980543
Nonf    -1.495818440 -1.6309537953
Proan    0.134092628 -0.3070875776
Color    0.355055710  0.2532306865
Hue     -0.818036073 -1.5156344987
Abs     -1.157559376  0.0511839665
Proline -0.002691206  0.0028529846

Proportion of trace:
   LD1    LD2
0.6875 0.3125
```


- Discriminadores lineares dos dados



- Discriminadores:
  - √ Combinações lineares que melhor discriminam os grupos
- Figura indica distinção clara entre os grupos

Função de Discriminação Logística

Regressão Logística

- Método de regressão para distinguir entre dois grupos diferentes
- Critério de classificação:
  - √ Baseado na probabilidade de indivíduo pertencer a grupo
- Resposta:
  - √ Binária
  - √ Condicionada a vetor de variáveis explicativas **x**.
- i-ésima observação: $(y_i, \boldsymbol{x}_i)$

Modelo Logístico

$$\ln\left(\frac{\mathrm{P}\{Y=0|\mathbf{x}\}}{\mathrm{P}\{Y=1|\mathbf{x}\}}\right) = \boldsymbol{\beta}'\mathbf{x}\,.$$

- √ log da chance é função linear das variáveis explicativas
- Probabilidade da resposta

$$\mathrm{P}\{Y=0|\mathbf{x}\} = \frac{\exp\{\boldsymbol{\beta}'\mathbf{x}\}}{1+\exp\{\boldsymbol{\beta}'\mathbf{x}\}}\,.$$

  - √ $\boldsymbol{\beta}$ pode ser estimado por máxima verossimilhança
  - √ Interpretação de $\boldsymbol{\beta}$:
    - – Alteração no log da chance para cada variação de uma unidade na variável explicativa

Generalização do Modelo

- √ Mais de dois grupos
- Abordagem mais comum:
  - √ Selecionar uma categoria como referência , comparando com todos os outros grupos

- Generalização do modelo para 3 grupos:

$$\ln\left(\frac{\mathrm{P}\{\mathrm{Grupo}=1|\mathbf{x}\}}{\mathrm{P}\{\mathrm{Grupo}=2|\mathbf{x}\}}\right) = \boldsymbol{\beta}_1'\mathbf{x}\,.$$

$$\ln\left(\frac{\mathrm{P}\{\mathrm{Grupo}=3|\mathbf{x}\}}{\mathrm{P}\{\mathrm{Grupo}=2|\mathbf{x}\}}\right) = \boldsymbol{\beta}_3'\mathbf{x}\,.$$

  - √ Probabilidades são comparadas separadamente com a da categoria de referência (Grupo = 2), através de seus coeficientes

Exemplo

- Cultivares de vinho
- Ajuste de modelo logístico multinomial
  - √ Covariáveis:
    - Alcohol, Ash, Alcal, Abs, proline
    - Julga-se que elas conduzirão a um bom ajuste
  - √ Categoria de referência: 2

```
> vinho$grupo <- as.factor(vinho$Class)
> vinho$rClass <- relevel(vinho$grupo, ref = 2)
    # define categoria de referencia
```


- Modelo:
  - √ rClass ~ Alcohol + Ash + Alcal + Abs + Proline:

```
> ajuste.logit <- multinom(rClass ~ Alcohol + Ash + Alcal + Abs + Proline,
+  data = vinho, maxit = 200)
> print(aj.lg <- summary(ajuste.logit), digits = 4)
Call:
multinom(formula = rClass ~ Alcohol + Ash + Alcal + Abs + Proline,
    data = vinho, maxit = 200)

Coefficients:
  (Intercept) Alcohol    Ash   Alcal    Abs  Proline
1     -124.00   6.213 22.849 -3.3478 10.354 0.029383
3      -46.46   2.927  6.192  0.3032 -7.483 0.005955

Std. Errors:
  (Intercept) Alcohol    Ash  Alcal    Abs  Proline
1      0.1139  1.6304 0.1403 1.0161 0.2612 0.032351
3      0.3588  0.4539 2.8564 0.2566 1.8766 0.003968

Residual Deviance: 26.65138
AIC: 50.65138
```

  - √ Coeficientes:
    - Alteração no log da chance, quando a cada unidade de variação explicativa

- Teste de significância dos coeficientes:
  - √ Estatística t e p-valor:

```
> razao.t <- aj.lg$coefficients / aj.lg$standard.errors
> # p-valores da t com gl = $edf
> print(2 * (1 - pt(abs(razao.t), df = aj.lg$edf)), digits = 4)
  (Intercept)   Alcohol     Ash    Alcal       Abs Proline
1           0 2.482e-03 0.00000 0.006404 4.308e-14  0.3816
3           0 3.169e-05 0.05101 0.260270 1.802e-03  0.1592
```

  - √ Distribuição multinomial ajustada a cada observação

```
> head(aj.lg$fitted.values)
             2         1            3
1 6.405531e-18 1.0000000 2.109756e-21
2 3.942623e-16 1.0000000 1.247350e-20
3 3.273230e-11 1.0000000 2.887328e-12
4 1.979074e-20 1.0000000 8.690855e-21
5 4.198875e-03 0.9944199 1.381191e-03
6 1.013497e-18 1.0000000 9.108334e-18
```

    - Cada linha soma 1

Outros Métodos de Discriminação

Métodos de Discriminação

- Método do vizinho mais próximo
- Classification and Regression Tress – CART
- Support Vector Machine – SVM
- Método dos núcleos estimadores
- Redes neurais artificiais

Método do Vizinho mais Próximo

- *Nearest neighbor discriminant analysis*
  - √ Não é um método paramétrico
    - Não depende da suposição de normalidade multivariada
- Procedimento de discriminação:
  - √ Encontra-se vizinho mais próximo
    - (distância de Mahalanobis)
  - √ Classifica-se a observação na população do vizinho
- Variação:
  - √ Método dos k vizinhos mais próximos

Árvores de Regressão e Classificação

- CART(Classification and Regression Tress)
  - √ Trata simultaneamente variáveis contínuas e não contínuas
- Procedimento de classificação:
  - √ Baseado em informações das distribuições isoladas de cada variável
  - √ Resulta numa árvore com vários nós

Support Vector Machine

- Grupo de algoritmos de classificação, incluem ampla variedade de modelos paramétricos e não paramétricos
  - √ Modelos lineares e métodos de regressão
  - √ Técnicas de suavização por núcleo
- Estimativa de probabilidade de classificação:
  - √ Validação cruzada ou por amostras de treinamento

Referências

Bibliografia Recomendada

- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- ZELTERMAN, D. *Applied Multivariate Statistics with R*. Springer, 2015.