



UFAM

MISTURAS FINITAS DE DENSIDADES BETA E DE DIRICHLET APLICADAS EM
ANÁLISE DISCRIMINANTE

Sarah Pinheiro Barbosa

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática

Orientador: José Raimundo Gomes Pereira

Manaus

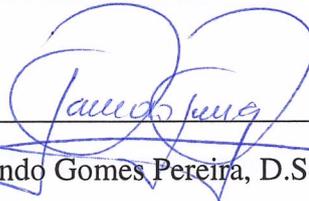
Março de 2018

MISTURAS FINITAS DE DENSIDADES BETA E DE DIRICHLET
APLICADAS EM ANÁLISE DISCRIMINANTE

Sarah Pinheiro Barbosa

DISSERTAÇÃO DE MESTRADO APRESENTADA AO PROGRAMA DE PÓS-GRADUAÇÃO
EM MATEMÁTICA, DA UNIVERSIDADE FEDERAL DO AMAZONAS, COMO PARTE DOS
REQUISITOS NECESSÁRIOS À OBTENÇÃO DO TÍTULO DE MESTRE EM MATEMÁTICA
(M.Sc).

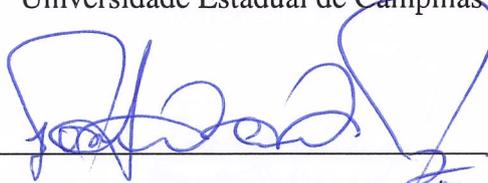
BANCA EXAMINADORA



Prof. José Raimundo Gomes Pereira, D.Sc. - Orientador
Universidade Federal do Amazonas



Prof. Larissa Avila Matos, Dr.Sc.
Universidade Estadual de Campinas



Prof. José Mir Justino da Costa, Dr.Sc.
Universidade Federal do Amazonas

MANAUS, AM- BRASIL

MARÇO de 2018

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

B238m Barbosa, Sarah Pinheiro
Misturas finitas de densidades beta e de dirichlet aplicadas em
análise discriminante / Sarah Pinheiro Barbosa. 2018
128 f.: il. color; 31 cm.

Orientador: José Raimundo Gomes Pereira
Dissertação (Mestrado em Matemática - Estatística) -
Universidade Federal do Amazonas.

1. Análise Discriminante. 2. Mistura Finita de Densidades. 3.
Distribuição Beta. 4. Distribuição Dirichlet. 5. Simulação
Computacional. I. Pereira, José Raimundo Gomes II. Universidade
Federal do Amazonas III. Título

*Este trabalho dedico à minha
família.*

Agradecimentos

Agradeço ao Meu Deus pela minha vida, pelo ar que respiro, por todas as dádivas recebidas, pelos dons que me deste, por ouvir minhas orações e pela força que me ergue a cada dia.

Ao meu pai, sempre calado no seu canto, porém, nunca deixou de incentivar, aconselhar e apoiar meus sonhos e planos. À minha mãe, pelas orações e seu amor, por dormir segurando a minha mão quando a minha vida parecia ter chegado ao fim. As minhas queridas irmãs Perla, Patrícia e Sabrina, que desde criança me ajudam a arquitetar sonhos merabolantes. Aos meus lindos sobrinhos Rebeca, Miguel, Rikelme, Querén e Benjamin, obrigada pelas alegrias, enfim, a todos da minha família, pois são meus alicerces.

Ao meu Orientador, pelas direções oferecidas, pelos ensinamentos, pelas oportunidades de desenvolvimento propiciadas e, principalmente, pelo exemplo de pessoa e profissional que certamente me guiará nesta nova etapa da minha vida.

Ao professor Diego Souza pela sua significativa contribuição no R do modelo proposto neste trabalho. A professora Maria Ivanilde por me inserir no ramo da pesquisa com o Programa de Iniciação Científica. A todos os professores do curso de Estatística da Universidade Federal do Amazonas por me proporcionar o conhecimento.

Agradeço aos colegas e amigos do curso de mestrado em Matemática, em especial, Renata, Alice, Regina e Guilherme, pelas angústias e alegrias compartilhadas, pela oportunidade de transpor a amizade para além do ambiente acadêmico, pela parceria, pelo apoio, pela confiança e pela contribuição em minha caminhada.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro nesses dois anos de estudos.

Enfim a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

“Não fui eu que ordenei a você? Seja forte e corajoso! Não se apavore nem desanime, pois o Senhor, o seu Deus, estará com você por onde você andar.”
(Josué:1-9).

Resumo da Dissertação apresentada ao Programa de Pós-Graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do grau de Mestre em Matemática. (M.Sc.)

MISTURAS FINITAS DE DENSIDADES BETA E DE DIRICHLET APLICADAS EM ANÁLISE DISCRIMINANTE

Sarah Pinheiro Barbosa

Março/2018

Orientador: José Raimundo Gomes Pereira

Linha de Pesquisa: Estatística

Em muitas aplicações de Análise Discriminante (AD) as observações das variáveis no vetor de características são confinadas ao intervalo $(0,1)$, por exemplo, classificação de pixels em imagens digitais. Neste trabalho, investigamos o emprego do Classificador de Bayes (CB) para estas aplicações, modelando as distribuições nas classes com emprego de Misturas Finitas de Densidades Betas e de Dirichlet. Para investigar e avaliar esta modelagem, desenvolvemos um estudo de simulação, analisando a estimação das densidades e dos parâmetros, assim como, as Taxas de Erro (TE) de classificação. Foram simulados problemas com diferentes estruturas, relativas ao número de componentes, tamanho do conjunto de treinamento, à sobreposição e distribuição das classes. Os resultados do estudo sugerem que os modelos avaliados são capazes de se ajustar aos diferentes problemas considerados, desde os mais simples aos mais complexos, em termos de modelagem das observações para fins de classificação. Com dados reais, situações onde desconhecemos as formas das distribuições nas classes, os CB's com os modelos implementados apresentaram TE razoáveis quando comparados a outros classificadores mais usuais. Como uma limitação, a modelagem apresenta melhores desempenhos com um número relativamente alto de observações no conjunto de treinamento.

Palavras-chave: Análise Discriminante, Mistura Finita de Densidades, Distribuição Beta, Distribuição Dirichlet, Simulação Computacional.

Abstract of Dissertation presented to Postgraduate in Mathematics, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

FINITE MIXTURES OF BETA AND DIRICHLET DENSITIES APPLIED IN
DISCRIMINANT ANALYSIS

Sarah Pinheiro Barbosa

March/2018

Advisor: José Raimundo Gomes Pereira

Research lines: Statistics

In many Discriminant Analysis (DA) applications the observations of the variables in the characteristic vector are confined on the interval $(0,1)$, p.e, pixel classification in digital images. In this work, we investigated the use of the Bayes Classifier (BC) for these applications, modeling the distributions in the classes using Finite Mixture Density Betas and the Dirichlet. To investigate and evaluate this model, we developed a simulation study, analyzing the estimation of densities and the parameters, as well as the Classification Error Rates (ER). Problems were simulated with different structures, relative to the number of components, training set size, overlap and class distribution. The results of the study suggest that the models evaluated are able to adjust to the different problems considered, from the simplest to the most complex, in terms of modeling observations for classification purposes. With real data, situations where the class distributions are unknown, the BC's with the implemented models presented reasonable TE when compared to other more usual classifiers. As a limitation, the modeling presents better performances with a relatively high number of observations in the training set.

Keywords: Discriminant Analysis, Finite Density Mixing, Beta Distribution, Dirichlet Distribution, Computational Simulation.

Sumário

Lista de Figuras	x
Lista de Tabelas	xiii
1 Introdução	1
1.1 O problema abordado em Análise Discriminante	1
1.2 Proposta do Trabalho	2
1.3 Objetivos	4
1.4 Estrutura do Trabalho	4
2 Análise Discriminante	5
2.1 Modelagem do Problema	5
2.2 Classificador de Bayes	8
2.2.1 Função Discriminante baseada em Modelo Normal	10
2.2.2 Naive Bayes	12
2.2.3 Função Discriminante empregando Mistura Finita de Densidades	14
2.3 Avaliação de classificadores	15
2.3.1 Conjunto de treinamento e teste	15
2.3.2 Validação cruzada	15
3 Mistura Finita de Densidades Betas, de Produtoário de Betas e de Dirichlet	17
3.1 Distribuição Beta	17
3.2 Distribuição Dirichlet	22
3.3 Misturas Finitas de Densidades	25
3.3.1 Estimação dos Parâmetros	28
3.3.2 Algoritmo EM para Mistura Finita de Densidades	29

3.4	Mistura Finita de Densidades de Produtório de Betas	34
3.5	Mistura Finita de Densidades de Dirichlet	37
3.6	Naive Bayes com Mistura Finita de Densidades Betas	39
3.7	Seleção do Número de Componentes	41
4	Estudos de Simulação	43
4.1	Estudo de Simulação 1:	44
4.1.1	Estudo 1 para o modelo MFPB	45
4.1.2	Estudo 1 para o modelo MFD	54
4.1.3	Análise do Estudo de Simulação 1	63
4.2	Estudo de Simulação 2:	64
4.2.1	Estudo 2 para o modelo MFPB	64
4.2.2	Estudo 2 para o modelo MFD	67
4.3	Estudo de Simulação 3:	71
4.3.1	Simulação de Amostras de MFD	72
4.3.2	Simulação de Amostras de MFPB	76
4.3.3	Simulação com Amostras de MFPB Transformadas	79
4.3.4	Simulação de Amostras de MFBI	84
4.3.5	Simulação com Amostras de MFD Transformadas	87
5	Aplicação em Dados Reais	93
5.1	Noções sobre o espaço de cores RGB	93
5.2	Estudo de Caso 1: Imagens RGB de Pele e Não-Pele	94
5.3	Estudo de Caso 2: Imagens RGB de Baciloscopia	98
6	Considerações Finais	105
	Referências Bibliográficas	110

Lista de Figuras

3.1	Função densidade de probabilidade Beta para diferentes valores de (α, β) .	18
3.2	Gráfico da densidade Dirichlet para diferentes valores de α_3 , fixando $\alpha_1 = 2$ e $\alpha_2 = 2$.	23
3.3	Mistura Finita de Betas com 2 componentes, com $p = 2$.	26
3.4	Mistura Finita de Dirichlet com 3 componentes, $p = 2$	26
4.1	Histograma, dispersão e correlação na Situação 1.	46
4.2	Plot Ordenado da <i>dif</i> da Situação 1.	47
4.3	Dispersão da densidade estimada e a verdadeira da Situação 1.	49
4.4	Histograma, dispersão e correlação da Situação 2.	51
4.5	Plot Ordenado da <i>dif</i> da Situação 2.	52
4.6	Dispersão da densidade estimada e a verdadeira da Situação 2.	54
4.7	Histograma, dispersão e correlação da Situação 3.	55
4.8	Plot Ordenado da <i>dif</i> da Situação 3.	56
4.9	Dispersão da densidade estimada e a verdadeira da Situação 3.	58
4.10	Histograma, dispersão e correlação da Situação 4.	59
4.11	Plot Ordenado da <i>dif</i> da Situação 4.	61
4.12	Dispersão da densidade estimada e a verdadeira da Situação 4.	62
4.13	Histograma, dispersão e correlação da MFPB.	65
4.14	Histograma, dispersão e correlação da MFD.	68
4.15	Histograma, dispersão e correlação da Estrutura 1.	73
4.16	Histograma, dispersão e correlação da Estrutura 2.	73
4.17	Histograma, dispersão e correlação da Estrutura 3.	73
4.18	Média e IC (95%) da TE para Estrutura 2.	74
4.19	Média e IC (95%) da TE para Estrutura 3.	74

4.20	Histograma, dispersão e correlação da Estrutura 4.	77
4.21	Histograma, dispersão e correlação da Estrutura 5.	77
4.22	Histograma, dispersão e correlação da Estrutura 6.	77
4.23	Média e IC (95%) da TE para Estrutura 4.	78
4.24	Média e IC (95%) da TE para Estrutura 5.	78
4.25	Média e IC (95%) da TE para Estrutura 6.	78
4.26	Histograma, dispersão e correlação da Estrutura 4 - Dados Transformados.	80
4.27	Histograma, dispersão e correlação da Estrutura 5 - Dados Transformados.	80
4.28	Histograma, dispersão e correlação da Estrutura 6 - Dados Transformados.	80
4.29	Média e IC (95%) da TE para Estrutura 4 - Dados Transformados.	81
4.30	Média e IC (95%) da TE para Estrutura 5 - Dados Transformados.	81
4.31	Média e IC (95%) da TE para Estrutura 6 - Dados Transformados.	81
4.32	Histograma, dispersão e correlação da Estrutura 7.	85
4.33	Histograma, dispersão e correlação da Estrutura 8.	85
4.34	Histograma, dispersão e correlação da Estrutura 9.	85
4.35	Média e IC (95%) da TE para Estrutura 7.	86
4.36	Média e IC (95%) da TE para Estrutura 8.	86
4.37	Média e IC (95%) da TE para Estrutura 9.	86
4.38	Histograma, dispersão e correlação da Estrutura 7 - Dados Transformados.	88
4.39	Histograma, dispersão e correlação da Estrutura 8 - Dados Transformados.	88
4.40	Histograma, dispersão e correlação da Estrutura 9 - Dados Transformados.	88
4.41	Média e IC (95%) da TE para Estrutura 7 - Dados Transformados.	89
4.42	Média e IC (95%) da TE para Estrutura 8 - Dados Transformados.	89
4.43	Média e IC (95%) da TE para Estrutura 9 - Dados Transformados.	89
5.1	a) Modelo RGB e b) Cubo RGB de 24-bit. Fonte: Gonzales & Woods (2002)	94
5.2	Gráfico 3D para Dados Pele e Não-Pele. (a)Transformação 1 e (b) Trans- formação 2.	95
5.3	Histograma, dispersão e correlação para Dados Pele e Não-Pele. (a) T1 e (b) T2.	95
5.4	Histogramas dos dados da classe Pele. (a) RT1, (b) GT1, (c) BT1, (d) RT2, (e) GT2 e (f) BT2.	96

5.5	Histogramas dos dados da classe Não-Pele. (a) RT1, (b) GT1, (c) BT1, (d) RT2, (e) GT2 e (f) BT2.	96
5.6	Ambiente para a aquisição das imagens. Fonte: Kimura Junior (2010). . .	99
5.7	Imagem obtida por microscopia com os bacilos marcados por um especialista. Fonte: Costa <i>et al.</i> (2008).	99
5.8	Gráfico 3D para Dados de Baciloscopia. (a) Transformação 1 e (b) Transformação 2.	100
5.9	Histograma, dispersão e correlação para Dados de Baciloscopia. (a) T1 e (b) T2.	100
5.10	Histogramas dos dados da classe Bacilo. (a) RT1, (b) GT1, (c) BT1, (d) RT2, (e) GT2 e (f) BT2.	101
5.11	Histogramas dos dados da classe Fundo. (a) RT1, (b) GT1, (c) BT1, (d) RT2, (e) GT2 e (f) BT2.	102

Lista de Tabelas

4.1	Estimativas dos parâmetros na Situação 1 para o Modelo 1.	46
4.2	Estimativas dos parâmetros da Situação 1 do Modelo 2.	46
4.3	Média, erro padrão e IC da <i>dif</i> da Situação 1.	48
4.4	Média da correlação (<i>r</i>) entre a densidade estimada e a verdadeira para Situação 1.	49
4.5	Estimativas dos parâmetros da Situação 2 do Modelo 1.	51
4.6	Estimativas dos parâmetros da Situação 2 do Modelo 2.	51
4.7	Média, erro padrão e IC da <i>dif</i> da Situação 2.	52
4.8	Média da correlação (<i>r</i>) entre a densidade estimada e a verdadeira para Situação 2.	53
4.9	Estimativas do parâmetros da Situação 3 do Modelo 1.	55
4.10	Estimativas do parâmetros da Situação 3 do Modelo 2.	56
4.11	Média, erro padrão e IC da <i>dif</i> da Situação 3.	57
4.12	Média da correlação (<i>r</i>) entre a densidade estimada e a verdadeira para Situação 3.	57
4.13	Estimativas do parâmetros da Situação 4 do Modelo 1.	60
4.14	Estimativas do parâmetros da Situação 4 do Modelo 2.	60
4.15	Média, erro padrão e IC da <i>dif</i> da Situação 4.	61
4.16	Média da correlação (<i>r</i>) entre a densidade estimada e a verdadeira para Situação 4.	62
4.17	Média das estimativas dos parâmetros para MFPB - Modelo 1.	66
4.18	Média das estimativas dos parâmetros para MFPB - Modelo 2.	66
4.19	Média das estimativas dos parâmetros para MFPB - Modelo 3	66
4.20	EQM das estimativas dos parâmetros para MFPB.	67
4.21	Média das estimativas dos parâmetros para MFD - Modelo 1.	69

4.22	Média das estimativas dos parâmetros para MFD - Modelo 2.	69
4.23	Média das estimativas dos parâmetros para MFD - Modelo 3.	69
4.24	EQM dos Parâmetros Estimados para MFD.	70
4.25	Média e desvio-padrão da TE para Simulações de Amostras de MFD. . .	75
4.26	Média e desvio-padrão da TE das Simulações de Amostras MFPB.	83
4.27	Média e desvio-padrão da TE das Simulações de Amostras MFPB - Dados Transformados (DT).	84
4.28	Média e desvio-padrão da TE das Simulações de Amostras MFB Independentes.	91
4.29	Média e desvio-padrão da TE das Simulações de Amostras MFB Independentes - Dados Transformados (DT).	92
5.1	TE de classificação dos Dados Pele e Não-Pele da T1.	97
5.2	TE de classificação dos Dados Pele e Não-Pele da T2.	98
5.3	Resultado da classificação dos dados da Baciloscopia da T1.	103
5.4	Resultado da classificação dos dados da Baciloscopia da T2.	103

Capítulo 1

Introdução

Neste capítulo apresentamos o problema abordado em Análise Discriminante, seus objetivos e exemplos de aplicações, descrevemos, também, a proposta do trabalho e sua organização.

1.1 O problema abordado em Análise Discriminante

Os problemas de Análise Discriminante (AD) são muito recorrentes na área da Estatística. Nesta classe de problema a questão a ser resolvida consiste em classificar *objetos* à *classes* previamente definidas. Os objetos podem ser pacientes em tratamento médico, plantas, assinatura em documentos, pixels em uma imagem digital, entre outros. As classes são categorias para quais os objetos devem ser classificados, por exemplo, paciente portador ou não de uma dada doença, diferentes espécies de plantas, assinatura verdadeira ou falsa, o pixel pertence ou não a uma região de interesse na imagem. Este tipo de problema ocorrem em várias áreas da ciência e tecnologia (veja por exemplo Hastie *et al.* (2009)). Outros exemplos são:

- Detectar tipos de poluição industrial, num espaço geográfico;
- Detectar células anormais em imagens digitais de amostras de sangue;
- Selecionar e-mails sem vírus;
- Identificar peças defeituosas em um processo de produção por meio de imagens digitais.

Na abordagem estatística para os problemas em AD, os objetos são descritos por um conjunto de variáveis, estas modeladas conjuntamente como um vetor aleatório, denominado *vetor de características*. Em uma abordagem paramétrica, em cada classe é atribuída uma distribuição de probabilidade para este vetor, sendo esta distribuição denominada *distribuição condicional da classe*. A abordagem tem por finalidade desenvolver uma regra estatística capaz de proceder a alocação dos objetos às classes definidas no problema. Esta regra/procedimento é denominado de *função discriminante* ou *classificador*, veja por exemplo (Johnson & Wichern (2012)).

Na prática, dispomos de observações do vetor de características para cada uma das classes, as quais são denominadas de *conjunto de treinamento*, sendo estas observações empregadas para estimar os parâmetros dos modelos adotados no problema em questão e, em consequência, estimar o classificador. Além disso, são também aplicados procedimentos com estas observações para avaliar a eficiência do classificador, inferindo sua capacidade de alocar futuras observações cujas classes sejam desconhecidas.

Uma questão relevante em aplicações de AD, é a escolha das variáveis para compor o vetor de características que sejam capazes de distinguir as classes no problema. Dependendo do problema abordado, estas variáveis podem ser observadas diretamente sobre o objeto em estudo mas, em outros casos, podem ser necessários procedimentos adicionais para obtenção das observações. Estas questões pertencem às atividades de *seleção e extração de características* em AD. Embora seja de importância fundamental para a prática da AD, esta questão não será considerada neste trabalho, será abordado apenas o problema de modelagem do vetor de características, a estimação das mesmas, bem como, procedimento de avaliação do classificador obtido. Para uma abordagem sobre seleção e extração de características veja, por exemplo, Theodoridis *et al.* (2010).

1.2 Proposta do Trabalho

Em muitas aplicações de AD as observações nas classes apresentam multimodalidade, assimetria e, muitas vezes, valores considerados extremos. Para estes casos, as misturas finitas de densidades são extremamente eficientes, sendo capazes de modelar distribuições desconhecidas arbitrariamente complexas (veja, por exemplo, McLachlan & Peel (2000), Cabral *et al.* (2012) e Prates *et al.* (2013)).

Vários trabalhos empregam mistura finita de densidades como modelos para as distribuições condicionais das classes em AD. Podemos citar Fraley & Raftery (2002) que empregam mistura finita de normais multivariadas; Andrews & McNicholas (2012) que empregam misturas finitas de t multivariadas; Coelho (2013) que apresenta um estudo sobre emprego de misturas finitas de distribuições normais assimétricas e de t assimétricas.

Dos trabalhos mencionados acima, as densidades das misturas têm suporte não limitado, no entanto, em muitas aplicações temos problemas onde as observações são obtidas, ou transformadas para estarem, em um intervalo limitado, em particular, no intervalo (0,1). Por exemplo, é comum em problemas de classificação de pixels em imagens digitais no espaço de cores RGB, transformar as intensidades dos pixels para o intervalo (0,1) (veja, por exemplo, Frery & Perciano (2013)).

Alguns trabalhos já abordaram o emprego de mistura finita de densidades em problemas com observações no intervalo (0,1), Ji *et al.* (2005) consideram uma mistura finita de densidades Beta, modelo univariado, com o algoritmo EM (Expectation and Maximization Algorithm), empregando para estimação, incluindo uma função de maximização no passo maximização do algoritmo. Em Bouguila *et al.* (2006) empregam estimação bayesiana com uma mistura finita de densidades Beta, empregando um modelo univariado. Ma & Leijon (2011) consideram problema com dados multivariados, adotando uma mistura finita onde as densidades componentes são produtos de densidades Beta, estes produtos sendo empregado para modelar a distribuição conjunta das variáveis no vetor aleatório, com o procedimento de estimação implementado com uma abordagem bayesiana. Nguyen & Wu (2015) também consideram uma mistura finita cujas densidades componentes são produtos de Beta, porém, apresentam uma abordagem bayesiana diferenciada para estimação dos modelos.

Os trabalhos citados no parágrafo anterior não apresentam estudos de simulação computacional para avaliar as densidades estimadas, que são os elementos necessários em abordagens de estimação de um classificador em AD. Desta forma, desenvolvemos um estudo sobre a modelagem das distribuições condicionais das classes em AD, com dados no intervalo (0,1), empregando mistura finita de produtos de Betas e, também, misturas finitas de densidades de Dirichlet. Restringimos a dimensão máxima do vetor de característica ao caso tridimensional, visando o emprego em classificação de pixels em

imagens digitais no espaço de cores RGB.

1.3 Objetivos

O presente trabalho tem como objetivo principal investigar a eficiência de misturas finitas de densidades Beta e de Dirichlet como modelos para distribuição nas classes em Análise Discriminante. Analisaremos a performance através de um estudo de simulação e aplicação em dados reais. Para isso, realizamos as seguintes etapas:

- Revisão de literatura sobre todos os conceitos envolvidos no desenvolvimento desse trabalho;
- Simulação computacional de problemas com especificadas estruturas de distribuições nas classes;
- Aplicações com dados reais da literatura.

Na seção seguinte será descrita a organização do trabalho.

1.4 Estrutura do Trabalho

O presente trabalho está organizado em 6 capítulos. O Capítulo 1 com a introdução. No Capítulo 2 apresentamos uma breve introdução dos principais métodos para análise discriminante, descrevendo detalhadamente o método de classificação de Bayes.

No Capítulo 3, apresentamos uma breve revisão das distribuições Beta e Dirichlet. Para estas distribuições, foram construídos modelos de mistura finita de densidades Betas, de Produto de Betas e de Dirichlet. Apresentamos também, a metodologia para estimação dos parâmetros via algoritmo EM.

Nos Capítulos 4 e 5 são apresentados os resultados obtidos por meio dos estudos de simulações e aplicações em dados reais, respectivamente. Capítulo 6 apresentamos as considerações finais.

Capítulo 2

Análise Discriminante

Neste capítulo, abordaremos a classificação de objetos empregando a modelagem estatística via classificador de Bayes, algumas abordagens paramétricas, como Análise Discriminante Linear, Análise Discriminante Quadrática, Naive Bayes com distribuição Normal e Análise Discriminante com Mistura Finita de Densidades. Abordamos, também, a metodologia estatística mais usual para avaliar a performance dos classificadores.

2.1 Modelagem do Problema

Análise Discriminante (AD) é uma técnica estatística para alocar objetos em *categorias* ou *classes* previamente definida, muito utilizada em problemas de Reconhecimento de Padrões Supervisionado (RPS). Entendemos por objeto ou observação algo de nosso interesse que possamos descrever e classificar. O objeto é descrito por uma coleção de variáveis que, consideradas conjuntamente, recebe a denominação de *vetor de características* (\mathbf{X}), onde $\mathbf{X} = (X_1, X_2, \dots, X_p)$. As p -variáveis são denominadas de *variáveis preditoras*.

Considere uma caixa de e-mails, onde não se admita *spam* (Sending and Posting Advertisement in Mass). Em breves palavras o *spam* eletrônico é o uso de sistemas de mensagens eletrônicas para enviar uma mensagem não solicitada, especialmente publicidade, além de enviar mensagens repetidamente no mesmo site. Empregando para isso um dispositivo de identificação baseado na análise das informações do texto e sobre o remetente. Após uma fase de reconhecimento dos primeiros e-mails e o armazenamento das correspondentes informações, então, todas as vezes que um novo *spam*, ocorre o dis-

positivo deve ser capaz de checar as informações do remetente do e-mail, comparando com informações armazenadas em seu banco de dados. Esse é o paradigma do RPS, onde o objetivo é o reconhecimento dos padrões existentes nas classes predefinidas, a fim de criar uma regra, ou função, que discrimine ou classifique um novo objeto, do qual não se tem a informação da classe, em uma das classes predeterminadas.

Na abordagem estatística o vetor de características é modelado por uma distribuição de probabilidade condicionada a cada uma das classes, denotadas por G_1, G_2, \dots, G_N . Em cada G_j , para $j = 1, \dots, N$, modelamos \mathbf{X} de acordo com uma função densidade $f_j(\cdot)$, adequada às variáveis preditoras, onde $f_j(\cdot)$ pode ser uma função de probabilidade ou uma função de densidade de probabilidade. Considere ainda que $P(G_j) = P(Y = j)$ denota a probabilidade de um objeto provir da G_j , para $Y \in \{1, 2, \dots, N\}$. As densidades $f_j(\cdot)$ são denominadas *densidades condicionais das classes* e as probabilidades $P(G_j)$ denominadas *probabilidades a priori das classes*. Desta forma, o vetor de características é modelado como um par aleatório (\mathbf{X}, Y) .

Definição 2.1.1. *Uma função discriminante ou classificador é qualquer função*

$$w : \mathcal{X} \longrightarrow \Omega,$$

onde $\mathbf{X} \in \mathcal{X}$ é o espaço das características e $\Omega = \{G_1, \dots, G_N\}$ é o conjunto de N classes às quais um objeto pode ser alocado.

Assim, dado um classificador $w(\cdot)$ e um objeto para o qual observamos $\{\mathbf{X} = \mathbf{x}, w(\mathbf{x}) = j\}$, significa que o objeto é alocado para a classe G_j (Johnson & Wichern (2012)).

Para a construção de um classificador torna-se necessário um critério para avaliá-lo. Um "bom classificador" produz poucos erros de classificação. Do ponto de vista estatístico, a probabilidade do erro de classificação deve ser pequena, e podemos avaliar tal probabilidade construindo uma função de perda resultante do processo de classificação.

Suponha que em uma ultrassonografia diagnosticou-se a presença de nódulos cancerígenos quando na verdade eram apenas cistos, a gravidade do erro logo seria um alívio, porém, o custo do erro de não diagnosticar uma doença onde de fato ela existe é muito maior. Assim, classificar um objeto que pertence a uma classe G_1 , como pertencente à outra classe, denotada por G_2 , pode representar um erro mais grave do que classificando o objeto de G_2 como pertencente à G_1 . O método de classificação de objetos deve-se sem-

pre levar em consideração as perdas associadas a uma classificação errada. Uma maneira usual de formalizar um critério para avaliar classificadores, é estabelecer uma *função de perda*.

Seja $\varepsilon(i, j)$ a perda decorrente de classificar um objeto da classe G_i para a classe G_j , isto é, a perda decorrente do processo de classificação. Sendo que $\varepsilon(i, i) = 0$, para $i = 1, 2, 3, \dots, N$.

Temos que $\varepsilon(i, j) = \varepsilon(i, w(\mathbf{x}) = j)$, é uma observação de uma variável aleatória, e adotamos o valor esperado desta variável aleatória como um critério para desenvolver um classificador. Considere a função de perda $\varepsilon(\cdot, \cdot)$, então, empregamos a definição de *função de risco* e do *risco médio* (Ripley (2007)).

Definição 2.1.2. A perda esperada como função de G_i é a **Função de Risco**, definida por

$$\begin{aligned} R(w, i) &= E\{\varepsilon(i, w(\mathbf{X}))|Y = i\} \\ &= \sum_{j \neq i}^N \varepsilon(i, j)P(w(\mathbf{X}) = j|Y = i). \end{aligned} \quad (2.1)$$

Definição 2.1.3. A perda total esperada como função das variáveis aleatórias \mathbf{X} e Y é denominada de **Risco Médio**, ou **Risco Total**, ou seja,

$$\begin{aligned} R(w) &= E\{R(w, Y)\} \\ &= \sum_{i=1}^N R(w, i)P(Y = i) \\ &= \sum_{i=1}^N \sum_{j \neq i}^N \varepsilon(i, j)P(w(\mathbf{X}) = j|Y = i)P(Y = i). \end{aligned} \quad (2.2)$$

A finalidade é obter um classificador capaz de fazer o *Risco Médio* ser mínimo, para uma dada *função de perda* $\varepsilon(\cdot, \cdot)$.

Em muitos casos reais o custo das perdas $\varepsilon(i, j)$ não são simples para serem definidos. Então, sendo possível considerar os custos das perdas iguais, adota-se uma função denominada *função de perda 0 – 1*. Na expressão 2.2, se $\varepsilon(i, j) = k$, onde k é uma constante, a otimização não depende deste valor k , assim,

$$\varepsilon(i, j) = \begin{cases} 0 & \text{se } i = j \\ 1 & \text{se } i \neq j \end{cases} \quad (2.3)$$

No caso particular da função de perda 0 – 1, temos que

$$R(w, i) = \sum_{i \neq j=1}^N P(w(\mathbf{X}) = j | Y = i) \quad (2.4)$$

e

$$R(w) = \sum_{i=1}^N \sum_{i \neq j=1}^N P(w(\mathbf{X}) = j | Y = i) P(Y = i). \quad (2.5)$$

Portanto, a função perda 0 – 1, onde $R(w, i)$ é a probabilidade de classificação errada dos objetos da classe G_i e $R(w)$ é a probabilidade total de classificação errada do classificador w , também denominado de *erro de classificação de w* .

2.2 Classificador de Bayes

Recapitulando, seja $Y \in \{1, 2, 3, \dots, N\}$ uma variável indicadora das classes para os objetos. Em cada classe G_j , $j = 1, 2, 3, \dots, N$, \mathbf{X} é modelado por uma distribuição $f_j(\cdot) = f(\cdot | Y = j)$ adequada às variáveis preditoras, e $P(Y = j)$ a probabilidade de um objeto provir da classe G_j . Com os termos definidos acima, dado $\mathbf{X} = \mathbf{x}$, empregamos o Teorema de Bayes para determinar a *probabilidade a posteriori* da classe G_j , que é dada por

$$P(Y = j | \mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x} | Y = j) P(Y = j)}{\sum_{l=1}^N f(\mathbf{x} | Y = l) P(Y = l)}, \quad j = 1, \dots, N. \quad (2.6)$$

Definição 2.2.1. O classificador de Bayes é definido por:

$$w^*(\mathbf{x}) = h \quad \text{se} \quad \sum_{i=1}^N \varepsilon(i, h) f(\mathbf{x} | Y = i) P(Y = i) = \min_j \sum_{i=1}^N \varepsilon(i, j) f(\mathbf{x} | Y = i) P(Y = i). \quad (2.7)$$

Caso ocorra que duas ou mais classes atinjam o mínimo, o objeto é alocado em qualquer uma das classes, aleatoriamente.

A ideia subjacente ao desenvolvimento do *classificador de Bayes* se dá quando observações de uma determinada classe tenham uma maior ocorrência do que a outra, denominada de classe com *maior prevalência*. Se de alguma forma temos informações sobre as probabilidades de ocorrências das classes, é razoável considerarmos na construção do classificador essas probabilidades.

Para todas as classes a regra (2.7) pode ser estabelecida de maneira equivalente como

$$w^*(\mathbf{x}) = h \quad \text{se} \quad \sum_{i=1}^N \varepsilon(i, h)P(Y = i|\mathbf{x}) = \min_j \sum_{i=1}^N \varepsilon(i, j)P(Y = i|\mathbf{x}). \quad (2.8)$$

Observe que as expressões para as probabilidades *a posteriori* das classes, em (2.6), tem o mesmo denominador, sendo necessário somente analisar o numerador. O teorema a seguir estabelece a principal propriedade para o classificador em (2.7) ou na forma de (2.8). O teorema a seguir estabelece que o classificador de Bayes é ótimo no sentido de minimizar a *Função de Risco Total*, para sua demonstração veja Ripley (2007).

Teorema 2.2.1. *Para uma dada função de perda $\varepsilon(\cdot, \cdot)$, o classificador w^* minimiza o risco total, ou seja, $R(w^*) \leq R(w)$ para qualquer classificador w .*

Empregando a função de perda 0 – 1, em (2.8)

$$\sum_{i=1}^N \varepsilon(i, h)P(Y = i|\mathbf{x}) = \sum_{h \neq i=1}^N P(Y = i|\mathbf{x}) = 1 - P(Y = h|\mathbf{x}), \quad (2.9)$$

de (2.9) teremos o mínimo se tomarmos a classe G_h para a qual $P(Y = h|\mathbf{x}) = f(\mathbf{x}|Y = h)P(Y = h)$ é um máximo, isto é, a classe com a *maior probabilidade a posteriori*, (veja Ripley (2007)). Portanto, o classificador de Bayes com função de perda 0 – 1 pode ser expresso como

$$w^*(\mathbf{x}) = h \quad \text{se} \quad \frac{f_h(\mathbf{x})P(Y = h)}{f(\mathbf{x})} = \max_j f_j(\mathbf{x})P(Y = j). \quad (2.10)$$

Em teoria, o risco $e^* = R(w^*)$, denominado *Risco de Bayes*, pode ser determinado se conhecermos as probabilidades $P(Y = j)$ e as distribuições $f_j(\cdot|Y = j)$, para $j = 1, 2, 3, \dots, N$. O valor do risco de Bayes serve como referência para comparação de classificadores, pois ele é o menor valor que pode ser atingido por qualquer classificador. No caso da função de perda 0 – 1, e^* é equivalente ao erro de classificação de w^* . Em outras palavras, consideramos um bom classificador aquele que possuir um Risco Médio mais próximo possível ao do classificador de Bayes.

Neste trabalho, vamos considerar o caso da função de perda 0 – 1 para todos os métodos discriminantes analisados, entre os quais, alguns classificadores mais usuais

na literatura, empregando abordagens paramétricas, como Análise Discriminante Linear, Análise Discriminante Quadrática e Naive Bayes Normal.

Em situações que não é possível estimar as probabilidades *a priori*, por meio da função de perda $(0 - 1)$, podemos empregar o classificador de Bayes considerando $P(Y = 1) = P(Y = 2) = \dots = P(Y = N)$, que pode ser interpretada como sendo *priori não informativa*, o que leva a forma para as probabilidades *a posteriori* de (2.6) dadas por

$$P(Y = j | \mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x} | Y = j)}{\sum_{l=1}^N f(\mathbf{x} | Y = l)}, \quad j = 1, \dots, N. \quad (2.11)$$

Em (2.11), na escolha do máximo $P(Y = j | \mathbf{x})$, vemos uma estrutura de máxima verossimilhança e, dessa maneira, o classificador

$$w^*(\mathbf{x}) = h \quad \text{se} \quad f(\mathbf{x} | Y = h) = \max_j f(\mathbf{x} | Y = j) \quad (2.12)$$

é denominado *classificador de máxima verossimilhança*.

Como em aplicações reais as distribuições condicionais $f(\cdot | Y = j)$ e as probabilidades *a priori* $P(Y = j)$ não são conhecidas, um procedimento é postular modelos de probabilidade para os dados e tais modelos têm um conjunto de parâmetros a serem estimados, constituindo uma abordagem paramétrica para obter estimativas para as funções $f(\cdot | Y = j)$. Outra abordagem consiste em não atribuir modelos paramétricos para as distribuições, onde para as $f(\cdot | Y = j)$ são consideradas estimadores não paramétricos de densidades. Outra abordagem, com o emprego da *regressão logística*, pode ser adotada, em que as probabilidades $P(Y = j | \mathbf{x})$ são estimadas diretamente sem estimar as condicionais $f(\cdot | Y = j)$, ver Hastie *et al.* (2009).

2.2.1 Função Discriminante baseada em Modelo Normal

Classificadores baseados em modelos normais predominam na literatura estatística por causa de sua simplicidade e boa eficiência, em aplicações reais (Johnson & Wichern (2012)). As densidades $f(\cdot | Y = j)$ são modeladas por distribuições Normais p-variadas com vetor de parâmetros $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, denotada por $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, para cada classe G_j , com vetor de médias $\boldsymbol{\mu}_j$ e matriz de covariâncias $\boldsymbol{\Sigma}_j$ não singular, onde $j = 1, 2, 3, \dots, N$,

ou seja,

$$f(\mathbf{x}; \boldsymbol{\theta}_j) = f(\mathbf{x}|Y = j) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\right\}. \quad (2.13)$$

Aplicando a função logarítmica no numerador de (2.6) e, com algumas manipulações algébricas, obtemos

$$\ln\{f(\mathbf{x})P(Y = j)\} = -\frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln P(Y = j). \quad (2.14)$$

Para (2.14), vamos considerar duas situações: uma em que as matrizes de covariâncias das classes são consideradas iguais (modelo homocedástico), e outra em que essas matrizes são supostamente diferentes (modelo heterocedástico).

No caso em que $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$, $j = 1, 2, 3, \dots, N$, temos que expandindo a forma quadrática em (2.14), multiplicando por -2 e desprezando-se os termos que são constantes para todas as classes, obtemos a quantidade

$$d_j^{Lin}(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_h)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_h) - 2 \ln P(Y = j). \quad (2.15)$$

Devido ao fato da função $d_j^{Lin}(\mathbf{x})$ ser linear em \mathbf{x} , essa regra é denominada *Análise Discriminante Linear* (ADL). Aplicando 2.15 na estrutura definida em 2.12, o classificador de Bayes com perda 0 – 1 é dado por

$$w^*(\mathbf{x}) = h \quad \text{se} \quad d_h^{Lin}(\mathbf{x}) = \max_j d_j^{Lin}(\mathbf{x}).$$

O primeiro termo no segundo membro em (2.15) é, por definição, a distância de Mahalanobis ao quadrado, entre \mathbf{x} e $\boldsymbol{\mu}_h$. Se as probabilidades *a priori* forem iguais para todas as classes, então, para um objeto com observação \mathbf{x} , a regra seleciona a classe cujo vetor de médias é o mais próximo de \mathbf{x} em termos da distância de Mahalanobis. Vemos também que, se a matriz $\boldsymbol{\Sigma}$ for proporcional a matriz identidade, essa proximidade pode ser mensurada em termos da distância Euclidiana.

Para o caso do modelo heterocedástico, com as manipulações algébricas mencionadas, obtemos a quantidade

$$d_j^{Qua}(\mathbf{x}) = \ln|\boldsymbol{\Sigma}_j| + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) - 2 \ln P(Y = j) \quad (2.16)$$

Empregando 2.16 na estrutura definida em 2.12, o classificador de Bayes com perda 0 – 1 é dado por

$$w^*(\mathbf{x}) = h \quad \text{se} \quad d_h^{Qua}(\mathbf{x}) = \max_j d_j^{Qua}(\mathbf{x}).$$

Devido a $d_j^{Qua}(\mathbf{x})$ ser uma função quadrática em \mathbf{x} , o classificador em (2.16) é denominado *Análise Discriminante Quadrática* (ADQ).

Para emprego das regras citadas, anteriormente, faz-se necessário estimar os parâmetros $\theta_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ e as probabilidades $P(Y = j)$. Em geral, são empregados os estimadores de Máxima Verossimilhança, onde n_j é o total de observações na classe G_j . Os estimadores são dados por

$$\widehat{P}(C_j) = \frac{n_j}{n} \quad (2.17)$$

$$\widehat{\boldsymbol{\mu}}_j = \frac{1}{n} \sum_{i=1}^{n_j} \mathbf{x}_{ji} \quad (2.18)$$

$$\widehat{\boldsymbol{\Sigma}}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\mathbf{x}_{ji} - \widehat{\boldsymbol{\mu}}_j)(\mathbf{x}_{ji} - \widehat{\boldsymbol{\mu}}_j)'. \quad (2.19)$$

Note que, $\widehat{\boldsymbol{\mu}}_j$ e $\widehat{\boldsymbol{\Sigma}}_j$ são, respectivamente, o *vetor de médias amostrais* e a *matriz de covariâncias amostrais* para a classe j , $j = 1, 2, 3, \dots, N$. E ainda, que no denominador de $\widehat{\boldsymbol{\Sigma}}_j$, $n_j - 1$, corrige o vício desse estimador. Para mais detalhes, veja Johnson & Wichern (2012).

2.2.2 Naive Bayes

Naive Bayes é uma técnica simples para a construção de classificadores em AD. A suposição "ingênua" que o classificador Naive Bayes faz é que as variáveis envolvidas na classificação são modeladas como independentes. Sabe-se, entretanto, que, na maioria dos casos, a suposição de independência entre as variáveis é falsa. Mesmo assim, o classificador Naive Bayes produz resultados bastante satisfatórios, para mais informações veja Domingos & Pazzani (1997).

Em Webb & Ting (2005), os autores fazem uma comparação entre diversos classificadores, em vários conjuntos de dados, bastante conhecidos na literatura, e obtêm resultados relevantes com relação ao Naive Bayes. Devido a sua simplicidade, eficiência e eficácia, tem sido usado como ferramenta para uma solução empírica do problema da alta

dimensionalidade. E ainda, de modo prático, é importante ressaltar que esta abordagem apresenta menos parâmetros a serem estimados, como por exemplo, no modelo Normal.

O Naive Bayes ainda pode ser ótimo, com a função de perda 0-1, mesmo com a pressuposição de independência para alguns problemas onde há um alto grau de dependência entre características, veja por exemplo Domingos & Pazzani (1997).

Seja $\mathbf{X} = (X_1, X_2, \dots, X_p)$ um vetor aleatório p -variado, o procedimento do Naive Bayes assume que (X_1, X_2, \dots, X_p) são independentes, assim

$$\begin{aligned} f_j(\mathbf{x}) &= f_j(x_1, x_2, \dots, x_p) \\ &= \text{(por independência entre as variáveis)} \\ &= g_{j1}(x_1)g_{j2}(x_2) \dots g_{jp}(x_p) \\ &= \prod_{d=1}^p g_j(x_d) \end{aligned}$$

Assim, de modo semelhante a ADL e ADQ, vamos considerar que $f(\mathbf{x})$ é a densidade Normal multivariada, mas com X_1, X_2, \dots, X_p independentes, então

$$f_j(\mathbf{x}; \boldsymbol{\theta}_j) = f(\mathbf{x}|Y = j) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\sigma}_j|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\}. \quad (2.20)$$

Como os X_i 's são independentes, implica que $\boldsymbol{\Sigma} = (\sigma_1^2 \sigma_2^2 \dots \sigma_p^2) \mathbf{I}_p$, onde \mathbf{I}_p é a matriz identidade de ordem p . Logo,

$$\begin{aligned} f_j(\mathbf{x}; \boldsymbol{\theta}_j) &= \frac{1}{(2\pi)^{\frac{p}{2}} |(\sigma_{1,j}^2, \dots, \sigma_{p,j}^2)' \mathbf{I}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)' [(\sigma_{1,j}^2, \dots, \sigma_{p,j}^2)' \mathbf{I}]^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\} \\ &= \frac{1}{(2\pi)^{\frac{1}{2}} \sigma_{1,j}} \exp \left\{ -\frac{(x_1 - \mu_{1,j})^2}{2\sigma_{1,j}^2} \right\} \times \dots \times \frac{1}{(2\pi)^{\frac{1}{2}} \sigma_{p,j}} \exp \left\{ -\frac{(x_p - \mu_{p,j})^2}{2\sigma_{p,j}^2} \right\} \\ &= g_{j1}(x_1)g_{j2}(x_2) \dots g_{jp}(x_p) \end{aligned} \quad (2.21)$$

onde $g_{jd}(x_d)$ é uma função densidade $N(\mu_{j,d}, \sigma_{j,d})$, com $d = 1, \dots, p$ e $j = 1, \dots, N$. Esse procedimento é denominado de Naive Bayes Normal (NBN). O conceito do Naive Bayes vai além de uma abordagem paramétrica, se estendendo também a abordagem não-paramétrica das densidades condicionais, não postulando nenhum modelo específico para essas densidades e as estimativas são obtidas diretamente com base no conjunto de treinamento com uma função núcleo, como exemplo, temos o classificador Naive Bayes com

distribuição Não-Paramétrica com função núcleo de Epanechnikov (NBE), para mais detalhes veja Hastie *et al.* (2009).

O modelo Naive Bayes é bastante flexível e admite a imposição de vários modelos às distribuições marginais, possibilitando a criação de vários classificadores, que será uma das propostas desse trabalho, o qual abordará um modelo Naive Bayes com mistura finita de densidades Betas, demonstrando então sua ampla aplicabilidade.

2.2.3 Função Discriminante empregando Mistura Finita de Densidades

Agora, nesta abordagem, consideramos cada classe G_j , $j = 1, 2, 3, \dots, N$, com densidade $f_j(\cdot|Y = j)$, tal que $f_j(\cdot|Y = j)$ é uma mistura finita de densidades com L_j componentes (ver McLachlan & Peel (2000)). Portanto, as classes são modeladas por

$$f_j(\mathbf{x}; \Theta_j) = \sum_{l=1}^{L_j} \rho_{jl} g_{jl}(\mathbf{x}|\theta_{jl}), \quad (2.22)$$

onde $\{\rho_{jl} > 0; \sum_{l=1}^{L_j} \rho_{jl} = 1\}$ e $\{g_{jl}(\cdot) \geq 0; \int_{-\infty}^{\infty} g_{jl}(\mathbf{x}) d\mathbf{x} = 1\}$, $\forall l = 1, \dots, L_j$. Os parâmetros ρ_{jl} são denominados pesos ou proporções da mistura e as densidades g_{jl} são chamadas de componentes da mistura.

Segundo Cabral *et al.* (2012), as misturas têm sido amplamente aplicadas em diversas áreas científicas como uma ferramenta para modelagem de heterogeneidade da população e aproximar densidades de probabilidades complicadas, apresentando assimetria, multimodalidade e caudas pesadas, sendo um método flexível para as suposições sobre a distribuição dos dados.

Neste trabalho, foram consideradas misturas finitas de densidades com componentes *Beta* e *Dirichlet*, em contexto multivariado. O objetivo é flexibilizar a modelagem em situações que a distribuição normal não se ajusta. Como mencionado na introdução a proposta deste trabalho é investigar o uso de misturas finitas de densidade para modelar as distribuições condicionais $f_j(\cdot|Y = j)$, considerando as componentes da mistura Produtório de Beta e de Dirichlet, assim, como um modelo Naive Bayes com mistura finita de densidades Betas. Estes modelos serão descritos com mais detalhes no próximo Capítulo 3.

2.3 Avaliação de classificadores

2.3.1 Conjunto de treinamento e teste

Em AD dispomos de amostras aleatórias que são identificadas com relação à classe de onde foram observadas. Este conjunto de observações "rotuladas", como mencionado, forma o *conjunto de treinamento*. Lembrando que os objetos são descritos por um par de variáveis (\mathbf{X}, Y) , temos que

$$\mathfrak{S}_{(n)} = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}.$$

O conjunto de treinamento é dividido em dois subconjuntos: *conjunto de treino* e *conjunto de teste*. O *conjunto de treino* é usado para estimar o classificador, isto é, estimar os parâmetros dos modelos empregados para as distribuições condicionais das classes. As observações no *conjunto de teste* são empregadas para avaliar o classificador, isto é, são submetidas ao classificador estimado para obtermos as taxas de erro de classificação.

Para um classificador $w(\cdot)$ construído com um conjunto de treinamento $\mathfrak{S}_{(n_{tre})}$, temos que a probabilidade de erro de classificação é denotada por $e_{(n_{tre})}^{(w)}$. Com este método, a ideia é estimar $e_{(n_{tre})}^{(w)}$ através da contagem de classificações erradas efetuadas por w em um conjunto de teste $\mathfrak{S}_{(n_{tes})}$, ou seja,

$$\hat{e}_{(n_{tre}, n_{tes})}^{(w)} = \frac{1}{n_{tes}} \sum_{j=1}^N \sum_{i=1}^{n_{tes}^j} I_{\{w(\mathbf{x}_{j,i}) \neq j\}}(\mathbf{x}_{j,i})$$

Então, temos que $\hat{e}_{(n_{tre}, n_{tes})}^{(w)}$ é a proporção de classificações erradas das observações no conjunto teste.

2.3.2 Validação cruzada

A motivação da estimação por meio da validação cruzada, é usar o máximo de observações na construção do classificador. O método de validação cruzada denominado *k-fold (k-fold cross-validation)*, consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste, os $(k - 1)$ restantes são utilizados para estimação dos parâmetros (conjunto de treino) e calcula-se a precisão do classificador. Quando $k = n$ (o número de

observações), a validação cruzada de "k- fold" é exatamente a validação cruzada "leave-one-out". Em cada repetição é calculado

$$\hat{e}_{i(n-1,1)}^{(w)} = I_{\{w(\mathbf{x}_{j,i}) \neq j\}}(\mathbf{x}_{j,i}) \quad i = 1, 2, \dots, n.$$

O estimador para o erro de classificação é dado por

$$\hat{e}_{(n)}^{(CV)} = \frac{1}{n} \sum_{i=1}^n \hat{e}_{i(n-1,1)}^{(w)}.$$

Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. Ao final das k iterações calcula-se a precisão sobre os erros encontrados, através da equação descrita anteriormente, obtendo assim uma medida mais confiável sobre a capacidade do classificador. Temos que $\hat{e}_{(n)}^{(CV)}$ é um estimador não tendencioso para $e_{(n-1)}$, veja Hastie *et al.* (2009).

Capítulo 3

Mistura Finita de Densidades Betas, de Produtório de Betas e de Dirichlet

Neste capítulo, definimos o modelo de mistura finita de densidades e suas propriedades. Abordamos a estimação dos parâmetros, com ênfase na estimação de Máxima Verossimilhança via algoritmo EM. A Mistura Finita de Produtório de Betas (MFPB) e Mistura Finita de Dirichlet (MFD) são abordadas de forma particularizada, assim, como o modelo baseado em Naive Bayes com Mistura Finita de Betas (NBMFB).

3.1 Distribuição Beta

Na estatística, a distribuição Beta pertence a uma família de distribuições de probabilidade contínuas definidas no intervalo $(0,1)$, parametrizada por dois parâmetros positivos, que aparecem como expoentes da variável aleatória e controlam a forma da distribuição, denotada por $Beta(\alpha, \beta)$. A distribuição Beta é uma forma muito versátil de representar resultados como proporções ou probabilidades.

Definição 3.1.1. *Seja X uma variável aleatória absolutamente contínua . Com suporte no intervalo aberto $(0,1)$. Dizemos que X tem uma distribuição Beta com parâmetros de forma (α, β) se sua função de densidade de probabilidade for dada por*

$$\begin{aligned} f_X(x; \alpha, \beta) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} I_{(0,1)}(x), \\ &= \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} I_{(0,1)}(x), \end{aligned} \quad (3.1)$$

onde $\Gamma(\cdot)$ é a função **Gamma** e $B(\cdot)$ é a função **Beta**.

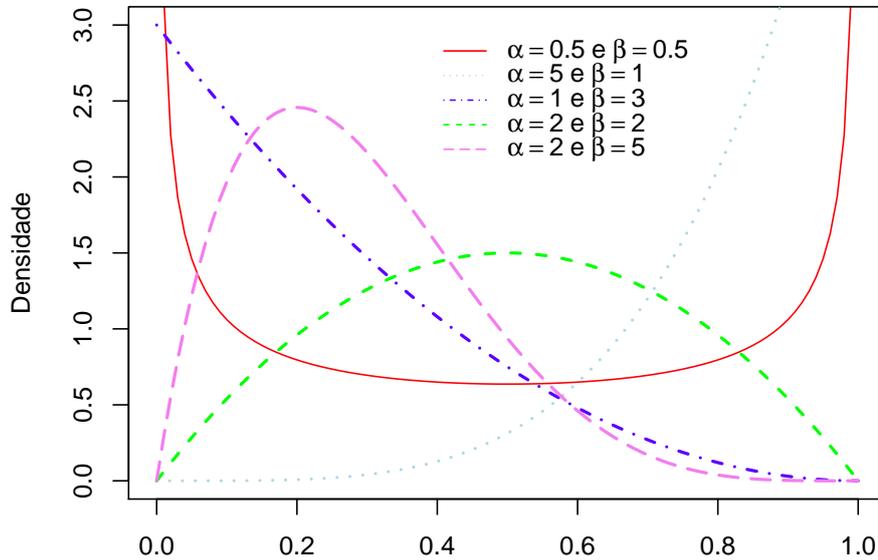


Figura 3.1: Função densidade de probabilidade Beta para diferentes valores de (α, β) .

A função *Beta* é estritamente relacionada a função *Gamma*. Uma das propriedades da função *Gamma* bastante utilizada é que $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. Fazendo uso de algumas propriedades da função *Gamma* para obtenção dos momentos da distribuição *Beta*, o cálculo torna-se bastante simples.

Pela definição de momentos, temos

$$\begin{aligned}
 E[X^t] &= \int_0^1 x^t f_X(x) dx \\
 &= \int_0^1 x^t \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\
 &= \frac{1}{B(\alpha, \beta)} \int_0^1 x^{(\alpha+t)-1} (1-x)^{\beta-1} dx \\
 &= \frac{1}{B(\alpha, \beta)} B(\alpha+t, \beta) \quad (\text{pela representação da integral da função Beta}) \\
 &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+t)\Gamma(\beta)}{\Gamma(\alpha + \beta + t)} \quad (\text{por definição da função Beta}) \\
 &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+t)}{\Gamma(\alpha + \beta + t)}. \tag{3.2}
 \end{aligned}$$

Fazendo $t = 1$ em 3.2, temos o valor esperado

$$\begin{aligned}
 E(X) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + \beta + 1)} \\
 &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\alpha \Gamma(\alpha)}{(\alpha + \beta) \Gamma(\alpha + \beta)} \quad (\text{propriedade da função } \Gamma) \\
 &= \frac{\alpha}{(\alpha + \beta)}. \tag{3.3}
 \end{aligned}$$

Também fazendo $t = 2$ em 3.2, obtemos o segundo momento,

$$\begin{aligned}
 E(X^2) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha + \beta + 2)} \\
 &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{(\alpha + 1) \alpha \Gamma(\alpha)}{(\alpha + \beta + 1)(\alpha + \beta) \Gamma(\alpha + \beta)} \quad (\text{propriedade da função } \Gamma) \\
 &= \frac{(\alpha + 1) \alpha}{(\alpha + \beta + 1)(\alpha + \beta)}. \tag{3.4}
 \end{aligned}$$

Então, poderemos facilmente calcular a variância,

$$\begin{aligned}
 \text{Var}(X) &= E(X^2) - [E(X)]^2 \\
 &= \frac{(\alpha + 1) \alpha}{(\alpha + \beta + 1)(\alpha + \beta)} - \frac{\alpha^2}{(\alpha + \beta)^2} \\
 &= \frac{\alpha \beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}. \tag{3.5}
 \end{aligned}$$

Notamos que a densidade *Beta* é apropriada para modelar proporções em virtude do seu domínio, o intervalo (0,1), e também devido a variedade de formas que a densidade pode assumir de acordo com os valores especificados para α e β , parâmetros da distribuição. Em especial se $\alpha = \beta = 1$, a densidade de Beta se reduz à uniforme (0,1).

Considere uma amostra aleatória X_1, X_2, \dots, X_n , de $X \sim \text{Beta}(\alpha, \beta)$. Os estimadores pelo método dos momentos para os dois parâmetros desconhecidos, denotados por $(\hat{\alpha}, \hat{\beta})$, podem ser obtidos empregando, a média amostral e variância amostral, como segue

$$\begin{aligned}
 \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i, \text{ e} \\
 \bar{V} &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.
 \end{aligned}$$

Fazendo, $E(X) = \bar{X}$ e $Var(X) = \bar{V}$, podemos estimar os parâmetros desconhecidos pela combinação de equações.

$$\bar{X} = \frac{\alpha}{(\alpha + \beta)}, \quad (3.6)$$

$$(\alpha + \beta) = \frac{\alpha}{\bar{X}}, \quad (3.7)$$

$$\beta = \frac{\alpha}{\bar{X}} - \alpha, \quad (3.8)$$

então,

$$\begin{aligned} \bar{V} &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\ \bar{V} &= \frac{\alpha}{(\alpha + \beta)} \frac{\beta}{(\alpha + \beta)(\alpha + \beta + 1)} \\ \left(\frac{\alpha}{\bar{X}} + 1\right) \frac{1}{(1 - \bar{X})} &= \frac{\bar{X}}{\bar{V}} \\ \frac{\alpha}{\bar{X}} + 1 &= \frac{\bar{X}(1 - \bar{X})}{\bar{V}} \\ \frac{\alpha}{\bar{X}} &= \frac{\bar{X}(1 - \bar{X})}{\bar{V}} - 1 \\ \hat{\alpha} &= \bar{X} \left[\frac{\bar{X}(1 - \bar{X})}{\bar{V}} - 1 \right] \end{aligned} \quad (3.9)$$

substituindo em (3.8),

$$\begin{aligned} \beta &= \frac{\bar{X} \left[\frac{\bar{X}(1 - \bar{X})}{\bar{V}} - 1 \right]}{\bar{X}} - \bar{X} \left[\frac{\bar{X}(1 - \bar{X})}{\bar{V}} - 1 \right] \\ \beta &= \left[\frac{\bar{X}(1 - \bar{X})}{\bar{V}} - 1 \right] - \bar{X} \left[\frac{\bar{X}(1 - \bar{X})}{\bar{V}} - 1 \right] \\ \hat{\beta} &= (1 - \bar{X}) \left[\frac{\bar{X}(1 - \bar{X})}{\bar{V}} - 1 \right]. \end{aligned} \quad (3.10)$$

Estimação dos Parâmetros por Máxima Verossimilhança

Conforme Casella (2002), as vantagens do uso do Estimador de Máxima Verossimilhança (EMV) são suas propriedades de suficiência, invariância e não ser viesado assintoticamente, entre outras.

Definição 3.1.2. *Seja X_1, X_2, \dots, X_n variáveis aleatórias independentes e identicamente distribuídas com distribuição dada por $f(\cdot; \Theta)$, com $\Theta \in \Omega$. A Função de Verossimilhança*

(FV) é definida como

$$\mathbb{L}(\Theta) = \mathbb{L}(\Theta | x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \Theta). \quad (3.11)$$

Definição 3.1.3. Se existe um único valor $\hat{\Theta} \in \Omega$ que maximiza a FV, então, $\hat{\Theta}$ é chamado de Estimador de Máxima Verossimilhança (EMV).

Seja X_1, X_2, \dots, X_n uma amostra aleatória, onde X é uma distribuição $Beta(\alpha, \beta)$.

A função de verossimilhança da expressão 3.1 é da forma

$$\mathbb{L}(\alpha, \beta; x_1, x_2, \dots, x_n) = \left(\frac{1}{B(\alpha, \beta)} \right)^n \left[\prod_{i=1}^n (x_i)^{\alpha-1} \right] \left[\prod_{i=1}^n (1-x_i)^{\beta-1} \right], \quad (3.12)$$

e a função logarítmica da verossimilhança é da forma

$$\log \mathbb{L}(\alpha, \beta; x_1, x_2, \dots, x_n) = -n \log B(\alpha, \beta) + (\alpha - 1) \sum_{i=1}^n \log(x_i) + (\beta - 1) \sum_{i=1}^n \log(1 - x_i).$$

Encontrar o máximo em relação a um parâmetro de forma envolve tomar a derivada parcial em relação ao parâmetro de forma e definir a expressão igual a zero resultando no estimador de máxima verossimilhança dos parâmetros de forma:

$$\begin{aligned} U(\alpha) &= \frac{\partial}{\partial \alpha} \log \mathbb{L}(\alpha, \beta; x_1, x_2, \dots, x_n) = -n \left(\frac{\partial}{\partial \alpha} \log B(\alpha, \beta) \right) + \sum_{i=1}^n \log(x_i) \\ U(\beta) &= \frac{\partial}{\partial \beta} \log \mathbb{L}(\alpha, \beta; x_1, x_2, \dots, x_n) = -n \left(\frac{\partial}{\partial \beta} \log B(\alpha, \beta) \right) + \sum_{i=1}^n \log(1 - x_i) \end{aligned}$$

onde:

$$\begin{aligned} \frac{\partial}{\partial \alpha} \log B(\alpha, \beta) &= -\frac{\partial}{\partial \alpha} \log \Gamma(\alpha + \beta) + \frac{\partial}{\partial \alpha} \log \Gamma(\alpha) + \frac{\partial}{\partial \alpha} \log \Gamma(\beta) = \psi(\alpha + \beta) + \psi(\alpha) \\ \frac{\partial}{\partial \beta} \log B(\alpha, \beta) &= -\frac{\partial}{\partial \beta} \log \Gamma(\alpha + \beta) + \frac{\partial}{\partial \beta} \log \Gamma(\alpha) + \frac{\partial}{\partial \beta} \log \Gamma(\beta) = \psi(\alpha + \beta) + \psi(\beta) \end{aligned}$$

O termo $\psi(\cdot)$ é conhecido como *função Digamma*, a qual é não linear, logo, a estimação por MV para dois parâmetros desconhecidos da distribuição $Beta(\alpha, \beta)$, não tem uma solução analítica. Portanto, é necessário métodos numéricos de maximização de funções, veja por exemplo Narayanan (1992).

3.2 Distribuição Dirichlet

A distribuição de Dirichlet, é uma família de distribuições de probabilidade multivariada contínuas, parametrizada por um vetor de parâmetros $\boldsymbol{\alpha} \in \mathbb{R}_+^p$, denotada por $Dir(\boldsymbol{\alpha})$. É uma generalização multivariada da distribuição Beta, podendo ser empregada no estudo da distribuição de vetores aleatórios, cuja as variáveis aleatórias estejam compreendidas no intervalo (0,1) e a soma é igual a 1, veja Kotz & Lovelace (1998).

Definição 3.2.1. *Seja $X = (X_1, \dots, X_p)$, um vetor aleatório definido no intervalo (0, 1), cuja as somas das coordenadas é 1, com vetor de parâmetros $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)$, onde $(\alpha_d > 0; d = 1, \dots, p)$. Dizemos que X tem distribuição Dirichlet se sua função de densidade de probabilidade for dada por*

$$f_X(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{d=1}^p \alpha_d\right)}{\prod_{d=1}^p \Gamma(\alpha_d)} \prod_{d=1}^{p-1} x_d^{\alpha_d-1} \left(1 - \sum_{d=1}^{p-1} x_d\right)^{\alpha_p-1}, \quad (3.13)$$

onde $x_p = 1 - \sum_{d=1}^{p-1} x_d$ é implicitamente definido em qualquer ponto do simplex

$$S_p = \left\{ (x_1, \dots, x_p) \in \mathbb{R}^p : 0 < x_d < 1; d = 1, \dots, p-1; \sum_{d=1}^{p-1} x_d < 1 \right\},$$

e assume zero em caso contrário.

Uma propriedade interessante é que as densidades marginais da distribuição *Dirichlet* são densidades *Betas*, veja Narayanan (1992). A função de momentos da distribuição *Dirichlet* (veja Kotz & Lovelace (1998)), é dada por

$$E(X_1^{r_1} \times \dots \times X_{p-1}^{r_{p-1}}) = \frac{\Gamma\left(\sum_{d=1}^p \alpha_d\right)}{\prod_{d=1}^p \Gamma(\alpha_d)} \frac{\Gamma(r_1 + \alpha_1) \times \dots \times \Gamma(r_{p-1} + \alpha_{p-1}) \Gamma(\alpha_p)}{\Gamma(r_1 + \alpha_1 + \dots + r_{p-1} + \alpha_{p-1} + \alpha_p)}. \quad (3.14)$$

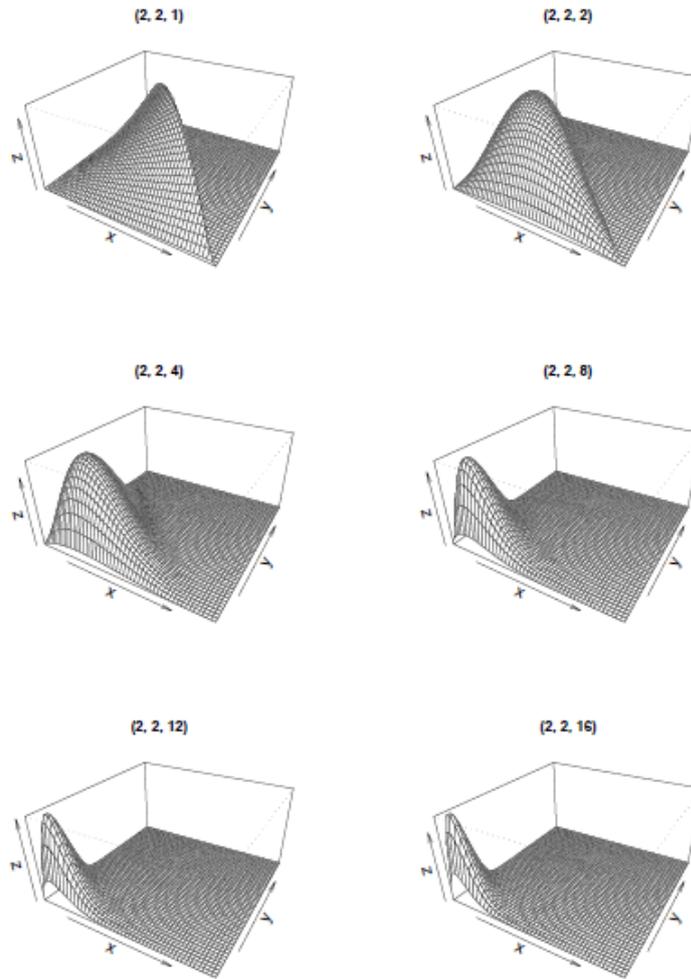


Figura 3.2: Gráfico da densidade Dirichlet para diferentes valores de α_3 , fixando $\alpha_1 = 2$ e $\alpha_2 = 2$.

A partir de (3.14) com $\delta = \sum_{d=1}^p \alpha_d$, temos algumas propriedades da distribuição *Dirichlet* onde temos que:

$$E(X_d) = \frac{\alpha_d}{\delta}, \quad d = 1, \dots, p; \quad (3.15)$$

$$\text{Var}(X_d) = \frac{\alpha_d(\delta - \alpha_d)}{\delta^2(\delta + 1)}, \quad d = 1, \dots, p; \quad (3.16)$$

$$\text{Cov}(X_d, X_q) = \frac{-\alpha_d \alpha_q}{\delta^2(\delta + 1)}, \quad d \neq q, \quad (d, q) = 1, \dots, p. \quad (3.17)$$

Estimação dos Parâmetros por Máxima Verossimilhança

Seja X_1, \dots, X_n uma amostra aleatória, onde $X \sim \text{Dir}(\boldsymbol{\alpha})$. Então, a *função de verossimilhança*, empregando a equação (3.13) é

$$\begin{aligned} \mathbb{L} &= \mathbb{L}(\boldsymbol{\alpha}; X_1, \dots, X_n) \\ &= \prod_{i=1}^n \left[\frac{\Gamma\left(\sum_{d=1}^p \alpha_d\right)}{\prod_{d=1}^p \Gamma(\alpha_d)} \prod_{d=1}^{p-1} X_{id}^{\alpha_d-1} \left(1 - \sum_{d=1}^{p-1} X_{id}\right)^{\alpha_p-1} \right]. \end{aligned} \quad (3.18)$$

De (3.18), as estatísticas suficientes para os parâmetros $\boldsymbol{\alpha}$ são as médias geométricas de X_{ip} , sintetizando todas as informações da amostra,

$$G_d = \left(\prod_{i=1}^n X_{id} \right)^{1/n}, \quad d = 1, \dots, p-1; \quad (3.19)$$

$$G_p = \left[\prod_{i=1}^n X_{id} \left(1 - \sum_{d=1}^{p-1} X_{id}\right) \right]^{1/n}. \quad (3.20)$$

A *função logarítmica da verossimilhança* é da forma

$$\log \mathbb{L} = \sum_{i=1}^n \left[\log \Gamma\left(\sum_{d=1}^p \alpha_d\right) - \sum_{d=1}^p \log \Gamma(\alpha_d) + \sum_{d=1}^p (\alpha_d - 1) \log G_d \right]. \quad (3.21)$$

Maximizando (3.21), obtemos os estimadores de máxima verossimilhança com respeito aos parâmetros α 's. Então, para α_d , temos

$$\frac{\partial}{\partial \alpha_d} \log \mathbb{L} = \sum_{i=1}^n \left\{ \frac{\partial}{\partial \alpha_d} \log \Gamma(\psi) - \frac{\partial}{\partial \alpha_d} \log \prod_{d=1}^p \Gamma(\alpha_d) + \sum_{d=1}^p (\alpha_d - 1) \log G_d \right\} \quad (3.22)$$

O valor de cada α_d , $d = 1, 2, \dots, p-1$, que torna 3.22 igual a zero é solução da equação

$$\log \Psi(\hat{\phi}) - \Psi(\hat{\alpha}_d) + G_d, \quad (3.23)$$

onde $\Psi(\hat{\phi}) = \frac{\partial}{\partial \alpha_d} \log \Gamma(\phi)$ e Ψ representa a *função Digamma*.

O sistema em (3.23) é não linear nos parâmetros α 's e não pode ser resolvido analiticamente, uma vez que a solução não possui forma fechada. É necessário, maximizar a função logarítmica da verossimilhança via um algoritmo iterativo, veja Narayanan (1992).

3.3 Misturas Finitas de Densidades

A proposta neste trabalho, como já mencionado, é modelar as distribuições condicionais $f_j(\cdot|Y = j)$ por mistura finita de densidades, considerando as componentes da mistura como Produto de Betas e de Dirichlet.

Definição 3.3.1. *Sejam g_1, g_2, \dots, g_L , funções de densidades e \mathbf{Y} um vetor aleatório, cuja função de densidade de probabilidade é dada por:*

$$h(\mathbf{y}) = \sum_{l=1}^L \rho_l g_l(\mathbf{y}), \quad (3.24)$$

onde $\rho_l > 0$, $l = 1, \dots, L$, $\sum_{l=1}^L \rho_l = 1$. Então, \mathbf{Y} é dito ter distribuição de mistura finita de densidades, com L componentes.

Os parâmetros ρ_1, \dots, ρ_L são denominados de *pesos* ou *proporções* da mistura e as densidades $g_1(\cdot), g_2(\cdot), \dots, g_L(\cdot)$ são chamadas de *componentes* da mistura.

Se as componentes $g_l(\cdot) \in \mathbb{G} = \{f : f(\mathbf{y}; \boldsymbol{\theta}), \mathbf{y} \in \mathbb{R}^p, \boldsymbol{\theta} \in \Theta\}$, uma família de densidade com espaço paramétrico Θ , a mistura pode ser reescrita como:

$$h(\mathbf{y}; \Phi) = \sum_{l=1}^L \rho_l g_l(\mathbf{y}; \boldsymbol{\theta}_l), \quad \mathbf{y} \in \mathbb{R}^p, \quad (3.25)$$

onde $\Phi = (\rho_1, \dots, \rho_L, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L)$.

Na maioria das aplicações, as componentes da mistura são membros de uma mesma família paramétrica de distribuições e, para esses casos, nós denotaremos a mistura finita de densidades em (3.25) por

$$h(\mathbf{y}; \Phi) = \sum_{l=1}^L \rho_l g(\mathbf{y}; \boldsymbol{\theta}_l), \quad \mathbf{y} \in \mathbb{R}^p. \quad (3.26)$$

Na Figura 3.3 temos observações bidimensionais de uma Mistura Finita de Betas

com 2 componentes.

$$h_1(\mathbf{x}; \Theta) = 0.6 * \text{Beta}(x_1; 5; 1) * \text{Beta}(x_2; 5; 5) + 0.4 * \text{Beta}(x_1; 1; 5) * \text{Beta}(x_2; 2; 8).$$

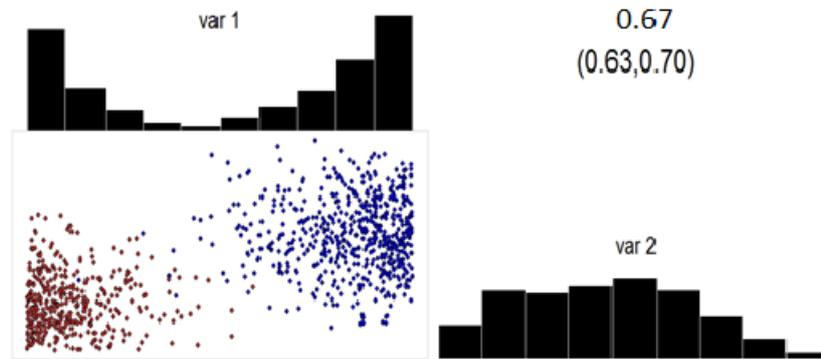


Figura 3.3: Mistura Finita de Betas com 2 componentes, com $p = 2$.

Na Figura 3.4 temos uma Mistura Finita de Dirichlet com 3 componentes, $p = 2$, com o modelo descrito por:

$$h_2(\mathbf{x}; \Theta) = 0,4 * \text{Dir}(\mathbf{x}; 10; 20) + 0,3 * \text{Dir}(\mathbf{x}; 5; 2) + 0,3 * \text{Dir}(\mathbf{x}; 20; 20).$$

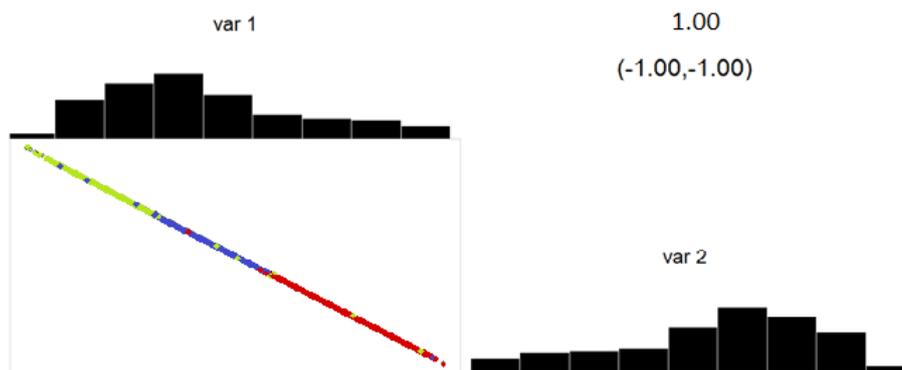


Figura 3.4: Mistura Finita de Dirichlet com 3 componentes, $p = 2$

Distribuições marginais em mistura finita de densidades com uma dada dimensão, também são mistura finita de mesma dimensão.

Identificabilidade

A condição de que a mistura finita seja identificável se faz necessária para que todos os parâmetros em Θ possam ser estimados de forma única. Em geral uma família paramétrica de funções densidades de probabilidades $f(\cdot, \Theta)$ é dita identificável se valores distintos de Θ determinam membros distintos da família de densidades. Isto é, considere a família de distribuições $\mathbb{F} = \{f(\cdot, \Theta) : \Theta \in \Omega\}$, onde Ω é o espaço paramétrico especificado. Então,

$$f(\cdot, \Theta) = f(\cdot, \Theta') \iff \Theta = \Theta'.$$

A identificabilidade no caso de misturas de densidades, apresenta uma característica mais específica, a permutabilidade, veja Titterington *et al.* (1985).

Seja $\mathbb{F} = \{\psi(\mathbf{y}; \theta) : \theta \in \Omega, \mathbf{y} \in \mathbb{R}^p\}$ uma família paramétrica de densidades e

$$\mathbb{P}\{f(\mathbf{y}; \Theta) : f(\mathbf{y}; \Theta) = \sum_{l=1}^L \rho_l \psi(\mathbf{y}; \theta_l), \quad \rho_l \geq 0,$$

$$\sum_{l=1}^L \rho_l = 1, \quad \psi(\mathbf{y}; \theta) \in \mathbb{F}, \quad \Theta = (\rho_1, \dots, \rho_L, \theta_1, \dots, \theta_L)\}$$

uma classe de família de MFD. A classe \mathbb{P} é dita identificável se, para quaisquer dois membros

$$f(\mathbf{y}; \Theta) = \sum_{l=1}^L \rho_l \psi(\mathbf{y}; \theta_l) \quad e \quad f(\mathbf{y}; \Theta') = \sum_{l=1}^L \rho'_l \psi(\mathbf{y}; \theta'_l),$$

tem-se que

$$f(\mathbf{y}; \Theta) = f(\mathbf{y}; \Theta')$$

e ainda podemos permutar os índices das componentes de forma que

$$\rho_l = \rho'_l \quad e \quad \psi(\mathbf{y}; \theta_l) = \psi(\mathbf{y}; \theta'_l), \quad l = 1, \dots, L.$$

A falta de identificabilidade não é preocupante, pois em AD trata-se de uma propriedade relativa ao modelo e não a algum método de estimação, porém se o modelo não é identificável a inferência pode ser dificultada, como exemplo, a estimação bayesiana, veja

em McLachlan & Peel (2000). Portanto, isto não se constitui um problema na estimação de máxima verossimilhança via algoritmo EM, principalmente, em aplicações onde o interesse é obter uma aproximação para o valor da densidade em observações específicas, como é o caso em AD.

3.3.1 Estimação dos Parâmetros

As duas abordagens de estimação dos parâmetros mais utilizadas são por *máxima verossimilhança* e *métodos Bayesianos*. Sendo a *estimação de máxima verossimilhança* (EMV) a mais popular em modelos de mistura de densidades (McLachlan & Peel (2000)).

A partir da definição em (3.12), temos que a *função de verossimilhança* para mistura finita de densidades é da forma

$$\begin{aligned} \mathbb{L}(\Theta | \mathbf{y}) &= \prod_{i=1}^n h(\mathbf{y}_i; \Theta) \\ \mathbb{L}(\Theta | \mathbf{y}) &= \prod_{i=1}^n \left\{ \sum_{l=1}^L \rho_l g_l(\mathbf{y}_i; \theta_l) \right\}, \quad \mathbf{y} \in \mathbb{R}^p. \end{aligned} \quad (3.27)$$

Na prática encontrar os EMV para Mistura Finita de Densidades, envolve várias dificuldades devido à complexidade da dependência da função de verossimilhança nos parâmetros. Existem dois inconvenientes inerentes associados ao problema de encontrar o máximo de uma função, pois pela falta de identificabilidade a função de verossimilhança atinge máximo local para diferentes valores de Θ , ou pode não ser limitada superiormente, então, o EMV para Θ pode não existir (McLachlan & Peel (2000)).

Por outro lado, como observado em Ripley (2007) (24,Cap.6), se o objetivo é estimar a densidade em pontos de interesse, como é o caso em AD, então não se constitui um problema. Sendo necessário somente obter uma "boa aproximação" para o valor estimado da densidade, logo, é possível encontrar uma estimativa para Θ que permita o emprego das Mistura Finita de Densidades para aproximarem as distribuições envolvidas no problema.

Diante de todas as dificuldades citadas, a função de verossimilhança (ou seu res-

pectivo log) não podem ser maximizadas analiticamente, porém, tendo como alternativa a estimação por máxima verossimilhança de forma iterativa. O algoritmo EM, neste contexto, é o mais utilizado para métodos numéricos de maximização de modelos de Mistura Finita de Densidades (McLachlan & Peel (2000)).

3.3.2 Algoritmo EM para Mistura Finita de Densidades

O algoritmo EM (do inglês "*Expectation and Maximization*"), é um método para encontrar estimativas de máxima verossimilhança dos parâmetros de uma distribuição de probabilidade.

No desenvolvimento do algoritmo EM, a ideia subjacente em estimação com dados incompletos, é considerar os dados observáveis (\mathbf{y}_i) como incompletos e aumentá-los com a inclusão de variáveis latentes (\mathbf{z}_i), variáveis não observáveis diretamente, de modo que a distribuição dos dados completos $\{(\mathbf{x}_i) = (\mathbf{y}_i, \mathbf{z}_i)\}$ simplifique as análises a serem desenvolvidas. Aplicando a ideia de estrutura de dados incompletos a Mistura Finita de Densidades, por hora, será tratada a questão de como gerar vetores pseudo aleatórios de uma mistura de densidades.

Seja $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ uma amostra aleatória de \mathbf{Y} , cuja distribuição é dada por uma Mistura Finita de Densidades,

$$h(\mathbf{y}; \Phi) = \sum_{l=1}^L \rho_l g_l(\mathbf{y}; \theta_l), \quad \mathbf{y} \in \mathbb{R}^p.$$

Seja \mathbf{Z}_i um vetor aleatório, onde $i = 1, \dots, n$, com $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{iL})$, onde as componentes da mistura são definidas como variáveis indicadoras da forma

$$Z_{il} = \begin{cases} 1, & \text{se } \mathbf{Y}_i \sim g_l(\mathbf{y}; \theta_l); \\ 0, & \text{se } \mathbf{Y}_i \sim g_m(\mathbf{y}; \theta_m), \quad m \neq l. \end{cases}$$

Neste contexto, os valores observados $\mathbf{y}_1, \dots, \mathbf{y}_n$ de $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, são considerados *dados incompletos*. Com os valores de $\mathbf{z}_1, \dots, \mathbf{z}_n$ para $\mathbf{Z}_1, \dots, \mathbf{Z}_n$, temos que a variável Z_{im}

como uma variável latente, não observável, associada ao vetor \mathbf{X}_i e indicando qual componente da mistura descreve sua distribuição. Obtemos os vetores de *dados completos* $\{\mathbf{x}_i = (\mathbf{y}_i, \mathbf{z}_i)\}$, para $i = 1, \dots, n$.

Consequentemente, sob essa abordagem o vetor aleatório \mathbf{Z}_i tem distribuição multinomial, considerando uma retirada em L categorias, com probabilidades $\boldsymbol{\rho}$. Portanto,

$$\mathbf{Z}_i \sim \text{Multinomial}(1; \boldsymbol{\rho}),$$

ou seja, para $\boldsymbol{\rho} = (\rho_1, \dots, \rho_L)$, temos

$$\begin{aligned} g(\mathbf{z}_i; \boldsymbol{\rho}) &= \rho_1^{z_{i1}} \times \dots \times \rho_L^{z_{iL}} \\ &= \prod_{l=1}^L \rho_l^{z_{il}}. \end{aligned} \quad (3.28)$$

Além disso, as proporções ρ_l denotam probabilidades *a priori* da i -ésima observação proveniente da l -ésima componente da mistura, com $i = 1, \dots, n$. Logo, a distribuição conjunta para os dados completos \mathbf{x}_i é da forma

$$\begin{aligned} h(\mathbf{x}_i; \Phi) &= g(\mathbf{y}_i | \mathbf{z}_i; \Theta) g(\mathbf{z}_i; \boldsymbol{\rho}) \\ &= \prod_{l=1}^L (g_l(\mathbf{y}_i; \boldsymbol{\theta}_l)^{z_{il}} g(\mathbf{z}_i; \boldsymbol{\rho})) \\ &= \prod_{l=1}^L (g_l(\mathbf{y}_i; \boldsymbol{\theta}_l)^{z_{il}} \rho_l^{z_{il}}). \end{aligned} \quad (3.29)$$

A função de verossimilhança para dados completos $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ é

$$\begin{aligned} \mathbb{L}(\Phi) = \mathbb{L}(\Phi | \mathbf{x}) &= \prod_{i=1}^n h(\mathbf{x}_i; \Phi) \\ &= \prod_{i=1}^n \prod_{l=1}^L g_l(\mathbf{y}_i; \boldsymbol{\theta}_l)^{z_{il}} \rho_l^{z_{il}}, \end{aligned} \quad (3.30)$$

e a função logarítmica da verossimilhança dos dados completos é da forma

$$\begin{aligned}\log \mathbb{L}(\Phi) &= \sum_{i=1}^n \sum_{l=1}^L z_{il} \{\log \rho_l + \log g_l(\mathbf{y}_i; \theta_l)\} \\ &= \sum_{i=1}^n \sum_{l=1}^L z_{il} \log \rho_l + \sum_{i=1}^n \sum_{l=1}^L z_{il} \log g_l(\mathbf{y}_i; \theta_l).\end{aligned}\quad (3.31)$$

O passo E do algoritmo EM consiste em determinar a função $Q(\Phi | \Phi^{(t)})$, como a esperança condicional do logaritmo da função de verossimilhança dos dados completos, condicionado aos dados observados e o t -ésimo passo ou ajuste para Φ , ou seja,

$$\begin{aligned}Q(\Phi | \Phi^{(t)}) &= E\{\log \mathbb{L}(\Phi) | \mathbf{y}, \Phi^{(t)}\} \\ &= \sum_{i=1}^n \sum_{l=1}^L E[Z_{il} | \mathbf{y}, \Phi^{(t)}] \ln \rho_l + \sum_{i=1}^n \sum_{l=1}^L E[Z_{il} | \mathbf{y}, \Phi^{(t)}] \ln g_l(\mathbf{y}_i; \theta_l^{(t)}),\end{aligned}\quad (3.32)$$

onde $\Phi^{(t)} = (\rho_1^{(t)}, \dots, \rho_L^{(t)}, \theta_1^{(t)}, \dots, \theta_L^{(t)})$ é uma aproximação para Φ .

De (3.32), vemos que é necessário determinar

$$\lambda_{il}^{(t)} \stackrel{\text{def}}{=} E[Z_{il} | \mathbf{y}, \Phi^{(t)}] = Pr[Z_{il} = 1 | \mathbf{y}, \Phi^{(t)}] = Pr[Z_{il} = 1 | \mathbf{y}_i, \Phi^{(t)}].\quad (3.33)$$

Considerando a distribuição dada em (3.28), temos que

$$Pr[z_{il} = 1 | \Phi^{(t)}] = Pr[z_{il} = 1, z_{ij} = 0, \forall j \neq l | \Phi^{(t)}] = \rho_l^{(t)}.\quad (3.34)$$

A partir de (3.29), temos que

$$g(\mathbf{y}_i | \mathbf{Z}_i, \Phi^{(t)}) = \prod_{l=1}^L [g_l(\mathbf{y}_i; \theta_l^{(t)})]^{z_{il}},\quad (3.35)$$

e, portanto, temos que

$$g(\mathbf{y}_i | Z_{il} = 1; \Phi^{(t)}) = g_l(\mathbf{y}_i; \theta_l^{(t)}).\quad (3.36)$$

Assim, usando o Teorema de Bayes, vemos que

$$\lambda_{il}^{(t)} = Pr[Z_{il} = 1 | \mathbf{y}_i, \Phi^{(t)}] = \frac{Pr[Z_{il} = 1 | \Phi^{(t)}] g(\mathbf{y}_i | Z_{il} = 1, \Phi^{(t)})}{g(\mathbf{y}_i; \Phi^{(t)})}.\quad (3.37)$$

Fazendo uso dos resultados de (3.34) e (3.36) em (3.37), obtemos

$$\lambda_{il}^{(t)} = \frac{\rho_l^{(t)} g_l(\mathbf{y}_i; \boldsymbol{\theta}_l^{(t)})}{\sum_{T=1}^L \rho_T^{(t)} g_T(\mathbf{y}_i; \boldsymbol{\theta}_T^{(t)})}. \quad (3.38)$$

Então, $\lambda_{il}^{(t)}$ é uma estimativa da probabilidade da observação \mathbf{y}_i provir da componente $g_l(\cdot; \boldsymbol{\theta}_l^{(t)})$ da mistura, com base em uma dada estimativa $\Phi^{(t)}$ do vetor de parâmetros Φ . Aplicando (3.38) em (3.32), podemos escrever

$$\begin{aligned} Q(\Phi | \Phi^{(t)}) &= \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \ln \rho_l + \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \ln g_l(\mathbf{y}_i; \boldsymbol{\theta}_l) \\ &= Q_1(\boldsymbol{\rho}) + Q_2(\boldsymbol{\Theta}), \end{aligned} \quad (3.39)$$

onde $\boldsymbol{\rho} = (\rho_1, \dots, \rho_L)$ e $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L)$.

O passo M é escolher $\Phi^{(t+1)}$ que maximize $Q(\Phi | \Phi^{(t)})$, ou seja,

$$Q_1(\boldsymbol{\rho}) = \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \ln \rho_l \quad (3.40)$$

e

$$Q_2(\boldsymbol{\Theta}) = \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \ln g_l(\mathbf{y}_i; \boldsymbol{\theta}_l). \quad (3.41)$$

Podemos aqui maximizar os dois termos de forma independente, considerando que eles não são relacionados.

$$\frac{\partial Q_1(\boldsymbol{\rho})}{\partial \boldsymbol{\rho}} = \mathbf{0} \quad \text{e} \quad \frac{\partial Q_2(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \mathbf{0}.$$

Para $Q_1(\boldsymbol{\rho})$ a maximização tem uma solução única, que é explicitamente determinada, independente da forma funcional das componentes da mistura. Embora temos que considerar as restrições sobre $\boldsymbol{\rho}$, onde os ρ_l são não negativos e somam 1, a solução é obtida com o emprego de multiplicadores de Lagrange, sendo dada por (ver Pereira *et al.* (2001)).

$$\rho_l^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \lambda_{il}^{(t)}, \quad l = 1, 2, \dots, L. \quad (3.42)$$

Para $Q_2(\boldsymbol{\Theta})$ a maximização, com relação a $\boldsymbol{\Theta}$, vemos que a solução depende da

forma funcional das componentes $f_l(\mathbf{y}_i; \theta_l)$, pois devemos solucionar

$$\sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \frac{\partial}{\partial \Theta} \ln g_l(\mathbf{y}_i; \theta_l) = \mathbf{0}. \quad (3.43)$$

Para a equação (3.43), dependendo da forma de $g_l(\mathbf{y}_i; \theta_l)$, nem sempre há uma solução analítica fechada para obter valores de θ_l .

Podemos, portanto sumarizar os passos do algoritmo EM para Mistura Finita de Densidades da seguinte forma:

1. Passo E: para $\Phi = \Phi^{(t)}$, calcular $Q(\Phi|\Phi^{(t)})$, tal que

$$Q(\Phi|\Phi^{(t)}) = E[\log \mathbb{L}(\Phi | \mathbf{x}) | \mathbf{y}; \Phi^{(t)}];$$

2. Passo M: Obter $\Phi^{(t+1)}$ maximizando $Q(\Phi|\Phi^{(t)})$.

Uma propriedade importante do algoritmo EM é que a forma como as aproximações $\Phi^{(t)}$ são determinadas, garante que $Q(\Phi^{(t+1)}|\Phi^{(t)}) \geq Q(\Phi^{(t)}|\Phi^{(t)})$. Gerando uma sequência monotônica, a qual implica em (ver McLachlan & Peel (2000))

$$\log \mathbb{L}(\Phi^{(t+1)}) \geq \log \mathbb{L}(\Phi^{(t)}), \quad (3.44)$$

portanto, as aproximações $\log \mathbb{L}(\Phi^{(t)})$ obtidas pelo algoritmo EM geram uma sequência $\{\log \mathbb{L}(\Phi^{(t)})\}$ não-decrescente.

Os passos E e M são repetidos até que um critério de convergência adequado seja atingido. Neste trabalho, o critério de convergência adotado é da forma

$$\left| \frac{\log \mathbb{L}(\Phi^{(t+1)})}{\log \mathbb{L}(\Phi^{(t)})} - 1 \right| < \varepsilon, \quad (3.45)$$

onde $|a|$ indica o valor absoluto de a , com $\varepsilon > 0$ e suficientemente pequeno. Neste trabalho usamos $\varepsilon = 10^{-6}$.

3.4 Mistura Finita de Densidades de Produtório de Betas

Supondo que $\mathbf{X} = (X_1, \dots, X_p)$ seja um vetor aleatório *i.i.d.*, para $d = 1, \dots, p$ temos que $X_d \sim \text{Beta}(\alpha_d, \beta_d)$. Seja \mathbf{X} uma mistura finita de densidades, com L componentes, logo, a sua função de densidade é da forma

$$f(\mathbf{x}; \Phi) = \sum_{l=1}^L \rho_l \mathbb{B}(\mathbf{x}; \theta_l) \quad \text{para} \quad \rho_l > 0; \sum_{l=1}^L \rho_l = 1 \quad (3.46)$$

onde,

$$\mathbb{B}(\mathbf{x}; \theta_l) = \prod_{d=1}^p \text{Beta}(x_d; \alpha_{ld}, \beta_{ld})$$

com $\Phi = \{\theta_1, \dots, \theta_L, \rho_1, \dots, \rho_L\}$ e $\theta_l = \{(\alpha_{l1}, \beta_{l1}), \dots, (\alpha_{lp}, \beta_{lp})\}$, $l = 1, \dots, L$. Ou seja, $(\alpha_{ld}, \beta_{ld})$ é um vetor de parâmetros de uma distribuição Beta da l -ésima componente para a variável X_d . A função definida em (3.46) será denominada de Mistura Finita de Densidades de Produtório de Betas (MFPB). No caso particular de $p = 1$, temos que X é univariada, sendo denominada de Mistura Finita de Densidades Betas (MFB).

Estimação dos parâmetros da MFPB

Para estimação de parâmetros do modelo MFPB, empregamos o método da máxima verossimilhança e solucionamos as equações deste método, implementando um algoritmo tipo EM.

Seja o conjunto de dados observados *i.i.d.* $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, com função de verossimilhança dada por

$$\mathbb{L}(\Phi | \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n \sum_{l=1}^L \rho_l \mathbb{B}(\mathbf{x}_i; \theta_l) \quad (3.47)$$

e a função log-verossimilhança é

$$\log \mathbb{L}(\Phi) = \sum_{i=1}^n \log \left[\sum_{l=1}^L \rho_l \mathbb{B}(\mathbf{x}_i; \theta_l) \right] \quad (3.48)$$

O logaritmo da função de verossimilhança para dados completos, pelo resultado

(3.31), é dado por

$$\begin{aligned}
\log \mathbb{L}_c(\Phi) &= \sum_{i=1}^n \sum_{l=1}^L z_{il} \log(\rho_l) + \sum_{i=1}^n \sum_{l=1}^L z_{il} \log(\mathbb{B}(\mathbf{x}_i; \theta_l)) \\
&= \sum_{i=1}^n \sum_{l=1}^L z_{il} \log(\rho_l) + \sum_{i=1}^n \sum_{l=1}^L z_{il} \left[\log \prod_{d=1}^p \text{Beta}(x_{di}; \alpha_{ld}, \beta_{ld}) \right] \\
&= \sum_{i=1}^n \sum_{l=1}^L z_{il} \log(\rho_l) + \sum_{i=1}^n \sum_{l=1}^L z_{il} \left[\sum_{d=1}^p \log \text{Beta}(x_{di}; \alpha_{ld}, \beta_{ld}) \right] \quad (3.49)
\end{aligned}$$

No passo E, obtemos a esperança condicional de $Q(\Phi | \Phi^{(t)}) = E_{\Phi^{(t)}} \{ \log \mathbb{L}_c(\Phi | \mathbf{X}) | \mathbf{X} \}$, substituindo em (3.38), temos que

$$\lambda_{il}^{(t)} = \frac{\rho_l^{(t)} \left[\prod_{d=1}^p \text{Beta}(x_{di}; \alpha_{ld}^{(t)}, \beta_{ld}^{(t)}) \right]}{\sum_{T=1}^L \rho_T^{(t)} \left[\prod_{d=1}^p \text{Beta}(x_{di}; \alpha_{Td}^{(t)}, \beta_{Td}^{(t)}) \right]}. \quad (3.50)$$

Portanto, o passo E do algoritmo é determinar a esperança condicional de

$$Q(\Phi | \Phi^{(t)}) = \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \log(\rho_l) + \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \log \left[\prod_{d=1}^p \text{Beta}(x_{di}; \alpha_{ld}, \beta_{ld}) \right],$$

ou seja,

$$Q_1(\boldsymbol{\rho}) = \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \log(\rho_l)$$

e

$$Q_2(\Theta) = \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \log \left[\prod_{d=1}^p \text{Beta}(x_{di}; \alpha_{ld}, \beta_{ld}) \right],$$

A partir da Seção 3.1 sabemos que a função $Q_2(\Theta)$ não pode ser resolvida de forma analítica, então, a equação será solucionada por meio de um método de otimização baseado no algoritmo de Newton-Raphson. Como, temos restrições quanto aos parâmetros do modelo MFPB, utilizamos o método L-BFGS-B de Byrd *et al.* (1995), o qual limitamos que as estimativas assumam somente valores positivos. Neste trabalho, implementamos

essa etapa da maximização no software R Core Team (2017), através da função *optim*.

Para a inicialização do algoritmo EM, aplicamos o método *K-means* para primeiro agrupar os dados e estimar os parâmetros para cada componente da mistura. Os valores iniciais de $\theta_l^{(0)} = \{(\alpha_{l1}^{(0)}, \beta_{l1}^{(0)}), \dots, (\alpha_{lp}^{(0)}, \beta_{lp}^{(0)})\}$, são calculados a partir das seguintes equações (ver artigo de Narayanan (1992)),

$$W'_{ld} = \frac{1}{n} \sum_{i=1}^n X_{ild} \quad (3.51)$$

e

$$W''_{ld} = \frac{1}{n} \sum_{i=1}^n X_{ild}^2. \quad (3.52)$$

Então, temos para $l = 1, \dots, L$ e $d = 1, \dots, p$

$$\hat{\alpha}_{ld}^{(0)} = \frac{(W'_{ld} - W''_{ld})W'_{ld}}{W''_{ld} - (W'_{ld})^2}$$

e

$$\hat{\beta}_{ld}^{(0)} = \frac{(W'_{ld} - W''_{ld})(1 - W'_{ld})}{W''_{ld} - (W'_{ld})^2}.$$

O valor inicial de $\rho_l^{(0)} = \frac{\text{n}^\circ \text{ de objetos alocados na componente } l}{n}$.

No passo M, vamos encontrar $\Phi^{(t+1)}$, maximizando $Q(\Phi | \Phi^{(t)})$. Atualize $\rho_l^{(t)}$ e $\theta_l^{(t)}$, com

$$\rho_l^{(t+1)} = \frac{1}{n} \sum \lambda_{ld}^{(t)}$$

e

$$\theta_l^{(t+1)} = \{(\alpha_{l1}^{(t+1)}, \beta_{l1}^{(t+1)}), \dots, (\alpha_{lp}^{(t+1)}, \beta_{lp}^{(t+1)})\}$$

onde $(\alpha_{ld}^{(t+1)}, \beta_{ld}^{(t+1)})$ são pares de parâmetros de distribuições Betas independentes. As iterações são repetidas até que o critério de convergência seja satisfeito, que neste trabalho consideramos como

$$\left| \frac{\log \mathbb{L}_c(\Phi^{(t+1)})}{\log \mathbb{L}_c(\Phi^{(t)})} - 1 \right| < 10^{-6}.$$

3.5 Mistura Finita de Densidades de Dirichlet

Seja $\mathbf{X} = (X_1, \dots, X_p)$ um vetor aleatório de uma mistura finita de densidades Dirichlet, com L componentes, sua função de densidade é dada por

$$f(\mathbf{x}; \Phi) = \sum_{l=1}^L \rho_l \text{Dir}(\mathbf{x}; \boldsymbol{\alpha}_l) \quad \text{para} \quad \rho_l > 0; \sum_{l=1}^L \rho_l = 1 \quad (3.53)$$

com $\Phi = \{\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_L, \rho_1, \dots, \rho_L\}$ e $\boldsymbol{\alpha}_l = \{\alpha_{l1}, \dots, \alpha_{lp}\}$, $l = 1, \dots, L$. Ou seja, (α_{ld}) é dos parâmetros de uma distribuição Dirichlet da l -ésima componente para a variável X_d com $d = 1, \dots, p$. A função definida em (3.53) será denominada de Mistura Finita de Densidades Dirichlet (MFD).

Estimação dos parâmetros da MFD

Para estimação de parâmetros do modelo MFD, empregamos o método da máxima verossimilhança via algoritmo EM.

Seja \mathbf{X} o conjunto de dados observados *i.i.d.* $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, com FV dada por

$$\mathbb{L}(\Phi | \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n \sum_{l=1}^L \rho_l \text{Dir}(\mathbf{x}_i; \boldsymbol{\alpha}_l) \quad (3.54)$$

e a função log-verossimilhança dada por

$$\log \mathbb{L}(\Phi) = \sum_{i=1}^n \log \left[\sum_{l=1}^L \rho_l \text{Dir}(\mathbf{x}_i; \boldsymbol{\alpha}_l) \right] \quad (3.55)$$

Então, a log-verossimilhança para os dados completos pelo resultado (3.31), é

$$\begin{aligned} \log \mathbb{L}_c(\Phi) &= \sum_{i=1}^n \sum_{l=1}^L z_{il} \log(\rho_l) + \sum_{i=1}^n \sum_{l=1}^L z_{il} \log(\text{Dir}(\mathbf{x}_i; \boldsymbol{\alpha}_l)) \\ &= \sum_{i=1}^n \sum_{l=1}^L z_{il} \log(\rho_l) + \sum_{i=1}^n \sum_{l=1}^L z_{il} \log [\text{Dir}(\mathbf{x}_i; \alpha_{i1}, \dots, \alpha_{ip})] \end{aligned} \quad (3.56)$$

Assim, pelos resultados (3.38), (3.40) e (3.41), temos que

$$\lambda_{il}^{(t)} = \frac{\rho_l^{(t)} \left[\text{Dir}(\mathbf{x}_i; \alpha_{l1}^{(t)}, \dots, \alpha_{lp}^{(t)}) \right]}{\sum_{T=1}^L \rho_T^{(t)} \left[\text{Dir}(\mathbf{x}_i; \alpha_{T1}^{(t)}, \dots, \alpha_{Tp}^{(t)}) \right]}, \quad (3.57)$$

$$Q_1(\boldsymbol{\rho}) = \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \log(\rho_l)$$

e

$$Q_2(\Theta) = \sum_{i=1}^n \sum_{l=1}^L \lambda_{il}^{(t)} \log \left[\text{Dir}(\mathbf{x}_i; \alpha_{l1}, \dots, \alpha_{lp}) \right],$$

A partir da Seção 3.2, sabemos que a função $Q_2(\Theta)$ não pode ser resolvida de forma analítica, então, a equação será solucionada por meio de um método de otimização baseado no algoritmo de Newton-Raphson. Como, temos restrições quanto aos parâmetros do modelo MFD, utilizamos o método L-BFGS-B de Byrd *et al.* (1995), o qual limitamos que as estimativas assumam somente valores positivos, através da função *optim* no software R Core Team (2017).

Para a inicialização do algoritmo EM, aplicamos o método *K-means* para primeiro agrupar os dados e estimar os parâmetros para cada componente da mistura. Os valores iniciais de $\boldsymbol{\alpha}_l^{(0)} = \{\alpha_{l1}^{(0)}, \dots, \alpha_{lp}^{(0)}\}$, onde W'_{ld} e W''_{ld} são da forma dada em 3.51 e 3.52. Então, temos para $l = 1, \dots, L$ e $d = 1, \dots, p$,

$$\hat{\alpha}_{ld}^{(0)} = \frac{(W'_{ld} - W''_{ld})W'_{ld}}{W''_{ld} - (W'_{ld})^2}, \quad \text{com } d = 1, 2, \dots, p-1$$

e

$$\hat{\alpha}_{lp}^{(0)} = \frac{(W'_{ld} - W''_{ld}) \left(1 - \sum_{d=1}^{p-1} W'_{ld} \right)}{W''_{ld} - (W'_{ld})^2}.$$

O valor inicial de $\rho_l^{(0)} = \frac{n^\circ \text{ de objetos alocados na componente } l}{n}$.

No passo M, vamos encontrar $\Phi^{(t+1)}$, maximizando $Q(\Phi | \Phi^{(t)})$. Atualize $\rho_l^{(t)}$ e

$\alpha_i^{(t)}$, com

$$\alpha_i^{(t+1)} = \{\alpha_{i1}^{(t+1)}, \dots, \alpha_{ip}^{(t+1)}\}$$

onde $(\alpha_{i1}^{(t+1)}, \dots, \alpha_{ip}^{(t+1)})$ é um vetor de parâmetros de uma distribuição Dirichlet. As iterações são repetidas até que o critério de convergência dado em (3.45) seja satisfeito.

3.6 Naive Bayes com Mistura Finita de Densidades Betas

Seja \mathbf{X} um vetor aleatório p -variado, com (X_1, \dots, X_p) independentes, onde cada X_d , $d = 1, \dots, p$, é uma Mistura Finita de Densidades Betas (MFB), com L_d componentes. Seja também $\Psi = \{\Phi_1, \dots, \Phi_p\}$, então, a função de densidade é da forma

$$\begin{aligned} f(\mathbf{x}; \Psi) &= MFB(\Phi_1) * MFB(\Phi_2) * \dots * MFB(\Phi_p) \\ &= h(x_1; \Phi_1)h(x_2; \Phi_2) \dots h(x_p; \Phi_p) \\ &= \prod_{d=1}^p h(x_d; \Phi_d) \end{aligned} \quad (3.58)$$

onde cada X_d é dado por

$$h(x_d; \Phi_d) = \sum_{l=1}^{L_d} \rho_{dl} \text{Beta}(x_d; \alpha_{dl}, \beta_{dl}), \quad \text{para } \rho_{dl} > 0; \sum_{l=1}^{L_d} \rho_{dl} = 1,$$

sendo que, $\Phi_d = \{(\alpha_{d1}, \beta_{d1}) \dots, (\alpha_{dL_d}, \beta_{dL_d}), \rho_{d1}, \dots, \rho_{dL_d}\}$, para $l = 1, \dots, L_d$. Logo, temos que, $(\alpha_{dl}, \beta_{dl})$ é um vetor de parâmetros da variável X_d pertencente a l -ésima componente de uma mistura finita de densidades Betas. A função definida em (3.58) será denominada Naive Bayes com Mistura Finita de Densidades Betas (NBMFB).

Estimação dos parâmetros do NBMFB

Para estimação de parâmetros do modelo NBMFB, basta considerar o modelo MFPB no caso $p = 1$, empregamos o método da máxima verossimilhança e solucionamos as equações deste método, implementando um algoritmo tipo EM para cada variável X_d , com $d = 1, 2, \dots, p$. Para X_d , sejam $\{x_{d1}, \dots, x_{dn}\}$, as observações da amostra aleatória,

então, a função de verossimilhança é dada

$$\mathbb{L}(\Phi_d | x_{d1}, \dots, x_{dn}) = \prod_{i=1}^n \sum_{l=1}^{L_d} \rho_{dl} \text{Beta}(x_{di}; \alpha_{dl}, \beta_{dl}) \quad (3.59)$$

e a função log-verossimilhança é

$$\log \mathbb{L}(\Phi_d) = \sum_{i=1}^n \log \left[\sum_{l=1}^{L_d} \rho_{dl} \text{Beta}(x_{di}; \alpha_{dl}, \beta_{dl}) \right].$$

O logaritmo da função de verossimilhança para dados completos, pelo resultado (3.31), é da forma

$$\log \mathbb{L}_c(\Phi_d) = \sum_{i=1}^n \sum_{l=1}^{L_d} z_{pil} \log(\rho_{dl}) + \sum_{i=1}^n \sum_{l=1}^{L_d} z_{dil} \log[\text{Beta}(x_{di}; \alpha_{dl}, \beta_{dl})] \quad (3.60)$$

Portanto, para cada variável X_d , com observações $\mathbf{x}_d = (x_{d1}, x_{d2}, \dots, x_{dn})$, serão realizados os passos do algoritmo EM. Pelos resultados (3.38), (3.40) e (3.41), temos que

$$\lambda_{dil}^{(t)} = \frac{\rho_{dl}^{(t)} [\text{Beta}(x_d; \alpha_{dl}^{(t)}, \beta_{dl}^{(t)})]}{\sum_{s=1}^{L_d} \rho_{ds}^{(t)} [\text{Beta}(x_d; \alpha_{ds}^{(t)}, \beta_{ds}^{(t)})]}, \quad (3.61)$$

$$Q_1(\boldsymbol{\rho}_d) = \sum_{i=1}^n \sum_{l=1}^{L_d} \lambda_{dil}^{(t)} \log(\rho_{dl})$$

e

$$Q_2(\Theta_d) = \sum_{i=1}^n \sum_{l=1}^{L_d} \lambda_{dil}^{(t)} \log[\text{Beta}(x_{di}; \alpha_{dl}, \beta_{dl})].$$

Como, temos restrições quanto aos parâmetros do modelo MFPB, consequentemente implica no modelo NBMFB, logo, utilizamos o método L-BFGS-B de Byrd *et al.* (1995), o qual implementamos o algoritmo no software R, através da função *optim*.

Aplicamos o método *K-means*, assim agrupamos os dados e a partir daí estimamos

os parâmetros para cada componente da mistura da variável \mathbf{X}_d . Os valores iniciais de $\theta_d^{(0)} = \{(\alpha_{d1}^{(0)}, \beta_{d1}^{(0)}), \dots, (\alpha_{dL_d}^{(0)}, \beta_{dL_d}^{(0)})\}$, para $l = 1, \dots, L_d$, são da forma dada em 3.51 e 3.52 Então,

$$\hat{\alpha}_{dl}^{(0)} = \frac{(W'_{dl} - W''_{dl})W'_{dl}}{W''_{dl} - (W'_{dl})^2}$$

e

$$\hat{\beta}_{dl}^{(0)} = \frac{(W'_{dl} - W''_{dl})(1 - W'_{dl})}{W''_{dl} - (W'_{dl})^2}.$$

O valor inicial de $\rho_{dl}^{(0)} = \frac{\text{n}^\circ \text{ de observações alocados na componente (grupo) } l}{n}$.

No passo M, vamos encontrar $\Phi_d^{(t+1)}$, maximizando $Q(\Phi_d | \Phi_d^{(t)})$. As iterações são repetidas até que o critério de convergência dado em 3.45 seja satisfeito.

Então, pelo modelo MFB, serão obtidos os parâmetros estimados para cada variável X_d de forma independente. Assim, temos que para o modelo NBMFB a densidade conjunta de \mathbf{X} é da forma

$$\hat{f}(\mathbf{x}) = MFB(x_1; \hat{\Phi}_1) * MFB(x_2; \hat{\Phi}_2) * \dots * MFB(x_p; \hat{\Phi}_p). \quad (3.62)$$

3.7 Seleção do Número de Componentes

Para selecionar o número de componentes considere uma função-critério cuja otimização indique o número de componentes do modelo adequado aos dados, isto é

$$\hat{L} = \arg \min_L \{C(\hat{\Theta}_{(L)}), L = L_{min}, \dots, L_{max}\},$$

onde $C(\hat{\Theta}_{(L)})$ é o valor da função-critério para o modelo estimado com dimensão L , no caso de mistura finita de densidades, o modelo com L componentes. Segundo Andrews & McNicholas (2012), um dos critérios de seleção de modelos mais empregado é o Critério de Informação Bayesiano, que denotaremos por BIC, devido ao termo em inglês *Bayesian Information Criterion*, que será adotado neste trabalho.

O desenvolvimento para obter o valor da função-critério está descrito em Basso

et al. (2009), sendo o resultado da forma

$$BIC(L) = -2l(\hat{\Theta}_L) + \kappa(L) \log n, \quad (3.63)$$

onde $l(\hat{\Theta}_{(.)})$ é o *logaritmo da função de máxima verossimilhança* para o modelo com vetor de parâmetros estimados $\hat{\Theta}_{(L)}$, e $\kappa(L)$ é o número de parâmetros livres no modelo de dimensão L . Portanto, o procedimento seleciona o modelo com o número de componentes que apresenta o menor valor para o BIC.

O critério BIC assintoticamente tende a selecionar o modelo de dimensão correta, por isso é denominado *consistente em ordem*, veja Ripley (2007). Apesar de não ter sido desenvolvido para modelos de misturas, na prática este critério tem apresentado resultados satisfatórios para selecionar o número de componentes da mistura ao aproximar densidades desconhecidas empregando misturas finitas de densidades.

No próximo capítulo são apresentados os estudos de simulação para observar o comportamento e desempenho dos modelos MFPB, MFD e NBMFB, comparando com outros classificadores mais conhecidos da literatura em AD.

Capítulo 4

Estudos de Simulação

Neste capítulo apresentaremos as diversas simulações realizadas. Serão descritos os modelos empregados para simulação, bem como, a análise das estimativas obtidas para os parâmetros do modelo e para as estimativas dos valores pontuais das densidades.

Três estudos foram realizados para observar o comportamento da estimação dos modelos propostos neste trabalho, assim como, avaliar o desempenho dos classificadores abordados neste trabalho, comparando com outros classificadores mais conhecidos da literatura de AD. Os estudos desenvolvidos foram:

Estudo 1: Análise do comportamento da densidade estimada para diferentes dimensões e componentes;

Estudo 2: Análise do comportamento dos parâmetros estimados para diferentes graus de separação entre as componentes da mistura;

Estudo 3: Análise dos modelos num contexto de classificação, com ênfase para o comportamento dos classificadores para diferentes graus de sobreposição entre as classes.

Para selecionar o número de componentes das misturas, empregamos o Critério de Informação Bayesiano (BIC). Os estudos 1, 2 e 3 foram realizados para os Modelos de

Misturas Finitas de Produtório de Densidades Betas (MFPB) e Misturas Finitas de Densidades de Dirichlet (MFD). No Estudo 3 foi incluído também o Naive Bayes com Mistura Finita de Densidade Betas (NBMFD). Todos os estudos de simulação foram desenvolvidos utilizando o software R Core Team (2017), empregando o pacote *gtools*, veja Warnes *et al.* (2015) e a função "*optim*" na implementação do algoritmo EM.

4.1 Estudo de Simulação 1:

Este estudo tem o objetivo de avaliar o comportamento da densidade estimada para diferentes dimensões e componentes, variando o tamanho de amostras para estimação dos parâmetros dos modelos. As simulações foram realizadas para MFPB e estimadas via MFPB, assim como, simulações para MFD e estimadas via MFD. Os passos das simulações são descritos a seguir.

- 1) Escolha da dimensão, do número de componentes e dos parâmetros.
- 2) Estimativa dos parâmetros via EM para diferentes tamanhos de amostra treino:
 $n = 100, 1.000, 5.000$ e 10.000 .
- 3) Gerar uma amostra de $n=1.000$ novas observações e gerar a densidade "verdadeira".
- 4) Obter o módulo da diferença relativa (*dif*) entre a densidade verdadeira ($f(\mathbf{x})$) e a densidade estimada ($\hat{f}(\mathbf{x})$):

$$dif = \left| \frac{\hat{f}(\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})} \right| * 100.$$

- 5) Repetir o processo 1.000 vezes.

As estimativas dos parâmetros serão apresentadas para uma avaliação superficial do seu comportamento. Os gráficos mostram a dispersão e histograma dos modelos de misturas simulados, gráficos da *dif*, e seus respectivos histogramas, ressaltamos que quanto mais próximos do valor zero, melhor será a estimação da densidade.

4.1.1 Estudo 1 para o modelo MFPB

Como definida na Seção 3.4 empregamos o modelo MFPB, dado por (3.46). Resaltamos que neste estudo o principal objetivo é observar o comportamento da diferença entre a densidade estimada e a "real", esta obtida com os parâmetros fixados em relação a variação do número de componentes, levando em consideração o tamanho da amostra treino.

Situação 1 : MFPB para observações em 2-dimensões

Nesta Situação 1 consideramos um caso simples em análise de dados multivariados, o conjunto de dados é de dimensão $p = 2$, isto é, $\mathbf{x}' = (x_1, x_2)$. Após serem fixados os parâmetros, simulamos dois conjuntos de dados variando o número de componentes, sendo um com 2-componentes e outro com 3-componentes.

Modelo 1:

$$f^1(\mathbf{x}; \Theta) = 0,6 * Beta(x_1; 5; 1) * Beta(x_2; 5; 5) + 0,4 * Beta(x_1; 1; 8) * Beta(x_2; 2; 8).$$

Modelo 2:

$$f^2(\mathbf{x}; \Theta) = 0.33 * Beta(x_1; 5; 1) * Beta(x_2; 5; 5) + 0.34 * Beta(x_1; 1; 8) * Beta(x_2; 2; 8) \\ + 0.33 * Beta(x_1; 5; 5) * Beta(x_2; 3; 1).$$

A seguir os gráficos de dispersão e histograma, gerados com 1.000 observações dos dois modelos são apresentados, assim, como, a correlação entre as variáveis. A intenção deste estudo é verificar a qualidade de estimação da densidade quanto a variação do número de componentes.

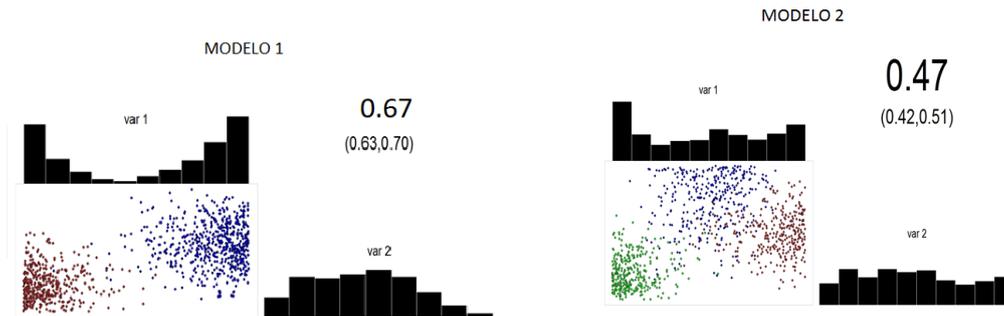


Figura 4.1: Histograma, dispersão e correlação na Situação 1.

Na Figura 4.1 é possível verificar a reprodução das misturas de densidades tanto no Modelo 1 quanto no Modelo 2 por meio dos histogramas de cada variável. Pelo gráfico de dispersão podemos visualizar as componentes nos modelos. No Modelo 1 o maior valor da correlação entre as variáveis é de 0,67 enquanto que no Modelo 2 é 0,47.

Para uma análise inicial apresentamos a estimação dos parâmetros para cada modelo, referente aos resultados de uma única repetição do experimento variando o tamanho do conjunto de treino.

Tabela 4.1: Estimativas dos parâmetros na Situação 1 para o Modelo 1.

PARÂMETROS		ρ_1	ρ_2	α_{11}	β_{11}	α_{12}	β_{12}	α_{21}	β_{21}	α_{22}	β_{22}
Valor Fixado		0,60	0,40	5,00	1,00	5,00	5,00	1,00	8,00	2,00	8,00
Estimativa	n=100	0,60	0,40	6,49	1,11	5,23	5,28	0,92	10,52	2,66	11,20
	n=1000	0,61	0,39	4,95	1,00	4,80	4,65	1,02	7,38	2,14	8,17
	n=5000	0,60	0,40	4,74	0,97	4,92	4,92	1,03	8,33	2,13	8,58
	n=10000	0,60	0,40	4,78	0,98	5,10	5,03	1,01	8,06	2,01	8,07

Tabela 4.2: Estimativas dos parâmetros da Situação 1 do Modelo 2.

PARÂMETROS		ρ_1	ρ_2	ρ_3	α_{11}	β_{11}	α_{12}	β_{12}	α_{21}	β_{21}	α_{22}	β_{22}	α_{31}	β_{31}	α_{32}	β_{32}
Valor Fixado		0,33	0,34	0,33	5,00	1,00	5,00	5,00	1,00	8,00	2,00	8,00	5,00	5,00	3,00	1,00
Estimativa	n=100	0,55	0,29	0,16	1,94	0,93	4,43	3,77	1,19	9,59	3,34	13,04	8,79	10,01	6,95	0,82
	n=1000	0,34	0,33	0,33	5,51	1,18	4,75	4,43	1,02	8,31	2,10	8,44	4,95	5,37	2,77	1,03
	n=5000	0,33	0,34	0,33	4,87	0,98	5,15	5,33	1,01	7,76	1,98	7,82	4,64	4,55	3,32	1,09
	n=10000	0,34	0,34	0,32	4,62	0,97	5,03	5,00	1,02	8,41	2,08	8,36	5,29	5,43	3,04	0,98

Ao analisarmos as Tabelas 4.1 e 4.2, notamos que os parâmetros estimados tendem a se aproximar do verdadeiro valor com o aumento do tamanho da amostra treino e que

algumas componentes são melhor estimadas, tanto no Modelo 1 quanto no Modelo 2. O estudo sobre a estimação dos parâmetros será abordado com maior profundidade a seguir, no Estudo 2.

Lembrando que o objetivo do Estudo 1 é observar a qualidade da densidade estimada quando variamos o número de componentes. Apresentamos nos gráficos a seguir o módulo da diferença entre a densidade verdadeira e a estimada (*dif*), usando diferentes tamanhos de amostras treino ($n = 100, 1000, 5.000, 10.000$) sobre amostra teste de tamanho $n = 1.000$. Nos gráficos da *dif* iremos observar que quanto mais próximo do valor zero melhor será a estimativa da densidade, a linha vermelha representa a média da *dif*, a linha verde é o eixo que corta o valor zero, para melhor visualização.

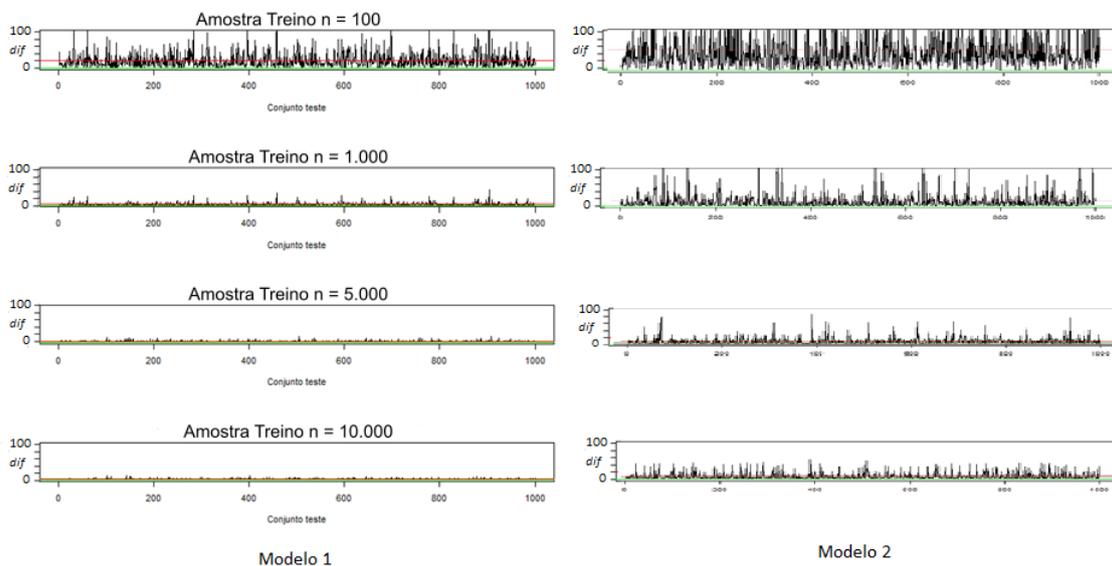


Figura 4.2: Plot Ordenado da *dif* da Situação 1.

Na Figura 4.2 são apresentados os resultados referentes ao Modelo 1 e 2. Notamos que, para os modelos, com o aumento do tamanho da amostra treino a densidade estimada fica mais próxima da densidade real, ou seja, melhor é a qualidade da densidade estimada, assim, como diminui a variabilidade da diferença entre a densidade estimada e a real.

Comparando os resultados para os Modelos 1 e 2, notamos que existe um decréscimo na qualidade da estimação da densidade quando aumentamos o número de componentes. Na Tabela 4.3 a seguir, com diferentes conjuntos de treino, são apresentadas a

média, erro padrão e intervalo de confiança obtidos com aproximação Normal para nossa medida *dif*, para 1.000 repetições do experimento.

Tabela 4.3: Média, erro padrão e IC da *dif* da Situação 1.

Modelo	Treino	Média	Erro Padrão	IC(95%)
1	100	21,21	21,00	[19,90;22,51]
	1.000	3,00	4,30	[2,73;3,26]
	5.000	3,10	3,11	[2,90;3,29]
	10.000	2,20	2,00	[2,07;2,32]
2	100	53,20	70,10	[48,85;57,54]
	1.000	14,01	20,00	[12,76;15,23]
	5.000	7,00	10,20	[6,36;7,63]
	10.000	7,04	10,00	[6,42;7,65]

Para o Modelo 1 e 2 o melhor ajuste se dá com o conjunto de treino de tamanho $n = 10.000$, como era esperado. De maneira geral observa-se que quanto menor o número de componentes e maior o tamanho da amostra treino, melhor será a qualidade de ajuste da densidade estimada. Observa-se que para o Modelo 1 conjunto de treino de tamanho $n = 1.000$ e $n = 5.000$ os IC's sobrepõem-se sugerindo não haver diferenças significativa na média da *dif* entre os dois conjuntos de treino. Os resultados sugerem que para o Modelo 2 os tamanhos de amostra treino $n = 5.000$ e $n = 10.000$ não tem diferenças significativas.

Obtivemos a média do Coeficiente de Correlação Linear de Pearson (r) entre a densidade estimada e a verdadeira, com base em 1.000 repetições do experimento, como mostra a seguir:

Tabela 4.4: Média da correlação (r) entre a densidade estimada e a verdadeira para Situação 1.

Modelo	Amostra Treino	r
1	100	0,981
	1.000	0,946
	5.000	0,935
	10.000	0,933
2	100	0,872
	1.000	0,909
	5.000	0,949
	10.000	0,942

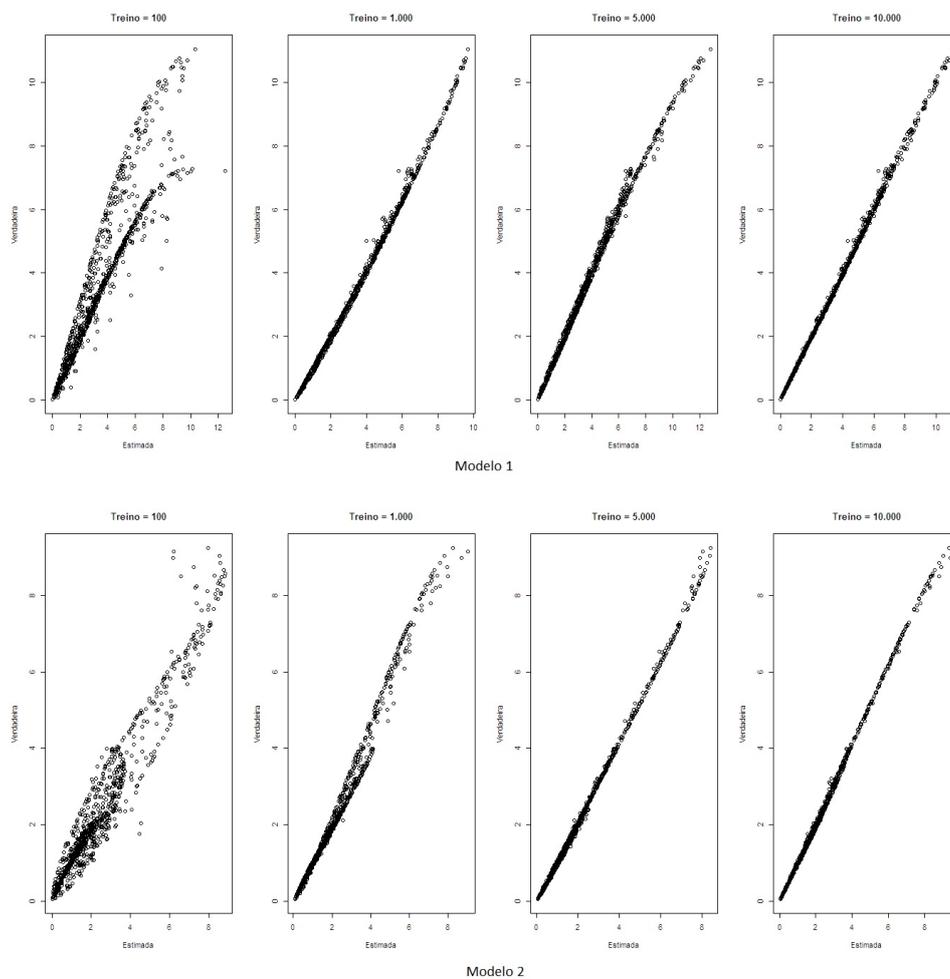


Figura 4.3: Dispersão da densidade estimada e a verdadeira da Situação 1.

Os resultados de análise de correlação sugerem que as estimativas das densidades estão altamente correlacionadas de forma positiva com as densidades reais, veja Figura 4.3. Sendo a menor média do coeficiente de correlação o apresentado no Modelo 2. Como em AD, empregamos o classificador de Bayes, a decisão é tomada em termos do máximo da densidade na observação a ser classificada, o emprego deste processo de estimação se mostra satisfatório para esse fim de classificação

Situação 2 : MFPB para observações em 3-dimensões

Nesta simulação aumentamos o número de variáveis, onde as observações simuladas tem dimensão $p = 3$, isto é, $\mathbf{x}' = (x_1, x_2, x_3)$. Simulamos dois conjuntos de dados variando o número de componentes, sendo um com 2-componentes e outro com 3-componentes.

Modelo 1:

$$f^1(\mathbf{x}; \Phi) = 0,6 * Beta(x_1; 5; 1) * Beta(x_2; 5; 5) * Beta(x_3; 3; 3) \\ + 0,4 * Beta(x_1; 1; 8) * Beta(x_2; 2; 8) * Beta(x_3; 3; 3).$$

Modelo 2:

$$f^2(\mathbf{x}; \Phi) = 0.34 * Beta(x_1; 1; 8) * Beta(x_2; 2; 8) * Beta(x_3; 2; 2) \\ + 0.33 * Beta(x_1; 5; 1) * Beta(x_2; 5; 5) * Beta(x_3; 2; 2) \\ + 0.33 * Beta(x_1; 5; 5) * Beta(x_2; 3; 1) * Beta(x_3; 2; 2).$$

Os gráficos de dispersão e histograma, dos dois modelos são apresentados, assim, como, a correlação entre as variáveis. Para uma análise inicial obtivemos a estimação dos parâmetros para cada modelo, referente aos resultados de uma única repetição do experimento variando o tamanho do conjunto de treino (veja Figura 4.4 e Tabela 4.5).

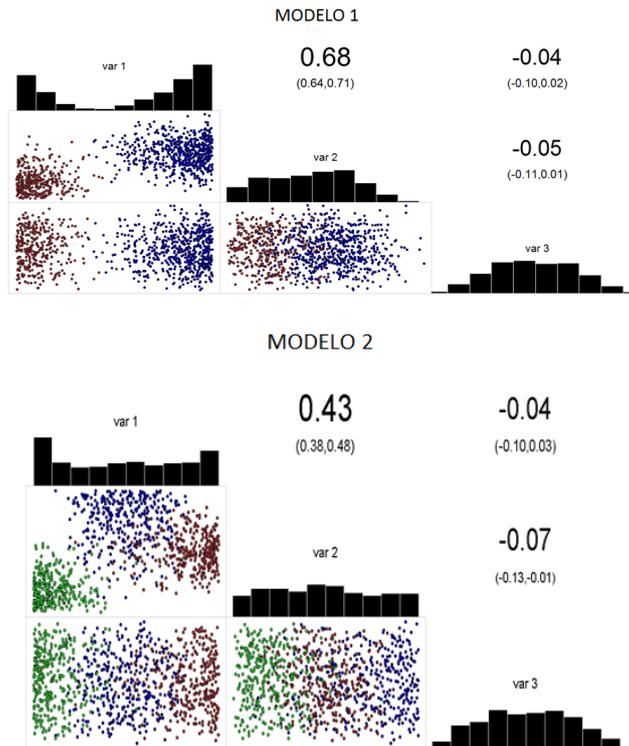


Figura 4.4: Histograma, dispersão e correlação da Situação 2.

Pelo gráfico de dispersão podemos visualizar as componentes. No Modelo 1 a maior valor da correlação entre as variáveis é de 0,68 enquanto que no Modelo 2 é 0,43.

Tabela 4.5: Estimativas dos parâmetros da Situação 2 do Modelo 1.

PARÂMETROS	ρ_1	ρ_2	α_{11}	β_{11}	α_{12}	β_{12}	α_{13}	β_{13}	α_{21}	β_{21}	α_{22}	β_{22}	α_{23}	β_{23}	
Valor Fixado	0,60	0,40	5,00	1,00	5,00	5,00	3,00	3,00	1,00	8,00	2,00	8,00	3,00	3,00	
Estimativa	n=100	0,65	0,35	3,81	0,85	5,45	4,72	3,30	3,44	1,25	9,37	1,88	8,82	3,15	3,33
	n=1000	0,59	0,41	5,63	0,97	5,26	5,23	3,33	3,48	1,04	7,75	1,94	7,44	2,98	3,03
	n=5000	0,59	0,41	4,95	1,01	4,95	4,96	2,93	2,96	0,98	7,71	1,93	7,60	3,03	3,10
	n=10000	0,60	0,40	5,11	1,01	5,00	5,01	3,01	3,04	0,97	7,89	2,01	8,00	2,96	2,94

Tabela 4.6: Estimativas dos parâmetros da Situação 2 do Modelo 2.

PARÂMETROS	ρ_1	ρ_2	ρ_3	α_{11}	β_{11}	α_{12}	β_{12}	α_{13}	β_{13}	α_{21}	β_{21}	α_{22}	β_{22}	α_{23}	β_{23}	α_{31}	β_{31}	α_{32}	β_{32}	α_{33}	β_{33}	
Valor Fixada	0,34	0,33	0,33	1,00	8,00	2,00	8,00	2,00	2,00	5,00	1,00	5,00	5,00	2,00	2,00	5,00	5,00	3,00	1,00	2,00	2,00	
Estimativa	n=100	0,28	0,37	0,35	0,94	7,70	1,89	6,54	2,25	2,43	8,74	1,09	6,38	6,84	2,55	2,30	4,90	4,31	3,49	0,94	2,18	1,82
	n=1000	0,33	0,36	0,30	0,90	6,91	1,98	8,12	2,08	2,14	5,21	1,03	4,89	4,78	2,07	2,31	5,64	5,73	3,56	1,06	2,03	1,95
	n=5000	0,33	0,37	0,31	1,02	7,98	1,99	7,85	1,99	1,93	3,98	0,95	5,11	5,16	1,80	1,86	4,94	5,11	3,52	1,01	2,09	2,09
	n=10000	0,34	0,32	0,34	0,99	8,20	1,99	7,95	1,99	2,02	5,49	1,03	5,09	5,13	1,98	1,98	4,93	4,99	3,00	1,00	1,93	1,99

Ao analisarmos as Tabelas 4.5 e 4.6, notamos que os parâmetros estimados tendem a se aproximar do verdadeiro valor com o aumento do tamanho da amostra treino e que

algumas componentes são melhor estimadas, tanto no Modelo 1 quanto no Modelo 2. Seguindo o mesmo foco de estudo da Situação 1.

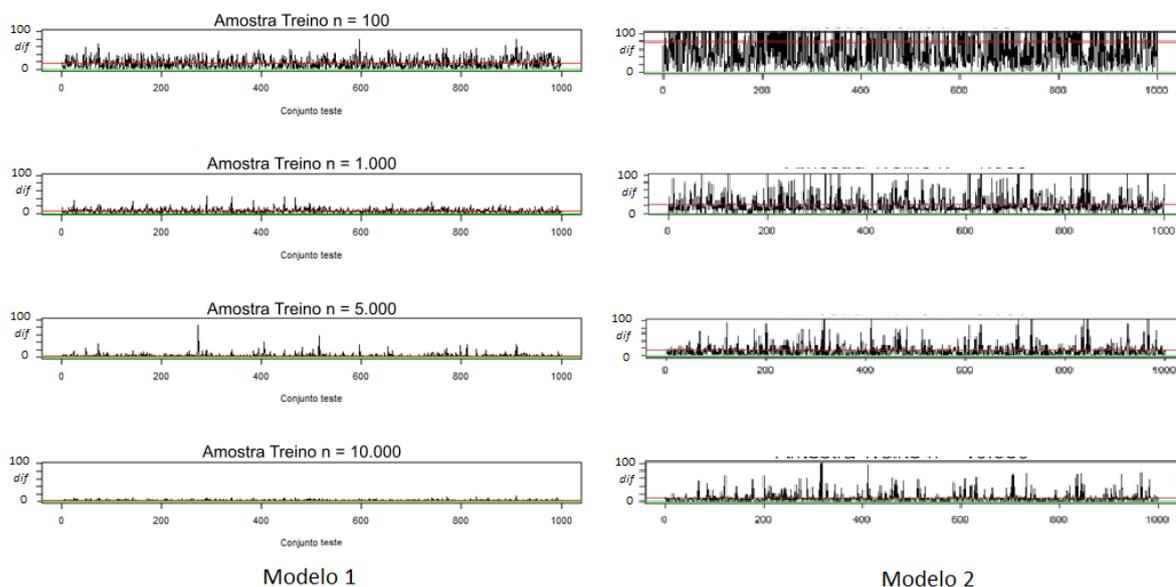


Figura 4.5: Plot Ordenado da *dif* da Situação 2.

Notamos na Figura 4.5 o mesmo comportamento da Situação 1, com o aumento do tamanho da amostra treino melhor é a qualidade da densidade estimada, isto se verificando tanto para o Modelo 1 quanto para o Modelo 2. Na Tabela 4.7 são apresentadas a média, erro padrão e intervalo de confiança para nossa medida *dif*, para 1.000 repetições do experimento.

Tabela 4.7: Média, erro padrão e IC da *dif* da Situação 2.

Número de Componentes	Treino	Média	Erro Padrão	IC(95%)
1	100	17,00	14,00	[16,13;17,86]
	1.000	7,10	6,30	[6,70;7,49]
	5.000	4,20	6,00	[3,82;4,57]
	10.000	3,11	2,00	[2,97;3,22]
2	100	80,00	99,90	[73,80;86,19]
	1.000	22,00	48,00	[19,02;24,97]
	5.000	12,00	18,00	[10,88;13,11]
	10.000	8,00	13,00	[7,19;8,80]

Observa-se que quanto menor o número de componentes e maior o tamanho da amostra treino, melhor será a qualidade de ajuste da densidade estimada. Diferentemente do que foi observado na Tabela 4.3, na Situação 2 nenhum dos IC's se sobrepõem para ambos os modelos, o que sugere que as diferenças na média da *dif* dos conjuntos de treino são significativas. Para ambos Modelo 1 e 2, o melhor ajuste se dá com o conjunto de treino de tamanho $n = 10.000$.

Na Tabela 4.8 a média do Coeficiente de Correlação Linear de Pearson (r) entre à densidade estimada e a verdadeira, a partir 1.000 repetições do experimento, como mostra a seguir:

Tabela 4.8: Média da correlação (r) entre a densidade estimada e a verdadeira para Situação 2.

Modelo	Amostra Treino	r
1	100	0,912
	1.000	0,996
	5.000	0,996
	10.000	0,998
2	100	0,873
	1.000	0,993
	5.000	0,998
	10.000	0,998

Verifica-se o mesmo comportamento da MFPB bidimensional, ou seja, sugerem que as estimativas das densidades estão altamente correlacionadas de forma positiva com as densidades verdadeiras, veja Figura 4.6 e Tabela 4.8. Mostrando-se satisfatório para esse fim de classificação.

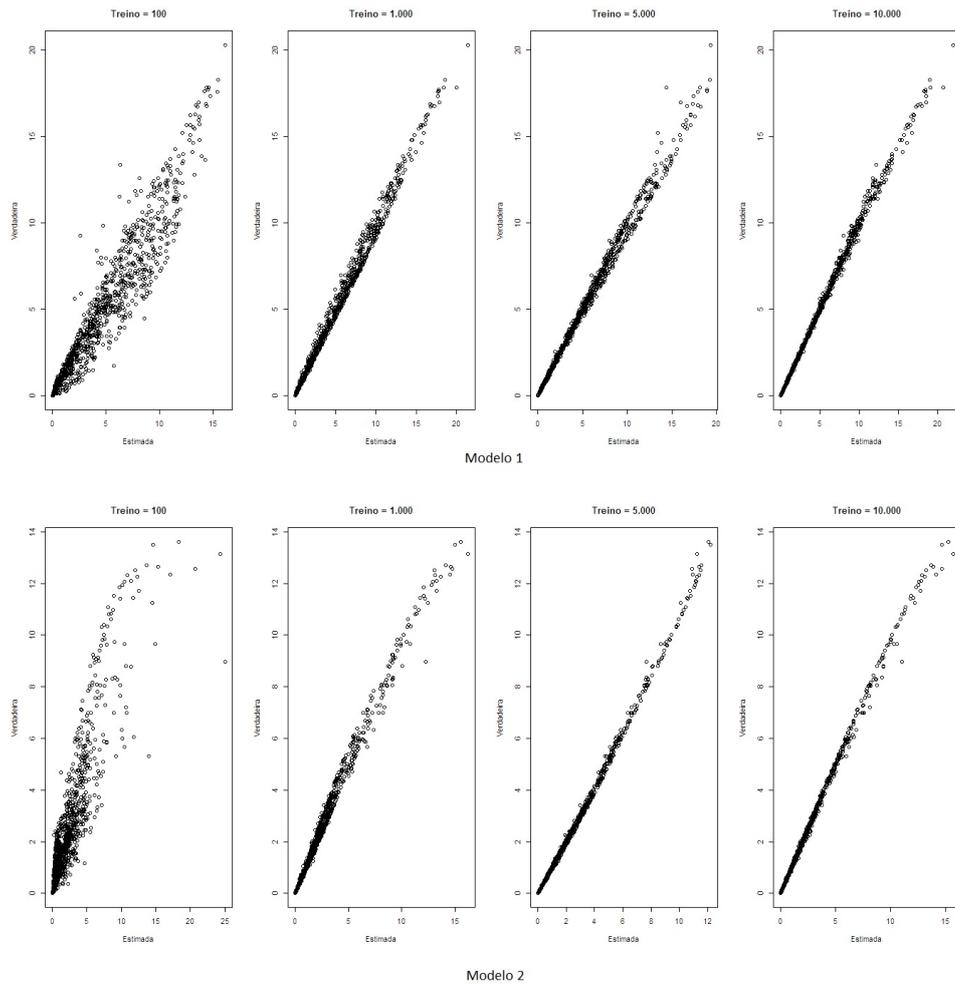


Figura 4.6: Dispersão da densidade estimada e a verdadeira da Situação 2.

4.1.2 Estudo 1 para o modelo MFD

Como definida na Seção 3.5 empregamos o modelo MFD, dado por (3.53). Foram realizados os mesmos experimentos feitos para MFPB, ou seja, observar o comportamento da diferença da densidade estimada e a verdadeira em relação a variação do número de componentes, levando em consideração o tamanho da amostra treino.

Situação 3 : MFD para observações em 2-dimensões

Na Situação 3 trazemos um casos simples em análise de dados multivariados, ou seja, o nosso conjunto de dados tem dimensão $p = 2$. Após serem fixados os parâmetros,

simulamos dois conjuntos de dados variando o número de componentes, sendo um com 2-componentes e outro com 3-componentes.

Modelo 1:

$$f^1(\mathbf{x}; \Phi) = 0,4 * Dir(\mathbf{x}_1; 10; 20) + 0,6 * Dir(\mathbf{x}_2; 5; 2).$$

Modelo 2:

$$f^2(\mathbf{x}; \Phi) = 0,4 * Dir(\mathbf{x}_1; 10; 20) + 0,3 * Dir(\mathbf{x}_2; 5; 2) + 0,3 * Dir(\mathbf{x}_3; 20; 20).$$

Na Figura 4.7 observamos que a MFD bidimensional, as variáveis são estritamente correlacionadas. A seguir os gráficos de dispersão e histograma, gerados a partir de um caso da amostra teste, assim, como, a correlação entre as variáveis. Nas Tabelas 4.9 e 4.10 temos as estimativas dos parâmetros para cada modelo, obtida de uma única repetição do experimento.



Figura 4.7: Histograma, dispersão e correlação da Situação 3.

Tabela 4.9: Estimativas do parâmetros da Situação 3 do Modelo 1.

PARÂMETROS		ρ_1	ρ_2	α_{11}	α_{12}	α_{21}	α_{22}
Valor Fixado		0,40	0,60	10,00	20,00	5,00	2,00
Estimativa	n=100	0,52	0,48	7,86	13,03	7,01	2,31
	n=1000	0,51	0,49	8,16	14,89	8,39	2,73
	n=5000	0,51	0,49	7,92	14,13	8,14	2,51
	n=10000	0,51	0,49	7,73	13,76	8,23	2,54

Tabela 4.10: Estimativas do parâmetros da Situação 3 do Modelo 2.

PARÂMETROS		ρ_1	ρ_2	ρ_3	α_{11}	α_{12}	α_{21}	α_{22}	α_{31}	α_{32}
Valor Fixado		0,40	0,30	0,30	10,00	20,00	5,00	2,00	20,00	20,00
Estimativa	n=100	0,36	0,21	0,43	10,00	22,94	14,89	4,12	25,04	23,60
	n=1000	0,39	0,23	0,38	12,63	26,18	11,52	2,99	27,05	25,17
	n=5000	0,39	0,22	0,39	12,98	27,30	10,97	2,81	25,35	24,13
	n=10000	0,40	0,21	0,39	12,74	27,04	11,85	2,97	23,94	22,52

Ao analisarmos as Tabelas 4.9 e 4.10, notamos que nem todos os parâmetros estimados tendem a se aproximar do verdadeiro valor com o aumento do tamanho da amostra treino e que algumas componentes são melhor estimadas, tanto no Modelo 1 quanto no Modelo 2. O estudo sobre a estimação dos parâmetros será abordado com maior profundidade a seguir, no Estudo 2. Apresentamos nos gráficos a seguir o módulo da diferença entre a densidade verdadeira e a estimada (*dif*), iremos observar que quanto mais próximo do valor zero melhor será a estimativa da densidade.

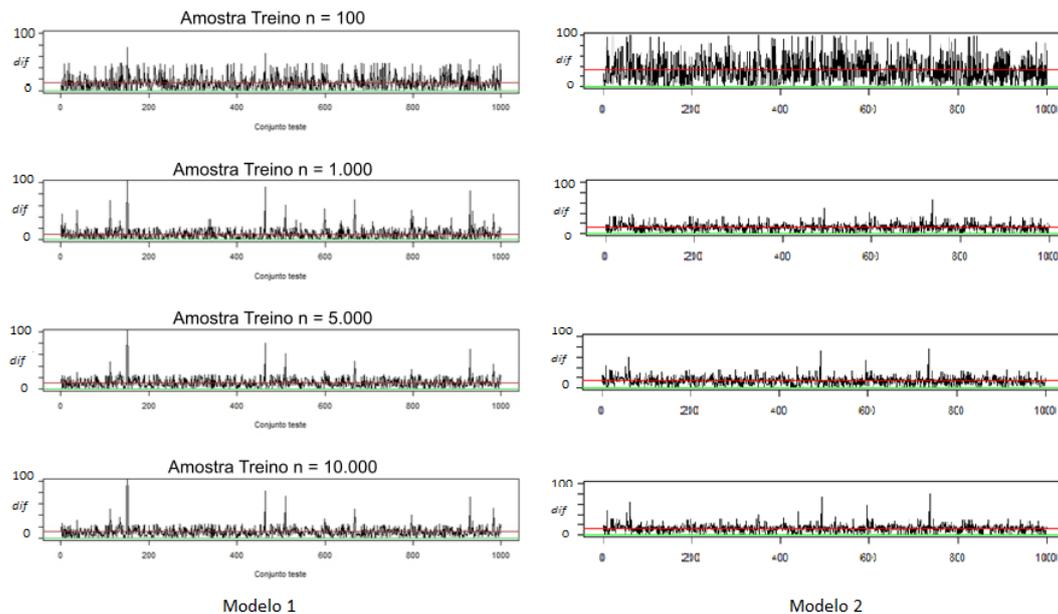


Figura 4.8: Plot Ordenado da *dif* da Situação 3.

Na Figura 4.8 notamos que com o aumento do tamanho da amostra treino a densidade estimada fica mais próxima da densidade verdadeira, este comportamento ocorrendo para os dois modelos. Ainda em relação a Figura 4.8 notamos que existe um decréscimo na qualidade da estimação da densidade quando aumentamos o número de componentes.

Na Tabela 4.11 quantificamos os resultados observados nos gráficos anteriores.

Tabela 4.11: Média, erro padrão e IC da *dif* da Situação 3.

Número de Componentes	Treino	Média	Erro Padrão	IC(95%)
1	100	14,20	13,52	[13,36;15,03]
	1.000	10,30	10,30	[9,66;10,93]
	5.000	11,02	9,98	[10,40;11,63]
	10.000	11,00	10,00	[10,38;11,61]
2	100	20,25	13,56	[19,40;21,09]
	1.000	14,41	9,31	[13,83;14,98]
	5.000	13,62	8,33	[13,10;14,13]
	10.000	13,25	8,01	[12,75;13,74]

A média da *dif* para o modelo bidimensional da MFD é maior do que o apresentado para MFPB. Para Modelo 1, em média o melhor *dif* se dá com o conjunto de treino de tamanho $n = 1.000$, enquanto que para o Modelo 2 é com o conjunto de treino de tamanho $n = 10.000$. Analisando os intervalos de confiança, os resultados sugerem que não há diferença significativa para a média da *dif* no Modelo 1 com o conjunto os conjuntos de treino 1.000, 5.000 e 10.000, o mesmo ocorre para o Modelo 2. A média do Coeficiente de Correlação Linear de Pearson (r) sugerem existir uma forte correlação entre a densidade estimada e a verdadeira, veja Tabela 4.12.

Tabela 4.12: Média da correlação (r) entre a densidade estimada e a verdadeira para Situação 3.

Modelo	Amostra Treino	r
1	100	0,981
	1.000	0,946
	5.000	0,935
	10.000	0,933
2	100	0,872
	1.000	0,909
	5.000	0,949
	10.000	0,942

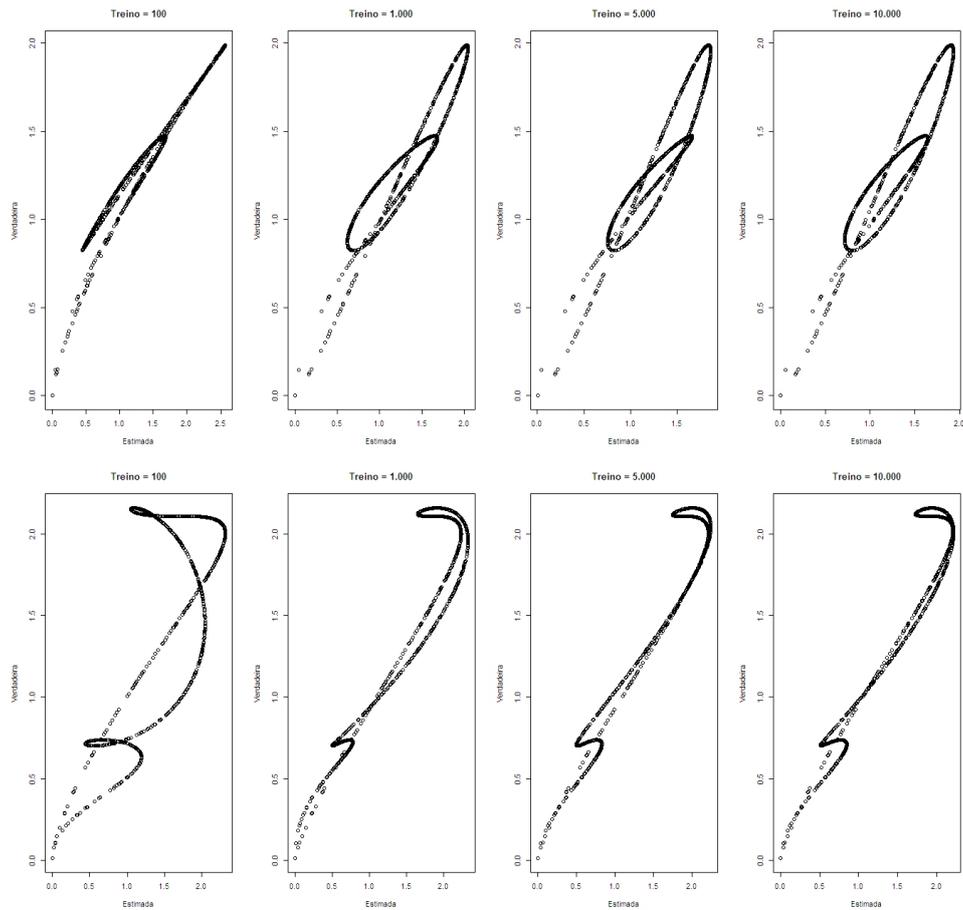


Figura 4.9: Dispersão da densidade estimada e a verdadeira da Situação 3.

Os resultados apresentados na Figura 4.9 não permitem estabelecer condições sobre o comportamento da *dif*, neste caso considerado. Assim, portanto, talvez sejam necessários outros procedimentos para avaliar as estimativas. No entanto, será possível verificar a qualidade destas estimativas no contexto de classificação no Estudo de Simulação 3.

Situação 4 : MFD para observações em 3-dimensões

Nesta simulação a dimensão do conjunto de $p = 3$. Simulamos dois conjuntos de dados variando o número de componentes, sendo um com 2-componentes e outro com 3-componentes.

Modelo 1:

$$f^1(\mathbf{x}; \Phi) = 0,6 * Dir(\mathbf{x}_1; 5; 2; 2) + 0,4 * Dir(\mathbf{x}_2; 10; 20; 2).$$

Modelo 2:

$$f^2(\mathbf{x}; \Phi) = 0,4 * Dir(\mathbf{x}_1; 10; 20; 2) + 0,3 * Dir(\mathbf{x}_2; 5; 2; 2) + 0,3 * Dir(\mathbf{x}_3; 20; 20; 2).$$

Na Figura 4.10 é possível verificar a reprodução das misturas de densidades tanto no Modelo 1 quanto no Modelo 2 por meio dos histogramas. Pelo gráfico de dispersão podemos identificar as componentes.

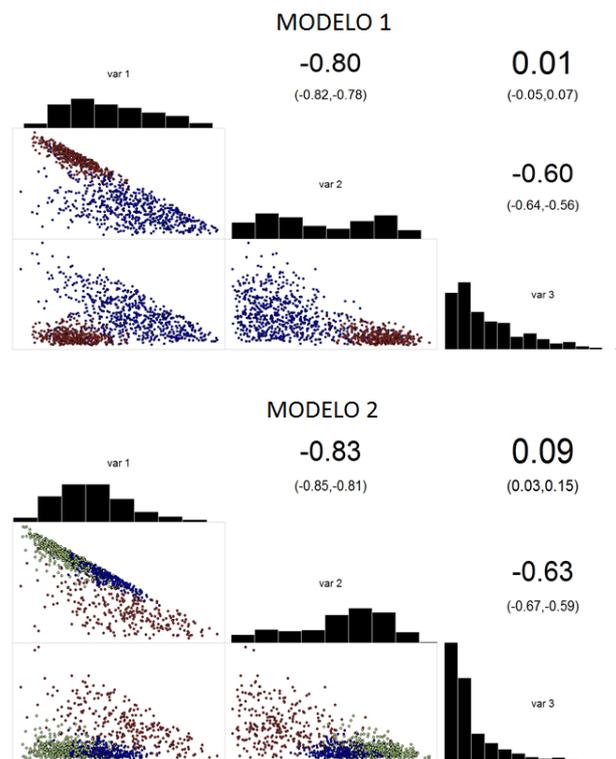


Figura 4.10: Histograma, dispersão e correlação da Situação 4.

A estimação dos parâmetros para cada modelo, referente aos resultados de uma única repetição do experimento são apresentados a seguir.

Tabela 4.13: Estimativas do parâmetros da Situação 4 do Modelo 1.

PARÂMETROS		ρ_1	ρ_2	α_{11}	α_{12}	α_{13}	α_{21}	α_{22}	α_{23}
Valor Fixado		0,6	0,4	5,00	2,00	2,00	10,00	20,00	2,00
Estimativa	n=100	0,48	0,52	6,60	2,25	2,68	8,73	14,49	1,89
	n=1000	0,58	0,42	5,57	2,09	2,25	8,54	16,54	1,81
	n=5000	0,56	0,44	5,51	2,02	2,20	8,73	16,62	1,87
	n=10000	0,56	0,44	5,31	1,98	2,16	9,03	17,35	1,90

Tabela 4.14: Estimativas do parâmetros da Situação 4 do Modelo 2.

PARÂMETROS		ρ_1	ρ_2	ρ_3	α_{11}	α_{12}	α_{13}	α_{21}	α_{22}	α_{23}	α_{31}	α_{32}	α_{33}
Valor Fixado		0,40	0,30	0,30	10,00	20,00	2,00	5,00	2,00	2,00	20,00	20,00	2,00
Estimativa	n=100	0,32	0,30	0,38	9,45	25,29	2,40	7,17	3,31	2,48	20,86	21,16	2,47
	n=1000	0,40	0,27	0,33	10,68	21,85	2,08	5,40	1,97	2,31	22,20	21,16	2,35
	n=5000	0,37	0,27	0,36	11,13	24,11	2,16	5,31	1,93	2,17	17,87	17,69	2,09
	n=10000	0,36	0,28	0,36	10,84	23,49	2,18	5,37	1,97	2,24	19,02	18,95	2,17

Ao analisarmos as Tabelas 4.13 e 4.14, observamos que tendem a ter o mesmo comportamento das simulações anteriores, onde alguns casos, as estimativas dos parâmetros se aproximam do valor fixado. A qualidade da densidade estimada melhora com o aumento da amostra treino (veja Figura 4.11). Para o Modelo 2 observamos o mesmo comportamento dos resultados do Modelo 1.

O melhor ajuste do Modelo 1 e 2 se dá com o conjunto de treino de tamanho $n = 10.000$. Analisando os IC's os dados sugerem não haver diferença significativas na média da *dif* para os conjuntos de treino $n = 1.000$, $n = 5.000$ e $n = 10.000$ para os modelos 1 e 2, veja Tabela 4.15.

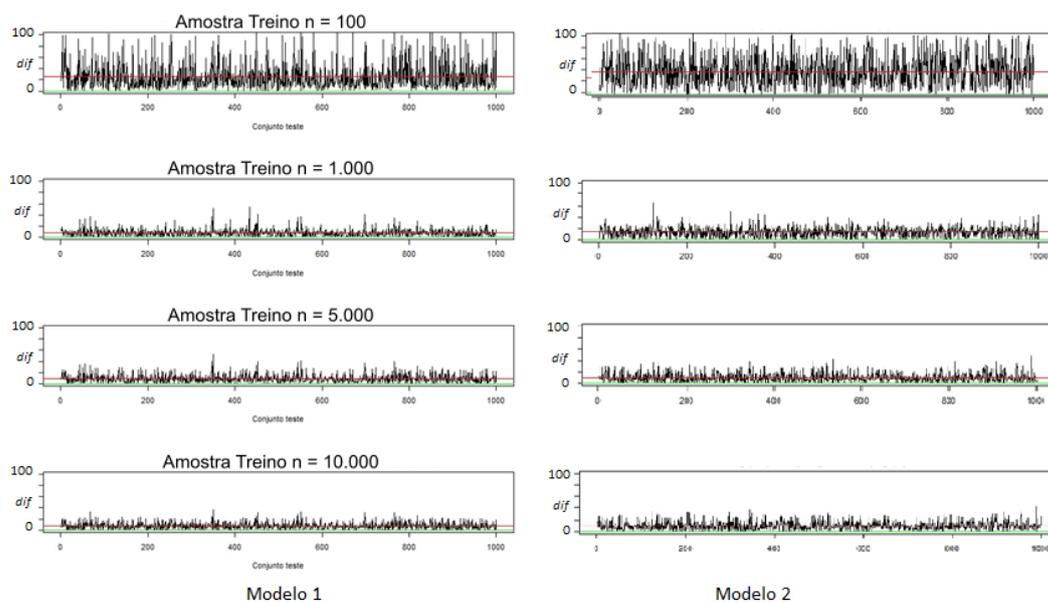


Figura 4.11: Plot Ordenado da *dif* da Situação 4.

Tabela 4.15: Média, erro padrão e IC da *dif* da Situação 4.

Número de Componentes	Treino	Média	Erro Padrão	IC(95%)
1	100	25,33	24,34	[23,82;26,83]
	1.000	8,99	7,41	[8,53;9,44]
	5.000	9,01	7,02	[8,57;9,44]
	10.000	8,19	6,04	[7,81;8,56]
2	100	32,50	27,30	[30,80;34,19]
	1.000	10,96	9,15	[10,39;11,52]
	5.000	11,01	9,08	[10,44;11,57]
	10.000	10,09	8,56	[9,55;10,62]

Verifica-se pelos resultados da Tabela 4.16 e Figura 4.12, o mesmo comportamento das simulações anteriores, onde as estimativas das densidades correlacionadas de forma positiva com as densidade verdadeira. Apesar de que em alguns pontos a densidade estimada não está bem ajustadas a verdadeira.

Tabela 4.16: Média da correlação (r) entre a densidade estimada e a verdadeira para Situação 4.

Modelo	Amostra Treino	r
1	100	0,956
	1.000	0,998
	5.000	0,992
	10.000	0,995
2	100	0,910
	1.000	0,982
	5.000	0,988
	10.000	0,990

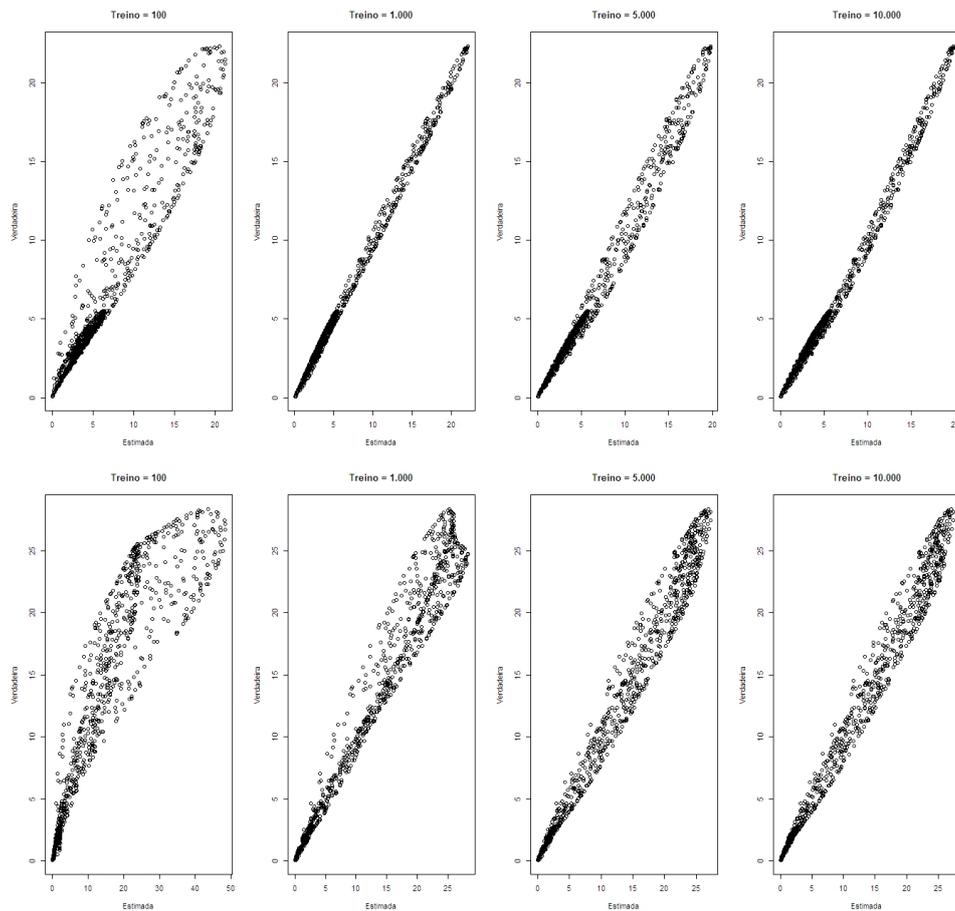


Figura 4.12: Dispersão da densidade estimada e a verdadeira da Situação 4.

4.1.3 Análise do Estudo de Simulação 1

A partir das estimativas da *dif*, tanto para MFPB (Situação 1 e 2) quanto para MFD (Situação 3 e 4), podemos afirmar que o aumento do número de componentes implica em um decréscimo na qualidade da estimação da densidade. Também, é notório que para o modelo MFPB o tamanho da amostra treino influencia de maneira significativa na qualidade da estimação da densidade, ou seja, quanto maior a amostra treino, melhor será a estimação da densidade. Porém, para MFD o aumento do tamanho da amostra treino maior do que 1.000 observações não influencia significativamente na estimação da *dif*. Em quase todas as situações houve diminuição da variabilidade da *dif* com o aumento da amostra treino, em um único caso não ocorreu tal fato, sendo no Modelo 1 da Situação 3.

Como esperado, em ambos modelos há evidências que existe diferença significativa quanto ao número de variáveis, pois os modelos tridimensionais apresentaram valores da *dif* maior do que os bidimensionais. Notamos, também, que o modelo MFPB tem uma melhor qualidade do ajuste da densidade estimada com a verdadeira, pois em média os valores da *dif* são maiores para os modelos MFD.

Em uma análise inicial, temos que as estimativas dos parâmetros na maioria dos casos apresentam-se bem ajustadas aos valores fixados, principalmente, as estimativas dos pesos das componentes. No entanto, em alguns casos as estimativas dos parâmetros apresentam muita discrepância aos valores fixados, sendo observado com maior relevância para o modelo MFD.

Como em AD, empregamos a classificação de Bayes, a decisão é tomada em termos do máximo da densidade na observação a ser classificada, o emprego deste processo de estimação se mostra satisfatório para esse fim de classificação. Os resultados do Coeficiente de Correlação sugerem que as estimativas das densidades são fortemente correlacionadas de forma positiva com as densidades verdadeiras em todas as situações apresentada.

4.2 Estudo de Simulação 2:

Neste estudo de simulação, os modelos MFPB e MFD tiveram seus parâmetros escolhidos com a finalidade de diminuir e/ou/ acentuar o grau de separação das componentes, simulando três cenários diferentes: nenhuma sobreposição, pouca sobreposição e muita sobreposição. O objetivo do estudo é analisar o comportamento das estimativas dos parâmetros nos diversos cenários. Considerando que a estimação via algoritmo EM para Misturas Finitas de Densidades pode levar a uma permutação nas componentes do modelo fixado, optamos por escolher os pesos da mistura e os parâmetros das componentes o mais diferente possível, para permitir associarmos as estimativas obtidas com os valores fixados.

As etapas desenvolvidas neste estudo de simulação foram:

- 1) Escolha dos parâmetros.
- 2) Obter as estimativas dos parâmetros via EM para diferentes tamanhos de amostra treino: $n = 100, 300, 500, 1.000, 5.000$ e 10.000 .
- 3) Comparar $\hat{\Theta}$ com Θ .
- 4) Repetir o processo 1.000 vezes.

O estudo será baseado na média das estimativas e os EQM dos parâmetros, assim como, na média das *dif* para cada tamanho de amostra.

4.2.1 Estudo 2 para o modelo MFPB

Para este estudo simulamos MFPB bidimensional, $\mathbf{x} = (x_1, x_2)$, com duas componentes, ou seja, empregamos o modelo dado em (3.46) com $p = 2$.

Modelo 1: Sem sobreposição entre as componentes

$$f^1(\mathbf{x}; \Phi) = 0,4 * Beta(x_1; 1; 8) * Beta(x_2; 2; 8) + 0,6 * Beta(x_1; 30; 1) * Beta(x_2; 30; 10).$$

Modelo 2: Pouca sobreposição entre as componentes

$$f^2(\mathbf{x}; \Phi) = 0,4 * Beta(x_1; 1; 8) * Beta(x_2; 2; 8) + 0,6 * Beta(x_1; 5; 1) * Beta(x_2; 5; 5).$$

Modelo 3: Muita sobreposição entre as componentes

$$f^3(\mathbf{x}; \Phi) = 0,4 * Beta(x_1; 2; 5) * Beta(x_2; 1; 5) + 0,6 * Beta(x_1; 2; 1) * Beta(x_2; 2; 2).$$

A seguir os gráficos de dispersão e histograma, gerados com 1.000 observações de cada modelo simulado, assim, como, a correlação entre as variáveis. A intenção deste estudo é verificar a qualidade da estimação dos parâmetros e observar a estimação da densidade quando variamos o grau de separação entre as componentes.

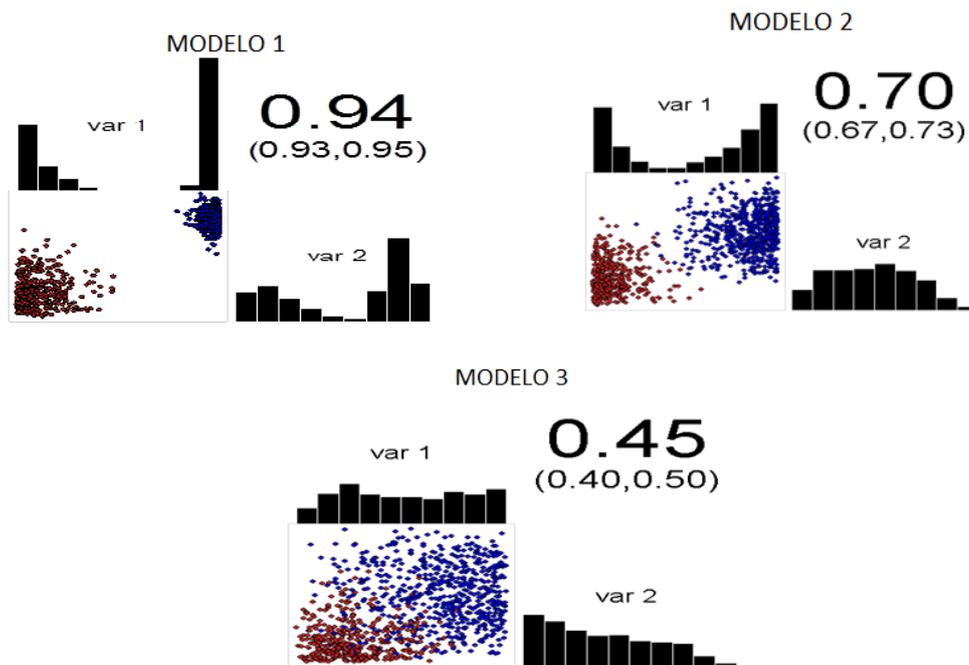


Figura 4.13: Histograma, dispersão e correlação da MFPB.

Na Figura 4.13 por meio do diagrama de dispersão e histogramas das variáveis é possível visualizar o grau de separação entre as componentes em cada modelo simulado. A correlação entre variáveis são diferentes em cada modelo, porém, não foi a intenção do estudo.

A seguir apresentamos à média das estimativas dos parâmetros dos Modelo 1, 2 e 3, variando o tamanho do conjunto treino. Lembrando que o experimento teve 1.000 repetições.

Tabela 4.17: Média das estimativas dos parâmetros para MFPB - Modelo 1.

PARÂMETROS		ρ_1	ρ_2	α_{11}	β_{11}	α_{12}	β_{12}	α_{21}	β_{21}	α_{22}	β_{22}
Valor Fixado		0,400	0,600	1,000	8,000	2,000	8,000	30,000	1,000	30,000	10,000
Estimativa	100	0,377	0,623	1,127	10,635	2,562	8,967	34,247	1,047	31,202	10,472
	300	0,383	0,617	0,837	6,884	1,969	8,384	30,280	1,047	28,280	9,466
	500	0,417	0,583	0,997	8,266	1,933	7,774	30,070	0,964	31,682	10,666
	1.000	0,405	0,595	0,971	8,014	1,944	8,164	29,693	1,016	32,534	10,883
	5.000	0,401	0,599	1,002	8,109	2,004	8,004	30,318	0,999	28,814	9,605
	10.000	0,399	0,601	0,986	7,948	1,984	7,877	30,035	1,002	29,845	9,913

Tabela 4.18: Média das estimativas dos parâmetros para MFPB - Modelo 2.

PARÂMETROS		ρ_1	ρ_2	α_{11}	β_{11}	α_{12}	β_{12}	α_{21}	β_{21}	α_{22}	β_{22}
Valor Fixado		0,400	0,600	1,000	8,000	2,000	8,000	5,000	1,000	5,000	5,000
Estimativa	100	0,397	0,603	1,193	8,624	2,253	8,664	5,197	1,215	5,223	5,370
	300	0,401	0,599	1,033	8,323	2,049	8,215	5,142	1,021	5,109	5,112
	500	0,401	0,599	1,017	8,216	2,021	8,102	5,027	1,005	5,065	5,069
	1.000	0,400	0,600	1,008	8,058	2,013	8,053	5,028	1,003	5,031	5,029
	5.000	0,400	0,600	1,003	8,018	2,003	8,013	5,004	1,000	5,009	5,007
	10.000	0,400	0,600	1,000	8,001	1,998	7,990	5,003	1,000	5,006	5,005

Tabela 4.19: Média das estimativas dos parâmetros para MFPB - Modelo 3

PARÂMETROS		ρ_1	ρ_2	α_{11}	β_{11}	α_{12}	β_{12}	α_{21}	β_{21}	α_{22}	β_{22}
Valor Fixado		0,400	0,600	2,000	5,000	1,000	5,000	2,000	1,000	2,000	2,000
Estimativa	100	0,373	0,627	3,400	5,680	1,567	4,495	1,722	1,871	1,572	2,139
	300	0,422	0,578	1,848	4,067	1,108	4,933	2,502	1,068	2,223	1,979
	500	0,363	0,637	1,994	5,004	1,066	6,096	1,897	1,065	2,015	2,062
	1.000	0,392	0,608	2,079	5,401	1,010	5,132	1,876	0,993	2,067	2,082
	5.000	0,398	0,602	1,951	4,796	1,000	4,994	1,957	0,974	2,027	2,022
	10.000	0,392	0,608	2,042	5,161	1,025	5,217	1,975	0,986	1,980	1,969

Analisando os resultados apresentados nas Tabelas 4.17, 4.18 e 4.19, em média as estimativas dos parâmetros ficam mais próximos dos valores fixados, para alguns casos o aumento do conjunto de treino, mesmo no caso do Modelo 3, onde temos uma sobreposição acentuada das componentes da mistura. Também, estimamos o Erro Quadrático

Méio (EQM) das estimativas dos parâmetros para uma melhor compreensão das mesmas. As estimativas do EQM estão apresentadas na Tabela 4.20.

Tabela 4.20: EQM das estimativas dos parâmetros para MFPB.

MODELO	n_{treino}	ρ_1	ρ_2	α_{11}	β_{11}	α_{12}	β_{12}	α_{21}	β_{21}	α_{22}	β_{22}
1	100	0,001	0,001	0,100	21,358	0,374	1,522	193,709	0,066	27,514	1,736
	300	0,001	0,001	0,031	1,415	0,035	0,233	44,100	0,042	5,393	0,693
	500	0,001	0,001	0,003	0,996	0,033	0,161	6,288	0,004	5,159	0,485
	1.000	0,001	0,001	0,001	0,907	0,009	0,042	1,625	0,004	4,242	0,213
	5.000	0,000	0,000	0,001	0,291	0,000	0,017	0,288	0,000	1,919	0,190
	10.000	0,000	0,000	0,000	0,005	0,000	0,005	0,071	0,000	0,636	0,077
2	100	0,002	0,002	0,710	8,253	0,687	5,951	2,009	1,210	1,314	1,424
	300	0,001	0,001	0,017	1,996	0,071	1,399	0,496	0,011	0,312	0,315
	500	0,001	0,001	0,009	1,180	0,035	0,728	0,264	0,006	0,180	0,181
	1.000	0,000	0,000	0,005	0,504	0,019	0,353	0,122	0,003	0,081	0,078
	5.000	0,000	0,000	0,001	0,102	0,004	0,068	0,024	0,001	0,017	0,017
	10.000	0,000	0,000	0,000	0,051	0,002	0,035	0,012	0,000	0,008	0,008
3	100	0,014	0,014	2,483	14,125	1,721	4,206	0,598	3,285	0,440	0,429
	300	0,003	0,003	0,145	1,558	0,026	2,077	0,455	0,008	0,256	0,044
	500	0,002	0,002	0,211	1,050	0,011	2,073	0,012	0,005	0,110	0,028
	1.000	0,001	0,001	0,076	1,007	0,000	0,039	0,017	0,000	0,021	0,024
	5.000	0,000	0,000	0,004	0,047	0,000	0,023	0,006	0,000	0,003	0,001
	10.000	0,000	0,000	0,001	0,044	0,000	0,010	0,002	0,000	0,001	0,000

A partir da Tabela 4.20, fica evidente que a qualidade da estimativa dos parâmetros, em termos do EQM, melhora com o aumento do tamanho da amostra de treino, isto é, para os três modelos considerados. Em geral, o resultado sugerem que para estimar os pesos da mistura o processo de estimação não apresenta dificuldades com nenhum dos tamanho de conjunto de treino para todos os três modelos.

4.2.2 Estudo 2 para o modelo MFD

Para este estudo simulamos MFD bidimensional, $\mathbf{x} = (x_1, x_2)$, com duas componentes, a partir do modelo dados em (3.53), com $p = 2$

Modelo 1: Nenhuma sobreposição entre as componentes

$$f^1(\mathbf{x}; \Phi) = 0,4 * Dir(\mathbf{x}; 200; 500) + 0,6 * Dir(\mathbf{x}; 10; 5).$$

Modelo 2: Pouca sobreposição entre as componentes

$$f^2(\mathbf{x}; \Phi) = 0,4 * Dir(\mathbf{x}; 10; 20) + 0,6 * Dir(\mathbf{x}; 5; 2).$$

Modelo 3: Muita sobreposição entre as componentes

$$f^3(\mathbf{x}; \Phi) = 0,4 * Dir(\mathbf{x}; 10; 3) + 0,6 * Dir(\mathbf{x}; 9; 5).$$

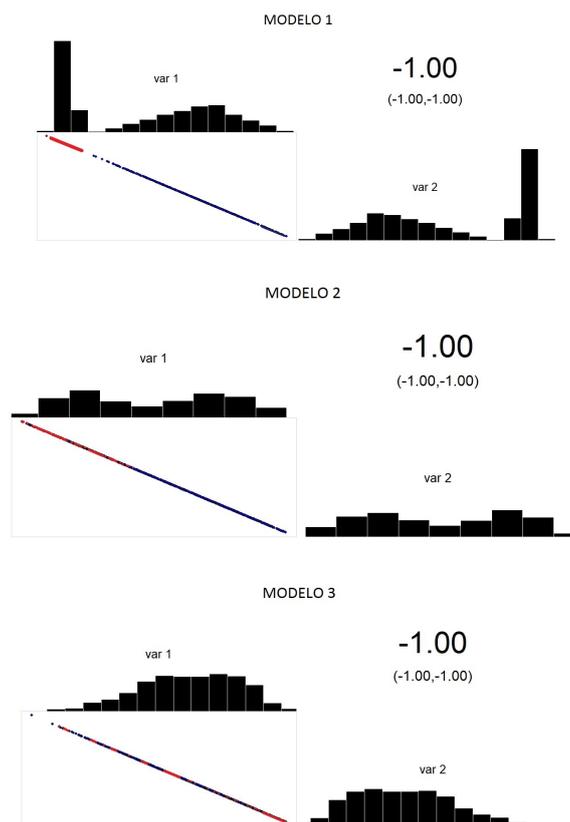


Figura 4.14: Histograma, dispersão e correlação da MFD.

Na Figura 4.14 é possível visualizar o grau de separação entre as componentes em cada modelo simulado por meio do diagrama de dispersão e histogramas das variáveis.

As variáveis são fortemente correlacionadas nos três modelos, característica da densidade MFD bidimensional. Para 1.000 repetições do experimento, obtivemos a média das estimativas dos parâmetros do Modelo 1, 2 e 3, variando o conjunto treino.

Tabela 4.21: Média das estimativas dos parâmetros para MFD - Modelo 1.

PARÂMETROS		ρ_1	ρ_2	α_{11}	α_{12}	α_{21}	α_{22}
Valor Fixado		0,4000	0,6000	200,0000	500,0000	10,0000	5,0000
Estimativa	100	0,4258	0,5742	56,2651	136,1519	12,4635	6,0549
	300	0,4207	0,5793	60,2637	145,6862	12,8780	5,9498
	500	0,4342	0,5658	55,5074	134,1287	11,8284	5,6612
	1.000	0,4316	0,5684	60,3469	145,9750	12,1654	5,7547
	5.000	0,4218	0,5782	63,2354	153,5320	11,9315	5,6765
	10.000	0,4215	0,5785	61,0340	148,0774	12,0410	5,7261

Tabela 4.22: Média das estimativas dos parâmetros para MFD - Modelo 2.

PARÂMETROS		ρ_1	ρ_2	α_{11}	α_{12}	α_{21}	α_{22}
Valor Fixado		0,4000	0,6000	10,0000	20,0000	5,0000	2,0000
Estimativa	100	0,4869	0,5131	7,0105	12,3333	9,0977	2,6260
	300	0,4914	0,5086	7,6866	13,7755	8,4071	2,5443
	500	0,4957	0,5043	7,9464	14,3532	8,4888	2,6277
	1.000	0,5116	0,4884	7,6904	13,6314	8,1758	2,5069
	5.000	0,5025	0,4975	7,6704	13,5830	8,2912	2,5396
	10.000	0,4955	0,5045	7,7416	13,8151	8,2051	2,5328

Tabela 4.23: Média das estimativas dos parâmetros para MFD - Modelo 3.

PARÂMETROS		ρ_1	ρ_2	α_{11}	α_{12}	α_{21}	α_{22}
Valor Fixado		0,4000	0,6000	10,0000	3,0000	9,0000	5,0000
Estimativa	100	0,4648	0,5352	19,2789	11,7558	17,4873	6,0746
	300	0,4278	0,5722	16,0360	12,0464	19,1376	5,3445
	500	0,4226	0,5774	15,2540	11,5031	16,6781	4,6371
	1.000	0,4454	0,5546	15,3674	11,2737	17,7144	4,8211
	5.000	0,4436	0,5564	15,6955	11,4584	17,5375	4,7762
	10.000	0,4437	0,5563	15,7105	11,4960	17,5274	4,7873

Analisando os resultados apresentados nas Tabelas 4.21, 4.22 e 4.23, observamos

que a média das estimativas dos parâmetros do modelo MFD apresenta desempenho inferior ao obtido para o modelo MFPB. As estimativas da componentes 1 do Modelo 1 tem valores muito discrepantes dos fixados. No entanto, estimamos o Erro Quadrático Médio (EQM) das estimativas dos parâmetros para uma melhor compreensão das mesmas. As estimativas do EQM estão apresentadas na Tabela 4.24.

Tabela 4.24: EQM dos Parâmetros Estimados para MFD.

MODELO	n_{treino}	ρ_1	ρ_2	α_{11}	α_{12}	α_{21}	α_{22}
1	100	0,0056	0,0056	21219,4345	135903,5650	9,8528	2,0253
	300	0,0004	0,0004	555,2850	3483,2364	0,9913	0,2123
	500	0,0004	0,0004	304,8521	1951,2335	1,4057	0,2281
	1.000	0,0001	0,0001	361,0702	2288,5161	0,3035	0,0353
	5.000	0,0002	0,0002	111,8106	708,7726	0,1153	0,0219
	10.000	0,0000	0,0000	22,9426	144,4347	0,0678	0,0143
2	100	0,0101	0,0101	10,0367	62,6970	21,8583	0,6721
	300	0,0015	0,0015	1,5588	6,5406	2,3509	0,1278
	500	0,0001	0,0001	0,9075	3,8158	0,5842	0,0657
	1.000	0,0008	0,0008	0,3512	1,7271	0,1574	0,0298
	5.000	0,0001	0,0001	0,0547	0,3138	0,0663	0,0079
	10.000	0,0001	0,0001	0,0589	0,3569	0,0657	0,0041
3	100	0,0069	0,0069	114,9079	86,2508	85,4402	11,4954
	300	0,0025	0,0025	13,2676	0,9122	12,0916	1,0718
	500	0,0018	0,0018	2,4324	1,4496	7,7526	0,6385
	1.000	0,0015	0,0015	1,0518	0,4438	2,7839	0,1260
	5.000	0,0001	0,0001	0,3425	0,2042	0,2091	0,0079
	10.000	0,0001	0,0001	0,2608	0,1224	0,1150	0,0079

A partir da Tabela 4.24, no caso de nenhuma sobreposição entre as componentes da mistura (Modelo 1), o maior EQM ocorre para o parâmetro α_{12} , pertencente a componente 1, o valor é considerado muito discrepante dos demais. No caso de pouca sobreposição entre as componentes da mistura (Modelo 2), o maior EQM ocorre para o parâmetro α_{12} , pertencente a componente 1. No caso de muita sobreposição entre as componentes da mistura (Modelo 3), o maior EQM ocorre para o parâmetro α_{11} , pertencente a componente 1. O EQM dos três modelos para $n_{treino} = 100$, apresenta valores muito alto,

principalmente para o modelo com nenhuma sobreposição das componentes (Modelo 1).

Em geral, os resultados sugerem que para estimar os pesos da mistura o processo de estimação não apresenta dificuldades com nenhum dos tamanho de conjunto de treino para todos os três modelos. Os resultados do modelo MFD são similares aos obtidos pelo modelo MFPB, onde a qualidade da estimativa dos parâmetros, em termos do EQM, melhora com o aumento do tamanho da amostra de treino, isto é, para os três modelos considerados.

4.3 Estudo de Simulação 3:

Neste estudo, queremos verificar se modelos diferentes conseguem aproximar as distribuições simuladas, observando o comportamento da estimação dos modelos propostos neste trabalho, com o objetivo de classificação, em três cenários diferentes com relação a sobreposição das classes: nenhuma, pouca e muita.

Neste estudo foram simuladas observações de dimensão $p = 3$, oriundas de Mistura Finita de Densidade de Dirichlet (MFD), Mistura Finita de Produtórios de Densidade Betas (MFPB) e uma situação onde as observações em cada dimensão são simuladas por Mistura Finita de Densidade Betas Independentes (MFBI).

Foram implementados classificadores de Bayes usando como modelo para as distribuições nas classes: Mistura Finita Densidade Dirichlet (MFD), Mistura Finita de Densidades de Produtórios Betas (MFPB), Naive Bayes de Mistura Finita de Densidades Betas (NBMFB), Naive Bayes com distribuição Normal (NBN), Naive Bayes com distribuição Não-Paramétrica com função núcleo de Epanechnikov (NBE), Análise Discriminante Linear (ADL) e Análise Discriminante Quadrática (ADQ). Ressaltando que para observações simuladas da MFD não foi possível empregar os classificadores ADL e ADQ, por problemas de multicolinearidade.

No total são abordados nove tipos de estruturas, oriundas de amostras de MFD, de MFPB e MFBI, sendo três estruturas de cada densidade mencionada. Para amostras

de MFPB e MFBI foram feitas transformações dos dados com a intenção de empregar o classificador de Bayes com MFD.

Empregamos o procedimento de conjunto de Treino e Teste para avaliação dos classificadores, repetindo o experimento várias vezes, determinamos a média, o desvio padrão e intervalos de confiança da Taxa de Erro (TE) de classificação para cada um dos classificador considerados. Todo o procedimento neste estudo de simulação foi utilizado os recursos de funções do software *R*. Os passos das simulações são descritos a seguir:

- 1) A Classe 1 foi gerada com probabilidade $P(Y = 1) = 0,5$ e a Classe 2 com probabilidade $P(Y = 2) = 0,5$;
- 2) Escolher o número de componentes e dos parâmetros do modelo de cada classe;
- 3) Gerar uma amostra teste de tamanho $n_{teste} = 20.000$;
- 4) Gerar amostras treinos, $n_{treino} = 200, 2.000, 10.000$ e 20.000 ;
- 5) Determinar as taxas de erro de classificação.
- 6) Faremos uso da transformação dos dados para cada classe, a fim de avaliar especialmente o modelo MFD:

$$v_i = \frac{x_i}{(x_1 + x_2 + x_3)}, i = 1, 2, 3.$$

- 7) Transformar as observações da amostra teste e treino;
- 8) Determinar as taxas de erro de classificação para os dados transformados;
- 9) Repetir o experimento 200 vezes.

4.3.1 Simulação de Amostras de MFD

As observações em cada classe foram geradas por Mistura Finita de Densidade Dirichlet, com dimensão $p = 3$. Os três cenários são chamados de Estrutura 1, 2 e 3. É

possível a análise do comportamento de cada variável através do histograma, assim como, a dispersão das classes e a correlação entre as variáveis. A cor verde representa a classe 1 e a cor vermelha a classe 2..

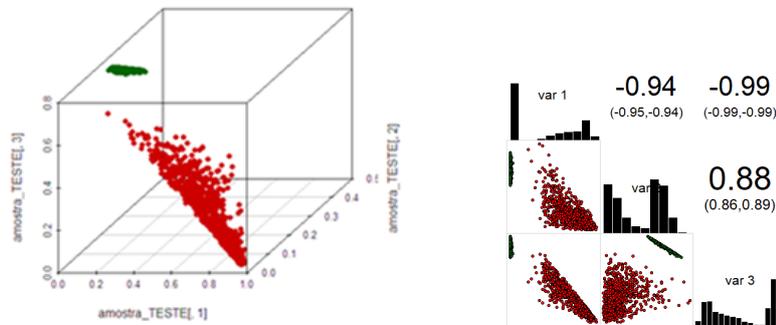


Figura 4.15: Histograma, dispersão e correlação da Estrutura 1.

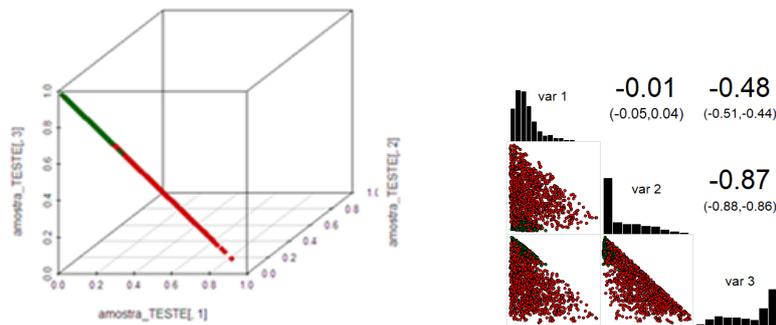


Figura 4.16: Histograma, dispersão e correlação da Estrutura 2.

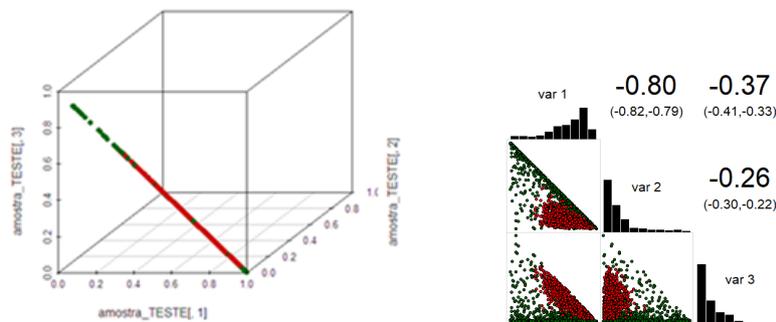


Figura 4.17: Histograma, dispersão e correlação da Estrutura 3.

O objetivo da Estrutura 1 é verificar o comportamento do classificador quando não

há sobreposição das classes, não admitindo erro de classificação. Essa separação entre as classes é possível verificar somente por meio do gráfico tridimensional (ver Figura 4.15). A Estrutura 2 é verificar o comportamento do classificador quando existe pouca sobreposição das classes, admitindo que exista um pequeno erro de classificação (ver Figura 4.16). A Estrutura 3 é verificar o comportamento do classificador quando há muita sobreposição das classes, admitindo que exista um grande erro de classificação (ver Figura 4.17). A seguir serão apresentadas as Taxas de Erros (TE) de classificação para cada estrutura.

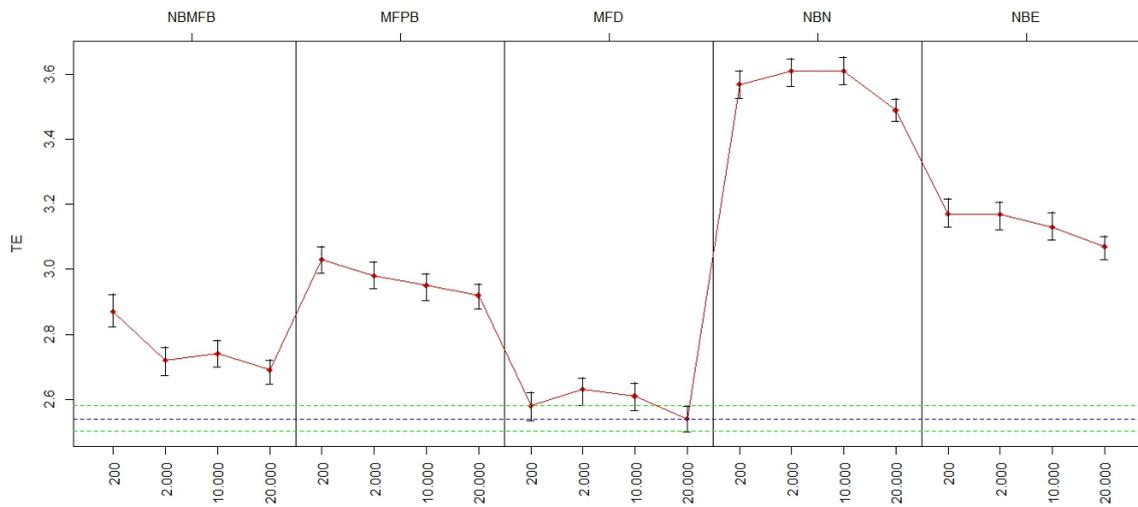


Figura 4.18: Média e IC (95%) da TE para Estrutura 2.

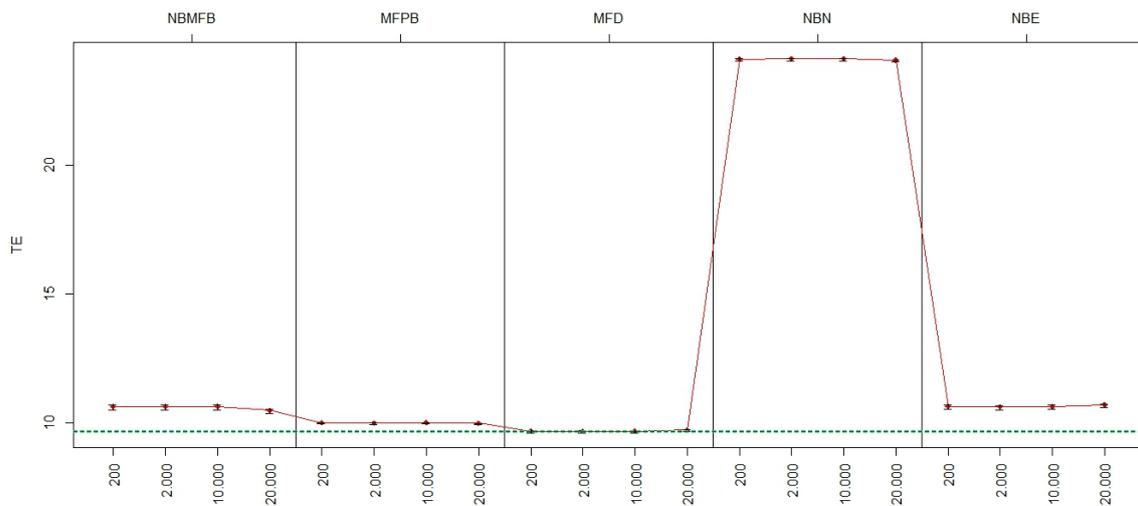


Figura 4.19: Média e IC (95%) da TE para Estrutura 3.

Para Estrutura 1, os resultados apresentaram que todos os modelos avaliados realizaram corretamente a classificação, ou seja, nas 200 repetições do experimento, a taxa de acerto foi de 100% para todos os classificadores.

Os resultados sugerem que a diferença do desempenho entre os classificadores é significativa, (veja Figura 4.18 e 4.19), ou seja, quando há muita sobreposição das classes o erro de classificação também aumenta.

Tabela 4.25: Média e desvio-padrão da TE para Simulações de Amostras de MFD.

Estrutura	Amostra Treino	NBMFB		MFPB		MFD		NBN		NBE	
		Média	DP								
1	200	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	2.000	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	10.000	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	20.000	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	200	2,87	0,36	3,03	0,30	2,58	0,31	3,57	0,30	3,17	0,31
	2.000	2,72	0,31	2,98	0,29	2,63	0,30	3,61	0,30	3,17	0,31
	10.000	2,74	0,29	2,95	0,30	2,61	0,29	3,61	0,31	3,13	0,31
	20.000	2,69	0,26	2,92	0,27	2,54	0,28	3,49	0,25	3,07	0,26
3	200	10,62	0,67	10,00	0,16	9,66	0,31	24,13	0,38	10,63	0,63
	2.000	10,62	0,68	9,98	0,16	9,66	0,31	24,14	0,38	10,61	0,61
	10.000	10,62	0,65	10,00	0,16	9,66	0,29	24,14	0,36	10,62	0,60
	20.000	10,47	0,57	9,97	0,08	9,71	0,18	24,09	0,23	10,69	0,50

Observando a Figura 4.18, notamos que a maior média da TE obtida foi para a distribuição com NBN (3,61%), com a menor amostra treino simulada. O ajuste que apresentou a menor média para TE foi o modelo verdadeiro, ou seja, a MFD (2,54%) para $n_{treino} = 20.000$, (ver Tabela 4.25). O modelo NBMFB teve a TE mais próxima do que foi observado no modelo verdadeiro, sendo a sua menor média da TE de 2,69% com um desvio padrão de 0,26% para $n_{treino} = 20.000$.

Para Estrutura 3 (veja Figura 4.19), a maior média da TE obtida foi para a distribuição com NBN (24,14%), com amostra treino simulada $n_{treino} = 2.000, 10.000$. O ajuste que apresentou a menor média para TE foi o modelo verdadeiro, ou seja, a MFD (9,66%) para $n_{treino} = 200, 2.000, 10.000$, (ver Tabela 4.25). O modelo MFPB teve a TE mais próxima do que foi observado no modelo verdadeiro, sendo a sua menor média da

TE de 9,97% com um desvio padrão de 0,16% para $n_{treino} = 20.000$. Porém, há diferença significativa entre os classificadores MFPB E MFD.

Para amostras simuladas a partir de MFD, os resultados na Tabela 4.25 sugerem que o melhor classificador é o verdadeiro (MFD) nas três Estruturas propostas. Na Estrutura 1, quando não existe sobreposição das classes, todos os modelos classificaram corretamente 100% das observações. Na Estrutura 2, quando existe pouca sobreposição das classes, a menor TE do modelo MFD ocorre para $n_{treino} = 20.000$, com média de 2,54% e desvio padrão de 0,28%. Na Estrutura 3, quando existe muita sobreposição das classes, a menor TE do modelo MFD acontece para $n_{treino} = 200, 2.000, 10.000$, com média de 9,66%, sendo que o menor desvio padrão ocorre para $n_{treino} = 10.000$, com 0,29%.

Com o aumento da amostra treino a média e a variabilidade da TE tendem a diminuir e o grau de separação entre as classes influencia, significativamente, na qualidade da classificação. Indicando que quanto maior for a sobreposição das classes maior será a TE.

4.3.2 Simulação de Amostras de MFPB

As observações em cada classe foram geradas por Mistura Finita de Densidades de Produtório de Betas, com dimensão $p = 3$. Os classificadores avaliados serão: MFD, MFPB (modelo verdadeiro), NBMFB, NBN, NBE, ADL e ADQ.

A Estrutura 4 tem o objetivo de verificar o comportamento do classificador quando as classes estão bastante separadas uma da outra, não admitindo erro de classificação (ver Figura 4.20). O objetivo da Estrutura 5 é verificar o comportamento do classificador quando existe pouca sobreposição das classes e o da Estrutura 6 é verificar o comportamento do classificador quando existe muita sobreposição das classes, admitindo um grande erro de classificação (ver Figuras 4.21 e 4.22). Essa separação entre as classes é possível verificar somente por meio do gráfico tridimensional. A cor verde representa a classe 1 e a cor vermelha a classe 2.

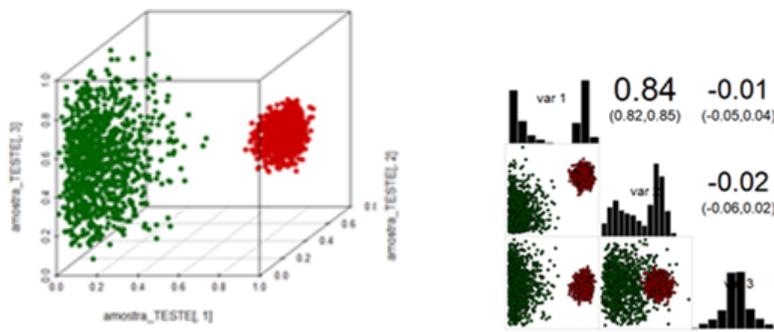


Figura 4.20: Histograma, dispersão e correlação da Estrutura 4.

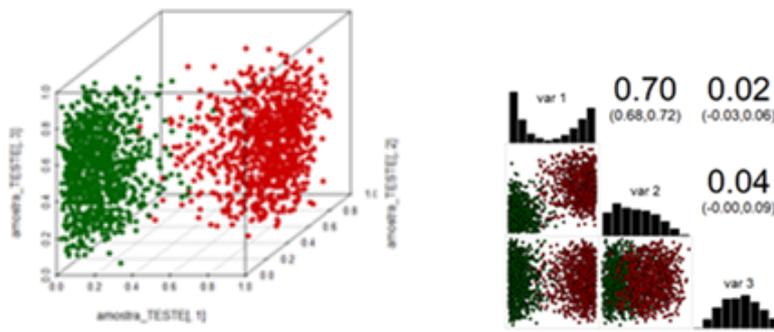


Figura 4.21: Histograma, dispersão e correlação da Estrutura 5.

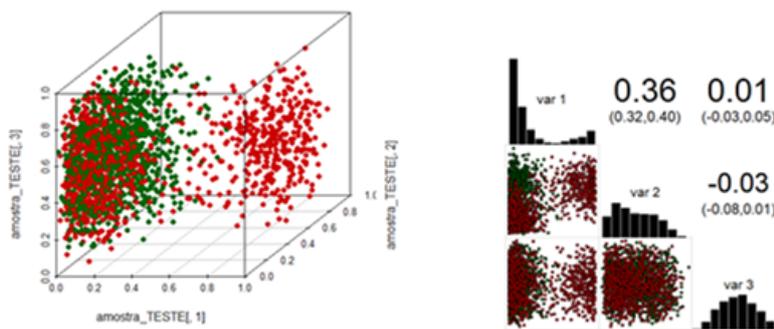


Figura 4.22: Histograma, dispersão e correlação da Estrutura 6.

Existe diferença significativa quanto a sobreposição das classe, aumentando o erro de classificação para o caso de muita sobreposição. Porém, não há diferença significativa para os tamanhos de amostras de treino dentro de cada classificador, veja Figuras 4.23, 4.24 e 4.25.

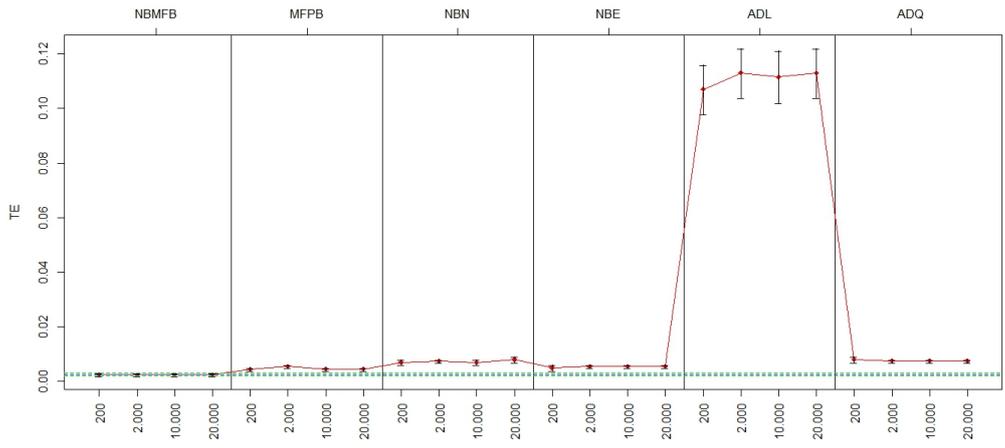


Figura 4.23: Média e IC (95%) da TE para Estrutura 4.

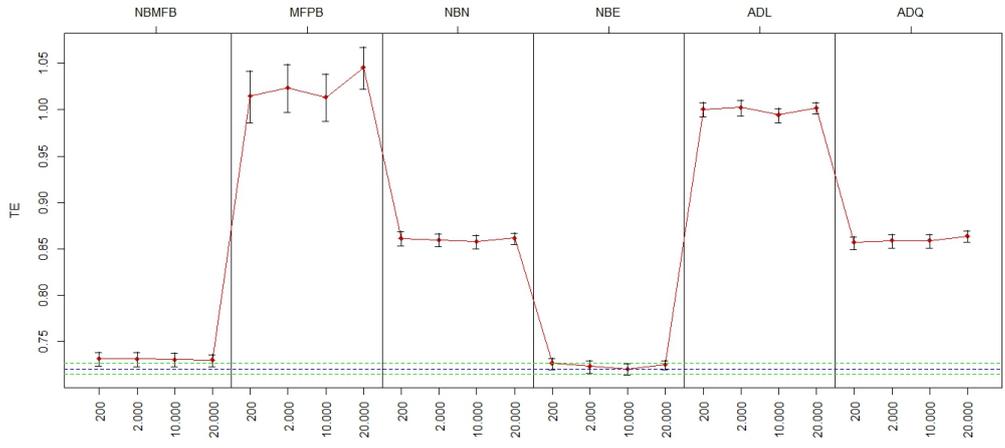


Figura 4.24: Média e IC (95%) da TE para Estrutura 5.

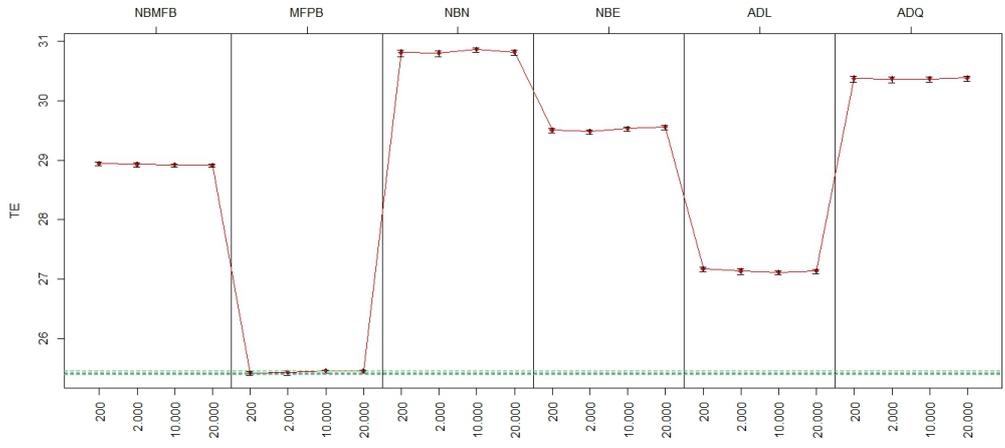


Figura 4.25: Média e IC (95%) da TE para Estrutura 6.

Para Estrutura 4, com os dados originais, a menor média da TE obtida foi com NBMFB (0,002%), com a amostra de treino $n_{treino} = 2.000$. O ajuste que apresentou a maior média para TE foi ADL (0,113%) para $n_{treino} = 2.000, 20.000$, apresentando uma grande variabilidade da TE, (veja Tabela 4.26).

Observando a Figura 4.24, referente a Estrutura 5, notamos que a maior média da TE obtida foi o modelo verdadeiro, ou seja, MFPB (1,05%), com a maior amostra de treino simulada ($n_{treino} = 20.000$). O ajuste que apresentou a menor média para TE foi NBE, com 0,72% para amostra treino ($n_{treino} = 10.000$). O modelo NBMFB apresenta TE próxima do que foi observado no modelo NBE, existindo uma interseção entre os intervalos de confiança dos dois modelos. A melhor média de TE do modelo NBMFB é de 0,73%, com desvio padrão 0,05%, ocorre para $n_{treino} = 20.000$. O classificador NBMFB para os dados originais pode ser tão bom quanto o NBE, (veja Tabela 4.26).

Para a Estrutura 6 (veja Figura 4.25), a maior média da TE foi para o NBN (30,859%), com amostra de treino $n_{treino} = 10.000$. O classificador que apresentou a menor média para TE foi com o modelo verdadeiro, ou seja, MFPB (25,432%) para $n_{treino} = 200$, (veja Tabela 4.26).

4.3.3 Simulação com Amostras de MFPB Transformadas

Com a transformação dos dados ocorre pouca sobreposição das classes. Essa mistura entre as classes é possível verificar somente por meio do gráfico tridimensional (veja Figura 4.26, 4.27 e 4.28). Fácil ver que existe um modelo de mistura em cada variável e que não há sobreposição das classes, sendo que existe uma correlação forte entre as variáveis X_1 e X_3 .

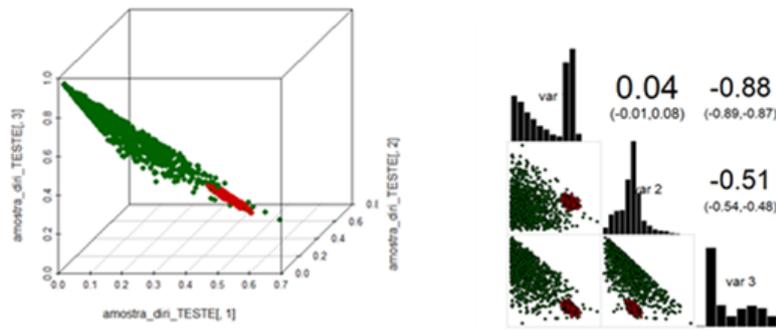


Figura 4.26: Histograma, dispersão e correlação da Estrutura 4 - Dados Transformados.

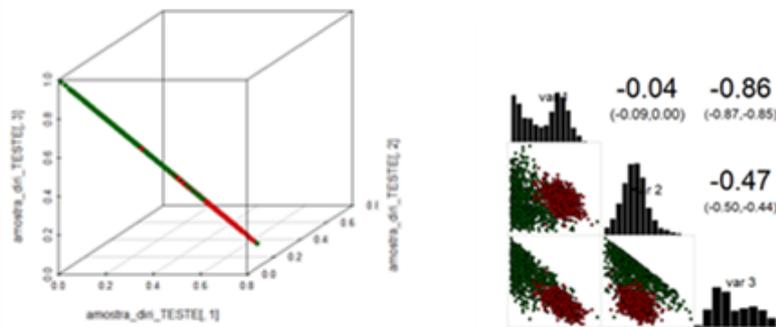


Figura 4.27: Histograma, dispersão e correlação da Estrutura 5 - Dados Transformados.

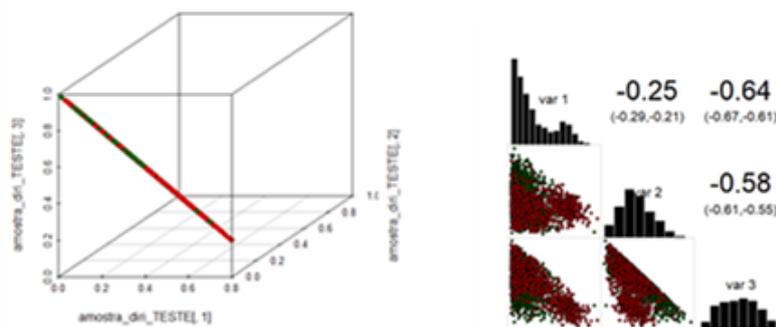


Figura 4.28: Histograma, dispersão e correlação da Estrutura 6 - Dados Transformados.

Até mesmo no caso que a intenção desse problema fosse que não existissem erros de classificação, após a transformação dos dados, nota-se que ocorre aumento significativo de erro de classificação (veja Figuras 4.29, 4.30 e 4.31).

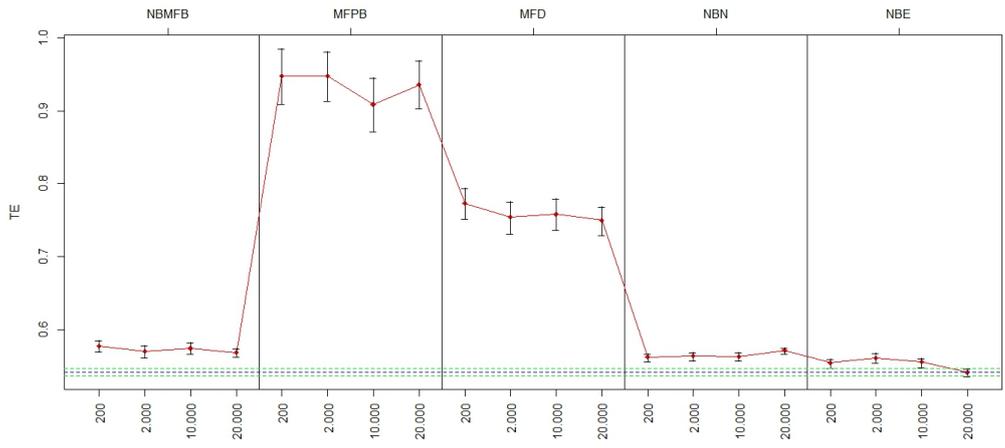


Figura 4.29: Média e IC (95%) da TE para Estrutura 4 - Dados Transformados.

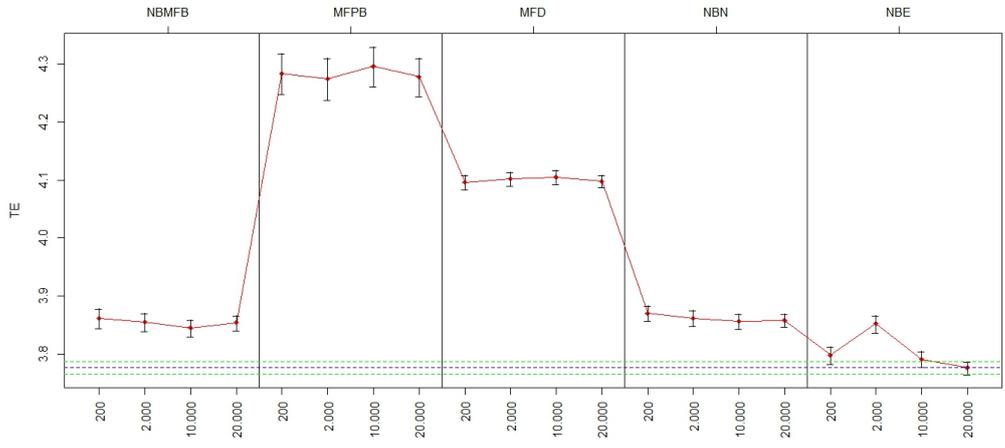


Figura 4.30: Média e IC (95%) da TE para Estrutura 5 - Dados Transformados.

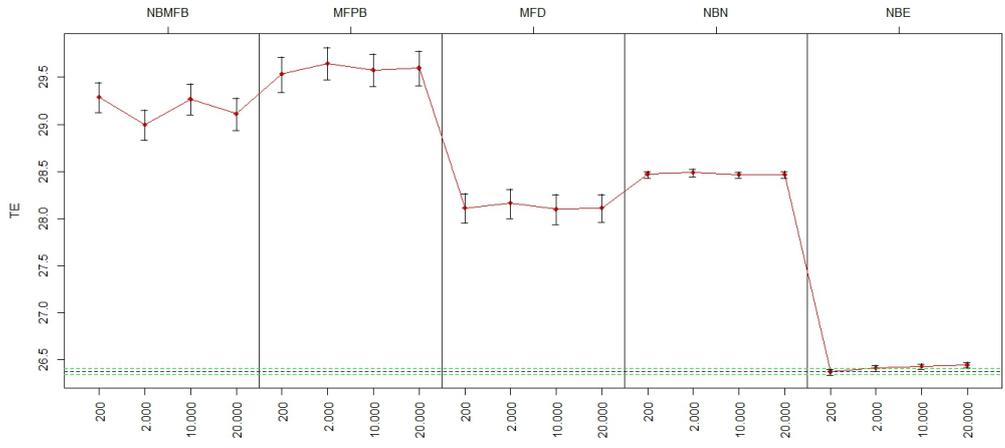


Figura 4.31: Média e IC (95%) da TE para Estrutura 6 - Dados Transformados.

Com os dados transformados para a Estrutura 4, a menor média da TE obtida foi com NBE (0,54%), com a maior amostra de treino. O classificador que apresentou a maior média para TE foi o modelo verdadeiro, MFPB (0,94%) para $n_{treino} = 200, 2.000$, (veja Tabela 4.27). Não existe diferença significativa entre a média das TE's dos modelos NBMFB (0,57%) e NBE (0,56%) para amostras treinos $n_{treino} = 2.000$, sendo considerados os melhores classificadores para os dados transformados, (veja Tabela 4.27). Para este tipo cenário da simulação, nota-se que com a transformação dos dados aumenta o erro de classificação.

Apesar da média das TE da Estrutura 5 (Dado Transformados), não terem diminuído significativamente, de maneira geral, com o aumento do conjunto de treinamento ocorre diminuição da variabilidade, (ver Figura 4.30). A maior média da TE obtida foi para o modelo verdadeiro MFPB, média de 4,29%, com amostra treino $n_{treino} = 10.000$. O classificador que apresentou a menor média para TE foi com o modelo NBE (3,77%) para $n_{treino} = 20.000$, (veja Tabela 4.27). Os resultados sugerem que quando temos a situação de amostras de MFPB com classes pouco sobrepostas, nota-se que com a transformação dos dados aumenta o erro de classificação.

A Figura 4.31, referente a Estrutura 6 (Dados Transformados), demonstra que os modelos NBMFB, MFPB e MFD tem grande variabilidade para a TE. Não existe influência significativa do tamanho da amostra treino. Com a transformação dos dados a maior média da TE obtida foi o modelo verdadeiro, MFPB (29,65%), com amostra de treino $n_{treino} = 2.000$. O ajuste que apresentou a menor média para TE foi NBE (26,376%) para $n_{treino} = 200$, (veja Tabela 4.27). Um fato interessante é que neste problema para os dados originais o ajuste pelo modelo verdadeiro obteve o melhor desempenho, porém, para os dados transformados o melhor foi NBE.

Tabela 4.26: Média e desvio-padrão da TE das Simulações de Amostras MFPB.

Estrutura	Amostra Treino	NBMFB		MFPB		NBN		NBE		ADL		ADQ	
		Média	DP										
4	200	0,00	0,00	0,01	0,00	0,01	0,01	0,01	0,00	0,11	0,06	0,01	0,01
	2.000	0,00	0,00	0,01	0,00	0,01	0,01	0,01	0,00	0,11	0,07	0,01	0,01
	10.000	0,00	0,00	0,01	0,00	0,01	0,01	0,01	0,00	0,11	0,07	0,01	0,01
	20.000	0,00	0,00	0,01	0,00	0,01	0,01	0,01	0,00	0,11	0,06	0,01	0,01
5	200	0,73	0,05	1,01	0,20	0,86	0,05	0,73	0,04	1,00	0,06	0,86	0,05
	2.000	0,73	0,05	1,02	0,18	0,86	0,05	0,72	0,05	1,00	0,06	0,86	0,05
	10.000	0,73	0,05	1,01	0,19	0,86	0,05	0,72	0,04	1,00	0,06	0,86	0,05
	20.000	0,73	0,05	1,05	0,16	0,86	0,04	0,73	0,03	1,00	0,04	0,86	0,04
6	200	28,95	0,23	25,43	0,22	30,80	0,36	29,51	0,28	27,18	0,30	30,37	0,36
	2.000	28,93	0,24	25,44	0,24	30,80	0,36	29,49	0,28	27,14	0,33	30,36	0,37
	10.000	28,92	0,20	25,47	0,19	30,86	0,30	29,53	0,25	27,12	0,25	30,36	0,32
	20.000	28,92	0,19	25,46	0,19	30,81	0,32	29,56	0,26	27,14	0,25	30,38	0,32

Para amostras simuladas a partir de MFPB, os resultados na Tabela 4.26 sugerem que o classificador construído com a densidade MFPB não obteve o melhor desempenho nas Estruturas 4 e 5, porém, para situação em que as classe são muito sobrepostas apresentou ser o melhor classificador. Com o aumento da amostra treino a média e a variabilidade da TE tendem a diminuir e o grau de separação entre as classes influencia, significativamente, na qualidade da classificação. Indicando que quanto maior for a sobreposição das classes maior será a TE.

Os resultados da Tabela 4.27 são referentes as simulações com os dados transformados. Na Estrutura 4 (DT), quando não existe sobreposição das classes, a menor média da TE obtida foi com NBE (0,54%), com a maior amostra de treino. Na Estrutura 5 (DT), quando existe pouca sobreposição das classes, o classificador que apresentou a menor média para TE foi com o modelo NBE (3,77%) para $n_{treino} = 20.000$. Na Estrutura 6 (DT), quando existe muita sobreposição das classes, a menor média para TE foi do NBE (26,38%) para $n_{treino} = 200$.

Tabela 4.27: Média e desvio-padrão da TE das Simulações de Amostras MFPB - Dados Transformados (DT).

Estrutura	Amostra Treino	MFD		NBMFB		MFPB		NBN		NBE	
		Média	DP								
4 (DT)	200	0,77	0,15	0,58	0,05	0,95	0,28	0,56	0,04	0,55	0,05
	2.000	0,75	0,16	0,57	0,06	0,95	0,25	0,56	0,04	0,56	0,05
	10.000	0,76	0,15	0,57	0,06	0,91	0,27	0,56	0,04	0,56	0,05
	20.000	0,75	0,14	0,57	0,04	0,94	0,24	0,57	0,03	0,54	0,03
5 (DT)	200	4,10	0,09	3,86	0,12	4,28	0,25	3,87	0,09	3,80	0,10
	2.000	4,10	0,09	3,86	0,11	4,27	0,26	3,86	0,09	3,85	0,11
	10.000	4,11	0,09	3,85	0,10	4,30	0,25	3,86	0,09	3,79	0,10
	20.000	4,10	0,08	3,85	0,09	4,28	0,23	3,86	0,08	3,78	0,08
6 (DT)	200	28,11	1,12	29,29	1,15	29,53	1,37	28,47	0,26	26,38	0,23
	2.000	28,16	1,13	29,00	1,14	29,65	1,25	28,49	0,28	26,41	0,24
	10.000	28,10	1,13	29,27	1,17	29,58	1,22	28,47	0,23	26,43	0,20
	20.000	28,11	1,07	29,11	1,20	29,60	1,30	28,47	0,25	26,45	0,20

Para a transformação dos dados provenientes de amostras simuladas de MFPB, os resultados na Tabela 4.27 sugerem que o classificador construído com a densidade MFPB obteve o pior desempenho nas três Estruturas propostas. O grau de separação entre as classes influencia, significativamente, na qualidade da classificação. Indicando que quanto maior for a sobreposição das classes maior será a TE.

4.3.4 Simulação de Amostras de MFBI

As observações em cada classe foram geradas por Mistura Finita de Densidade Betas Independentes, com dimensão $p = 3$. Os classificadores avaliados serão: MFD, MFPB, NBMFB (modelo verdadeiro), NBN, NBE, ADL e ADQ.

O objetivo da Estrutura 7 é verificar o comportamento do classificador quando as classes estão bastante separadas uma da outra, não admitindo erro de classificação (ver Figura 4.32). A Estrutura 8 é verificar o comportamento do classificador quando existe pouca sobreposição das classes, admitindo que exista um pequeno erro de classificação (ver Figura 4.33). A Estrutura 9 é verificar o comportamento do classificador quando

existe muita sobreposição das classes simuladas (veja Figura 4.34).

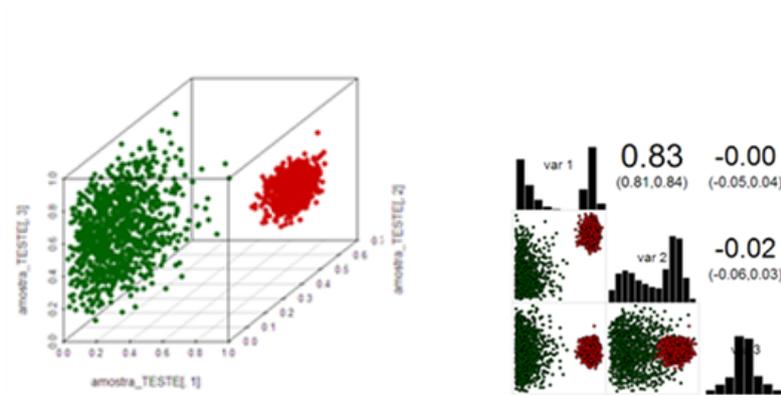


Figura 4.32: Histograma, dispersão e correlação da Estrutura 7.

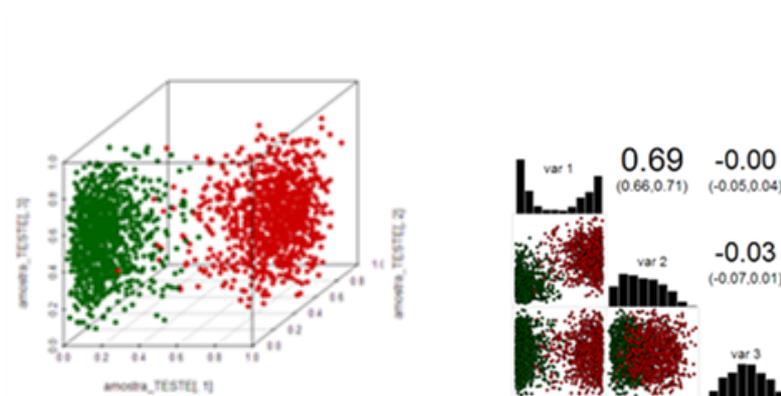


Figura 4.33: Histograma, dispersão e correlação da Estrutura 8.

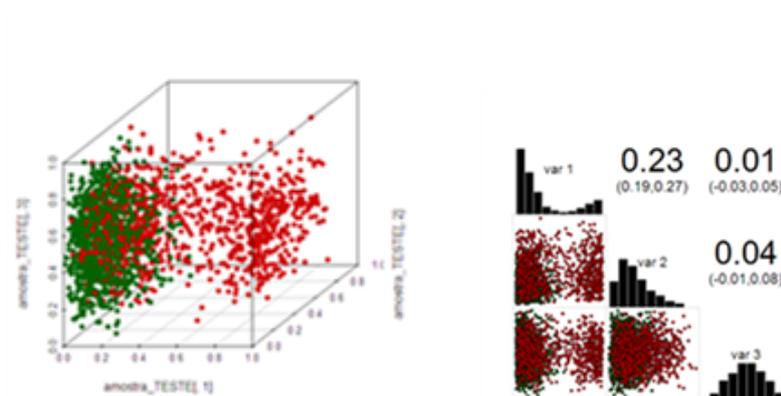


Figura 4.34: Histograma, dispersão e correlação da Estrutura 9.

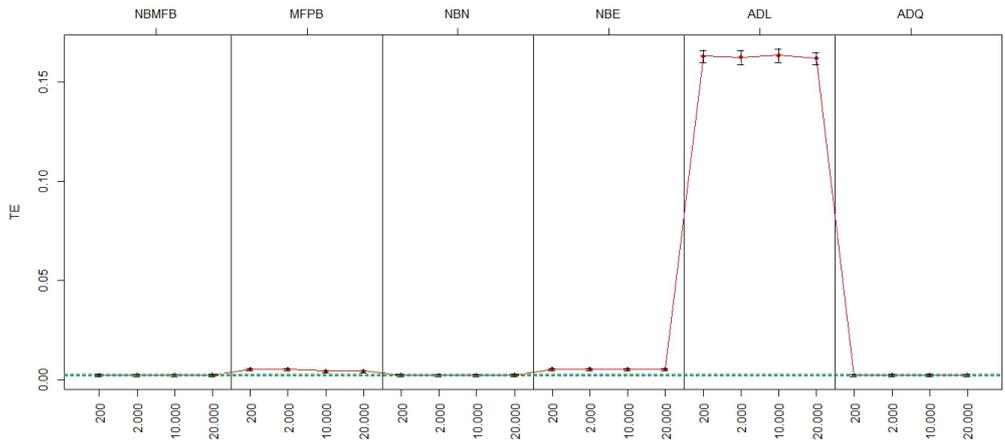


Figura 4.35: Média e IC (95%) da TE para Estrutura 7.

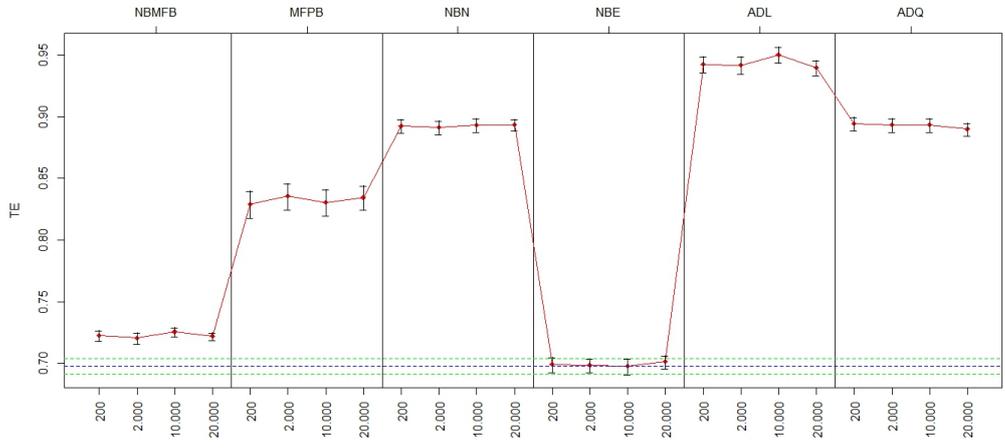


Figura 4.36: Média e IC (95%) da TE para Estrutura 8.

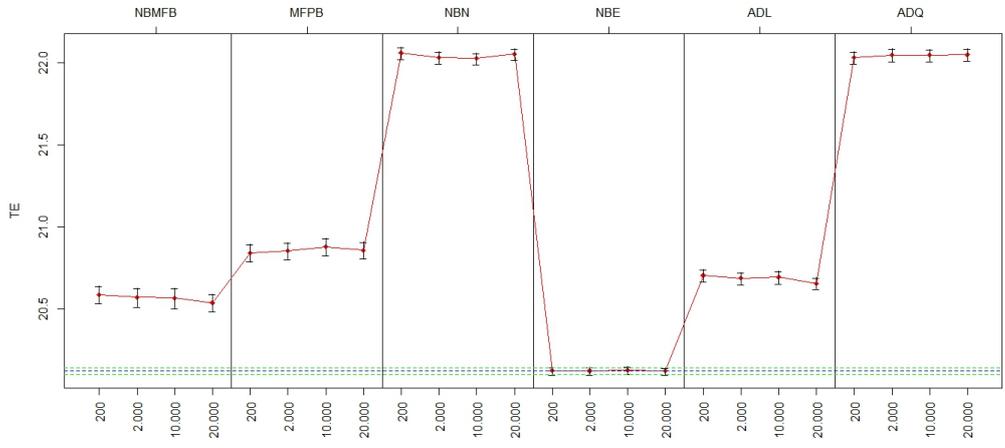


Figura 4.37: Média e IC (95%) da TE para Estrutura 9.

Os resultados sugerem que quanto mais sobrepostas estiverem as classes, maior será o erro de classificação e que muda o desempenho do classificador conforme for a sobreposição (veja Figuras 4.35, 4.36 e 4.37). Também, não existe diferença significativa para os tamanhos de amostras de treino dentro de cada classificador.

Para Estrutura 7, com os dados originais a menor média da TE obtida foi com os modelos NBMFB, NBN e ADQ, (0,002%). O ajuste que apresentou a maior média para TE foi ADL (0,163%) para $n_{treino} = 200, 2.000, 10.000$, (veja Tabela 4.28).

Observando a Figura 4.36, referente a Estrutura 8, a menor média da TE obtida foi NBE (0,70%), com amostra treino simulada ($n_{treino} = 200, 2.000, 10.000$). O ajuste que apresentou a maior média para TE foi ADL (0,95%) para $n_{treino} = 10.000$. O modelo verdadeiro NBMFB apresenta TE próxima do que foi observado no modelo NBE. A melhor média de TE do modelo NBMFB é de 0,72%, com desvio padrão 0,03%, ocorre para $n_{treino} = 2.000$, (veja Tabela 4.28).

A Estrutura 9, apresenta que a maior média da TE obtida foi para o NBN (22,06%), com amostra de treino $n_{treino} = 200$. O classificador que apresentou a menor média para TE foi NBE (20,123%) para o maior tamanho de amostra treino, (ver Tabela 4.28).

4.3.5 Simulação com Amostras de MFD Transformadas

Com a transformação dos dados ocorre pouca sobreposição das classes. Essa mistura entre as classes é possível verificar somente por meio do gráfico tridimensional (ver Figuras 4.38, 4.39 e 4.40). Fácil ver que existe um modelo de mistura em cada variável, sendo que existe uma correlação forte entre as variáveis X_1 e X_3 .

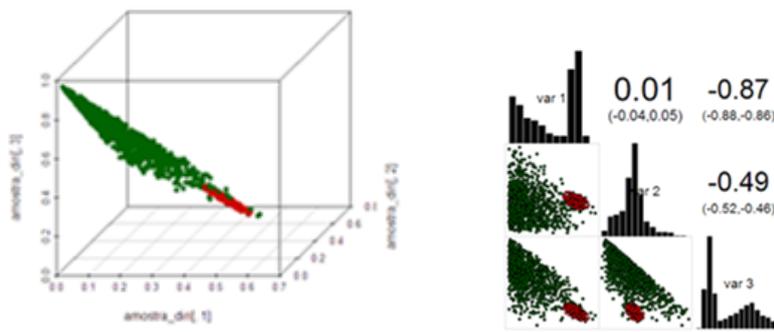


Figura 4.38: Histograma, dispersão e correlação da Estrutura 7 - Dados Transformados.

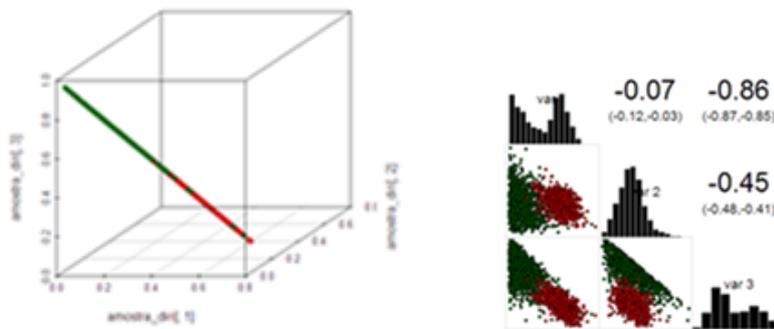


Figura 4.39: Histograma, dispersão e correlação da Estrutura 8 - Dados Transformados.

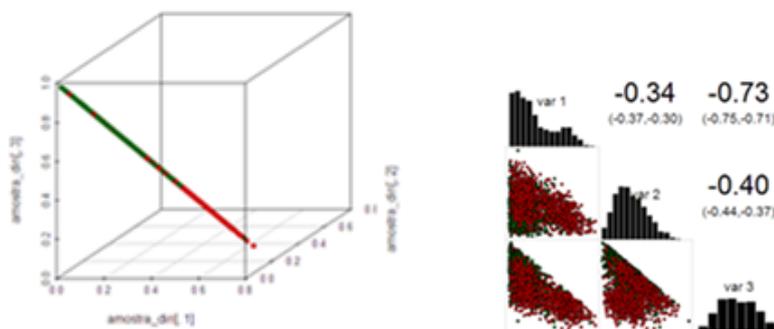


Figura 4.40: Histograma, dispersão e correlação da Estrutura 9 - Dados Transformados.

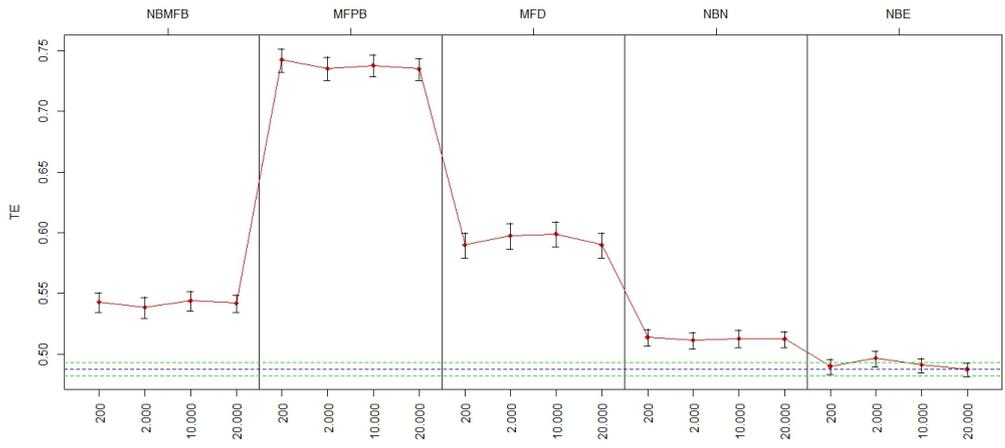


Figura 4.41: Média e IC (95%) da TE para Estrutura 7 - Dados Transformados.

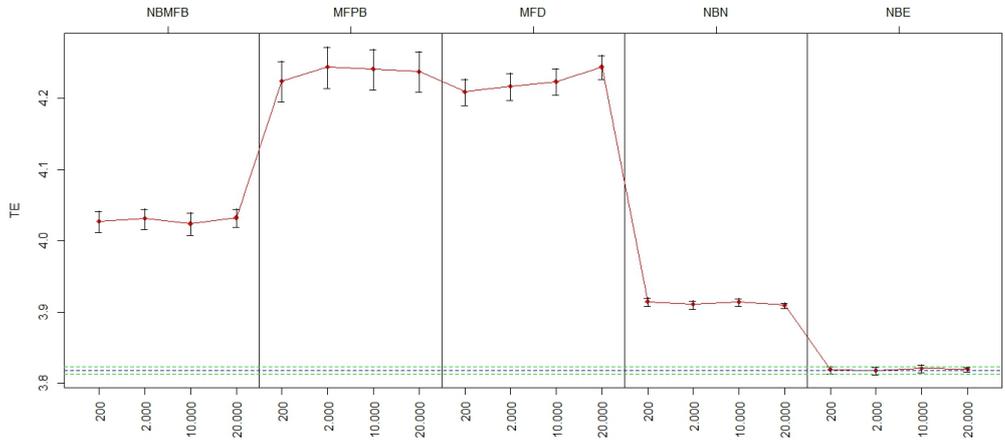


Figura 4.42: Média e IC (95%) da TE para Estrutura 8 - Dados Transformados.

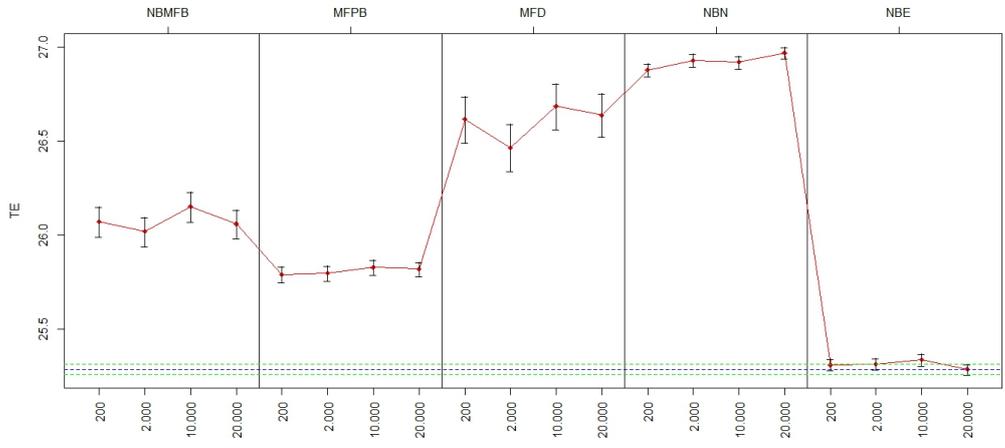


Figura 4.43: Média e IC (95%) da TE para Estrutura 9 - Dados Transformados.

Ainda para o caso que a intenção desse problema fosse que não existissem erros de classificação, após a transformação dos dados, nota-se que ocorreu aumento significativo de erro de classificação (veja Figuras 4.41, 4.42 e 4.43). Existe diferença entre os classificadores.

Para Estrutura 7 com os dados transformados a menor média da TE obtida foi com NBE (0,49%), com amostra de treino $n_{treino} = 200, 10.000, 20.000$. O classificador que apresentou a maior média para TE foi MFPB (0,74%) para todos os tamanhos de amostra treino, (veja Tabela 4.29).

Na Estrutura 8 (Dados Transformados), não existe diferença significativa quanto ao desempenho dos classificadores MFPB e MFD para os dados transformados, conforme podemos confirmar com os intervalos de confiança, (ver Figura 4.42), porém, não existe diferença significativa das TE's para os tamanhos de amostras treino dentro de cada classificador. A maior média da TE obtida foi com os modelos MFD e MFPB (4,24%), com amostra de treino $n_{treino} = 20.000$ e 2.000, respectivamente. O classificador que apresentou a menor média para TE foi com NBE (3,82%) para $n_{treino} = 2.000$, (ver Tabela 4.29). Quando comparamos a classificação dos dados originais com os transformados, nota-se um aumento significativo da TE para os dados transformados.

Mesmo com a transformação dos dados a Estrutura 9, as classes continuam muito sobrepostas, admitindo que exista um grande erro de classificação, (veja Figura 4.40). A Figura 4.43 mostra que não há evidências que as TE's diminuam com o aumento do conjunto de treinamento, porém, existe diferença entre os classificadores para os dados transformados. A maior média da TE obtida foi NBN (26,97%), com amostra treino simulada $n_{treino} = 20.000$. O ajuste que apresentou a menor média para TE foi com o modelo NBE (25,29%) para a maior amostra treino. Neste problema para os dados originais e para os dados transformados o melhor ajuste ocorre com o NBE. Os dados originais tem menor TE do que os dados transformados.

Tabela 4.28: Média e desvio-padrão da TE das Simulações de Amostras MFB Independentes.

Estrutura	Amostra Treino	NBMFB		MFPB		NBN		NBE		ADL		ADQ	
		Média	DP										
7	200	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,16	0,02	0,00	0,00
	2.000	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,16	0,02	0,00	0,00
	10.000	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,16	0,02	0,00	0,00
	20.000	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,16	0,02	0,00	0,00
8	200	0,72	0,03	0,83	0,08	0,89	0,04	0,70	0,04	0,94	0,05	0,90	0,04
	2.000	0,72	0,03	0,84	0,08	0,89	0,04	0,70	0,04	0,94	0,05	0,89	0,04
	10.000	0,73	0,02	0,83	0,07	0,89	0,04	0,70	0,05	0,95	0,05	0,89	0,04
	20.000	0,72	0,03	0,84	0,07	0,89	0,03	0,70	0,04	0,94	0,04	0,89	0,03
9	200	20,59	0,38	20,84	0,37	22,06	0,26	20,13	0,16	20,71	0,26	22,03	0,26
	2.000	20,57	0,42	20,85	0,36	22,03	0,28	20,12	0,16	20,69	0,25	22,04	0,27
	10.000	20,57	0,44	20,88	0,38	22,02	0,25	20,13	0,16	20,69	0,27	22,04	0,26
	20.000	20,54	0,38	20,86	0,36	22,05	0,25	20,12	0,16	20,66	0,24	22,05	0,26

Na Estrutura 7, quando não existe sobreposição das classes, os melhores classificadores são o NBMFB, NBN e ADQ, pois classificaram todas as observações corretamente. Na Estrutura 8, quando existe pouca sobreposição das classes, a menor média da TE obtida foi NBE (0,70%), com amostra treino simulada ($n_{treino} = 200, 2.000, 10.000$). Na Estrutura 9, quando existe muita sobreposição das classes, o classificador que apresentou a menor média para TE foi NBE (20,12%) para $n_{treino} = 2.000, 20.000$.

Para amostras simuladas a partir de NBMFB, os resultados na Tabela 4.28 sugerem que classificador construído com a densidade NBMFB não obteve o melhor desempenho nas Estruturas propostas, mas apresenta resultados muito próximo do NBE. Com o aumento da amostra treino a média e a variabilidade da TE tendem a diminuir e o grau de separação entre as classes influencia, significativamente, na qualidade da classificação. Indicando que quanto maior for a sobreposição das classes maior será a TE.

Tabela 4.29: Média e desvio-padrão da TE das Simulações de Amostras MFB Independentes - Dados Transformados (DT).

Estrutura	Amostra Treino	MFD		NBMFB		MFPB		NBN		NBE	
		Média	DP								
7 (DT)	200	0,59	0,07	0,54	0,06	0,74	0,07	0,51	0,05	0,49	0,04
	2.000	0,60	0,08	0,54	0,06	0,74	0,07	0,51	0,05	0,50	0,05
	10.000	0,60	0,07	0,54	0,06	0,74	0,06	0,51	0,05	0,49	0,04
	20.000	0,59	0,07	0,54	0,05	0,74	0,07	0,51	0,05	0,49	0,04
8 (DT)	200	4,21	0,13	4,03	0,11	4,23	0,21	3,92	0,04	3,82	0,04
	2.000	4,22	0,14	4,03	0,10	4,24	0,21	3,91	0,05	3,82	0,04
	10.000	4,22	0,13	4,02	0,11	4,24	0,21	3,91	0,04	3,82	0,04
	20.000	4,24	0,12	4,03	0,09	4,24	0,20	3,91	0,03	3,82	0,02
9 (DT)	200	26,62	0,90	26,07	0,58	25,79	0,30	26,88	0,24	25,31	0,22
	2.000	26,47	0,90	26,02	0,55	25,80	0,28	26,93	0,24	25,32	0,22
	10.000	26,69	0,87	26,15	0,58	25,83	0,29	26,92	0,25	25,34	0,23
	20.000	26,64	0,84	26,06	0,54	25,82	0,28	26,97	0,22	25,29	0,21

Os resultados da Tabela 4.29 são referentes as simulações com os dados transformados. Na Estrutura 7 (DT), quando não existe sobreposição das classes, a menor média da TE obtida foi com NBE (0,49%), com $n_{treino} = 200, 10.000, 20.000$. Na Estrutura 8 (DT), quando existe pouca sobreposição das classes, o classificador que apresentou a menor média para TE foi com o modelo NBE (3,82%) para $n_{treino} = 200, 2.000, 10.000$. Na Estrutura 9 (DT), quando existe muita sobreposição das classes, a menor média para TE foi do NBE (25,31%) para $n_{treino} = 200$.

Para a transformação dos dados provenientes de amostras simuladas de NBMFB, os resultados na Tabela 4.29 sugerem que classificador construído com a densidade NBMFB não obteve o melhor desempenho nas três Estruturas propostas. Um fato relevante é que o classificador MFPB obteve o pior desempenho para quando as classes não estão sobrepostas, porém, é o segundo melhor classificador quando as classes estão muito sobrepostas. O grau de separação entre as classes influencia, significativamente, na qualidade da classificação. Indicando que quanto maior for a sobreposição das classes maior será a TE.

Capítulo 5

Aplicação em Dados Reais

Nesta seção empregaremos a metodologia proposta neste trabalho considerando dois conjuntos de dados reais, referente a pixel de imagem em RGB. Foram avaliados os modelos mais usuais em RPS: Análise Discriminante Linear (ADL), Análise Discriminante Quadrática (ADQ), Naive Bayes com distribuição Normal (NBN) e Naive Bayes com distribuição não-paramétrica de Epanechnikov (NBE), assim como, os novos modelos abordados: Naive Bayes de Mistura Finita de Betas (NBMFB), Mistura Finita de Produtórios de Betas (MFPB) e Mistura Finita de Dirichlet (MFD). Além de separar as observações em conjunto de treinamento e teste, empregamos o procedimento conhecido como "*k-fold*".

5.1 Noções sobre o espaço de cores RGB

Os dados em RGB são baseados em um sistema de coordenadas cartesianas, sendo o espaço de cores um cubo. As cores primárias R (vermelho, "red"), G (verde, "green") e B (azul, "blue") estão em três vértices do cubo. As cores secundárias, ciano, magenta e amarelo estão em outros três vértices. O vértice junto à origem é o preto e o mais afastado da origem corresponde à cor branca, veja Figura 5.1. A escala de cinza se estende através da diagonal do cubo, ou seja, o segmento de reta que une origem (preto) até o vértice mais

distante (branco), tomando valores de [0;255] em cada cor.

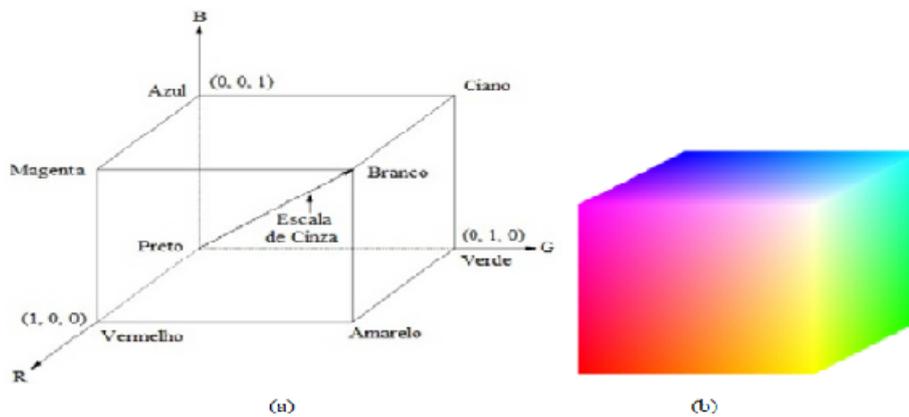


Figura 5.1: a) Modelo RGB e b) Cubo RGB de 24-bit.
Fonte: Gonzales & Woods (2002)

Para aplicação dos modelos proposto neste trabalho, foi realizado dois tipos de transformação desses valores, (veja Ma & Leijon (2010) e Bouguila *et al.* (2006)):

$$\text{Transformação 1 } (T1) = \frac{x_{id}}{255} = v_{id}, \quad i = 1, \dots, n; \quad d \in \{r, g, b\}. \quad (5.1)$$

$$\text{Transformação 2 } (T2) = \frac{x_{id}}{x_{ir} + x_{ig} + x_{ib}} = v_{id}, \quad i = 1, \dots, n; \quad d \in \{r, g, b\}. \quad (5.2)$$

Para resolver o problema de valores extremos, fizemos uso $\varepsilon = 0,003$, onde

- i) se $x_{id} = 0 \Rightarrow v_{id} = \varepsilon$;
- ii) se $x_{id} = 255 \Rightarrow v_{id} = 1 - \varepsilon$.

5.2 Estudo de Caso 1: Imagens RGB de Pele e Não-Pele

O conjunto de dados de imagens em RGB de Pele e Não -Pele foi coletado por amostragem aleatória a partir de imagens de rosto de vários grupos etários (jovens, adultos e idosos), grupos de raça (branco, preto e asiático) e gêneros, obtidos no "UCI Repository" (www.ics.uci.edu/mllearn/MLRepository.html), denominado "Skin Detection". O

tamanho total da amostra de aprendizagem é 245.057; dos quais 50.859 são amostras de Pele e 194.198 são amostras Não-Pele. Este conjunto de dados é de dimensão 245.057×4 , o qual tem três colunas que são os valores R, G, B e uma coluna de rótulos das classes. Na Figura 5.2 podemos verificar a dispersão dos dados após a transformação. Na Figura 5.2.(a) temos a transformação (T1), nota-se um grau razoável de separação entre as classes e na Figura 5.2.(b) temos a transformação (T2), onde observamos uma concentração dos dados referente a Pele, porém, a classe Não-Pele encontram-se bastante dispersos.

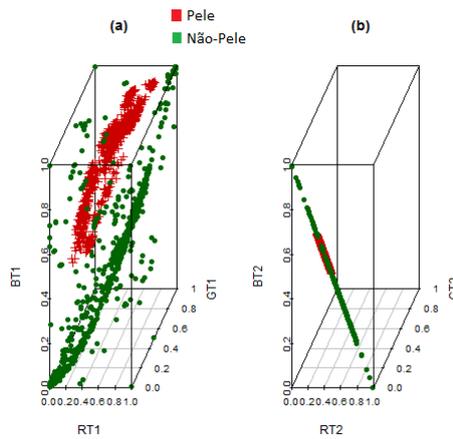


Figura 5.2: Gráfico 3D para Dados Pele e Não-Pele. (a) Transformação 1 e (b) Transformação 2.

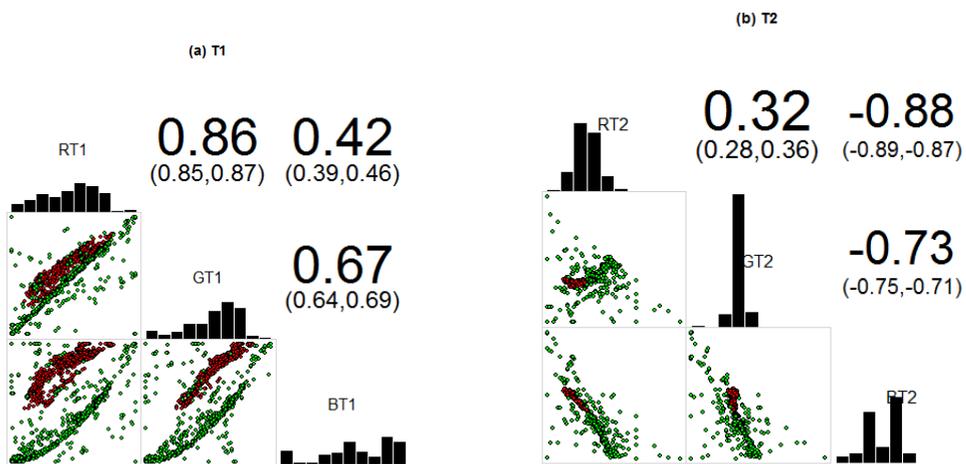


Figura 5.3: Histograma, dispersão e correlação para Dados Pele e Não-Pele. (a) T1 e (b) T2.

Na Figura 5.3.(a) temos os gráficos de histograma, dispersão e correlação dos Dados Pele e Não-Pele para T1, nota-se um grau razoável de separação entre as classes, forte

correlação entre as variáveis, pelo histograma nota-se que os dados seguem um modelo de mistura. Na Figura 5.3.(b) temos a transformação (T2), onde observamos uma concentração dos dados referente a Pele, forte correlação negativa entre algumas variáveis e os dados seguem um modelo de mistura. A seguir os gráficos de histograma de cada classe para os dados transformados T1 e T2.

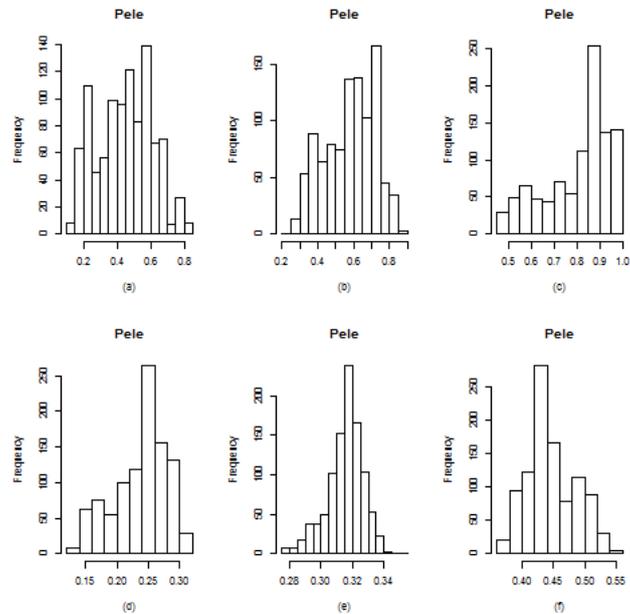


Figura 5.4: Histogramas dos dados da classe Pele. (a) RT1, (b) GT1, (c) BT1, (d) RT2, (e) GT2 e (f) BT2.

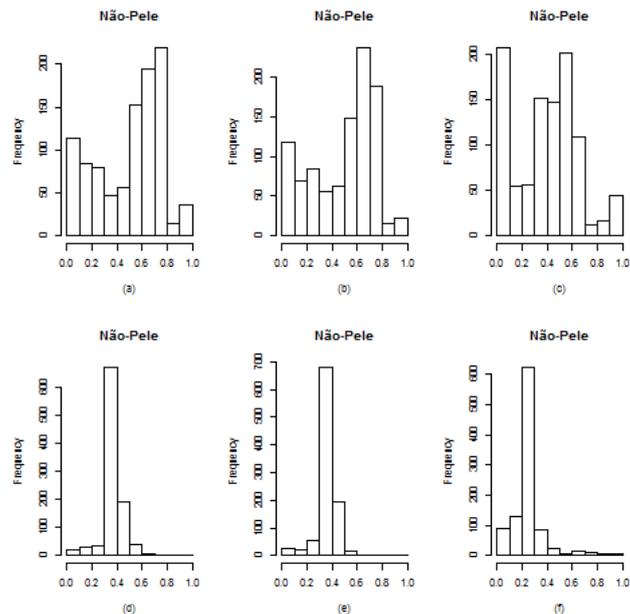


Figura 5.5: Histogramas dos dados da classe Não-Pele. (a) RT1, (b) GT1, (c) BT1, (d) RT2, (e) GT2 e (f) BT2.

Note que de uma maneira geral, as variáveis apresentam multimodalidade, característica intrínseca aos modelos de mistura finita de densidades, tanto para os dados T1 como para T2.

O procedimento para avaliar os classificadores foi baseado no conjunto Treino e Teste, para diferentes tamanho de conjunto de treinamento

$$n_{treino} = (200, 600, 1.000, 2.000, 10.000, 20.000),$$

sendo que cada conjunto de treino é formado de 21% da classe Pele e 79% da classe Não-Pele. A amostra teste de tamanho $n_{teste} = 20.000$, sendo 21% da classe Pele e 79% da classe Não-Pele. Foi realizada 200 repetições do experimento e obtida a média da Taxa de Erro (TE) e os seus respectivos intervalos de confiança (95%). Para a transformação T1 foram avaliados os modelos ADL, ADQ, NBN, NBE, NBMFB e MFPB; para a transformação T2 foram avaliados os modelos NBN,NBE, NBMFB, MFPB e MFD.

A seguir apresentamos os resultados obtidos para cada tamanho de conjunto treino segundo cada classificador com os dados da transformação T1.

Tabela 5.1: TE de classificação dos Dados Pele e Não-Pele da T1.

Treino	ADL (T1)	ADQ (T1)	NBN (T1)	NBE (T1)	NBMFB (T1)	MFPB (T1)
200	4,02 [3,97;4,06]	1,25 [1,21;1,28]	11,00 [10,95;11,04]	4,40 [4,31;4,48]	50,00 [49,73;50,26]	1,05 [1,00;1,09]
600	4,60 [4,55;4,64]	1,30 [1,26;1,33]	10,50 [10,45;10,54]	5,75 [5,66;5,83]	45,90 [45,63;46,16]	1,70 [1,65;1,74]
1.000	3,80 [3,75;3,84]	0,75 [0,71;0,78]	10,45 [10,40;10,49]	4,35 [4,26;4,43]	51,60 [51,33;51,86]	1,85 [1,80;1,89]
2.000	3,85 [3,80;3,89]	1,25[1,21;1,28]	11,00 [10,95;11,04]	4,30 [4,21;4,38]	49,35 [49,08;49,61]	2,15 [2,10;2,19]
10.000	3,90 [3,86;3,93]	0,75 [0,71;0,78]	10,40 [10,35;10,44]	4,15 [4,06;4,23]	50,05 [49,78;50,31]	1,65 [1,60;1,69]
20.000	3,90 [3,86;3,93]	0,80 [0,77;0,82]	10,55 [10,51;10,58]	4,10 [4,02;4,17]	49,90 [49,64;50,15]	1,70 [1,65;1,74]

Para este conjunto de dados com a transformação (T1), o aumento do tamanho do conjunto de treinamento, em alguns casos, os resultados sugerem uma diminuição da média das TE para alguns classificadores, (veja Tabela 5.1). A maior média da TE obtida foi para o modelo NBMFB (51,60%), com a amostra treino simulada $n_{treino} = 1.000$. O ajuste que apresentou a menor média para TE foi com o modelo a ADQ (0,75%) para $n_{treino} = 1.000$ e 10.000. Porém, o melhor modelo com o menor conjunto treino

$n_{treino} = 200$, foi o modelo MFPB com TE de (1,05%). Analisando so IC's estes sugerem que a diferença do desempenho entre os classificadores ADQ e MFPB nesta situação é significativa.

Tabela 5.2: TE de classificação dos Dados Pele e Não-Pelel da T2.

Treino	NBN (T2)	NBE (T2)	NBMFB (T2)	MFPB (T2)	MFD (T2)
200	0,55 [0,53;0,56]	0,95 [0,91;0,98]	0,35 [0,34;0,35]	0,90 [0,87;0,92]	0,65 [0,63;0,66]
600	0,35 [0,33;0,36]	0,40 [0,37;0,42]	0,40 [0,39;0,41]	0,60 [0,57;0,62]	0,80 [0,79;0,81]
1.000	0,30 [0,28;0,31]	0,60 [0,57;0,62]	0,45 [0,44;0,46]	0,45 [0,42;0,47]	0,60 [0,59;0,61]
2.000	0,35 [0,33;0,36]	0,30 [0,27;0,32]	0,40 [0,39;0,41]	0,45 [0,42;0,47]	0,70 [0,69;0,71]
10.000	0,35 [0,33;0,36]	0,50 [0,47;0,52]	0,35 [0,34;0,36]	0,45 [0,42;0,47]	0,60 [0,59;0,61]
20.000	0,35 [0,33;0,36]	0,50 [0,47;0,52]	0,40 [0,39;0,41]	0,45 [0,42;0,47]	0,75 [0,74;0,75]

Para os dados da transformação (T2), em alguns casos, o aumento do tamanho do conjunto de treinamento proporcionou uma diminuição da média das TE, (veja Tabela 5.2). A maior média da TE obtida foi para a distribuição com NBE (0,95%), com a amostra treino simulada $n_{treino} = 200$. Os ajustes que apresentaram a menor média para TE são os modelos NBN e NBE com TE de (0,30%). A análise dos IC's sugere que existe diferenças significativas entre os classificadores.

Comparando os resultados da T1 e T2, observamos uma melhora significativa na classificação dos pixel em todos os classificadores com T2. O modelo NBMFB, apresentou ser o melhor modelo com o menor conjunto treino $n_{treino} = 200$, com TE de 0,35%.

5.3 Estudo de Caso 2: Imagens RGB de Baciloscopia

O conjunto de imagens utilizado foi obtido por Junior *et al.* (2010). As imagens correspondem a campos de lâminas contendo esfregaço de secreção pulmonar de pacientes. Esses esfregaços foram preparados com a coloração Kinyoun, no Instituto Nacional de Pesquisas da Amazônia (INPA). Esse conjunto, constituído de 120 imagens de baciloscopia de campo claro, com resolução espacial de 2816x2112 pixels, é proveniente de

12 lâminas baciloscópicas, veja Figuras 5.6 e 5.7.

O objetivo é classificar os pixels nessas imagens como pertencente a uma região de "bacilo" *Mycobacterium tuberculosis* ou a região de "não bacilo".

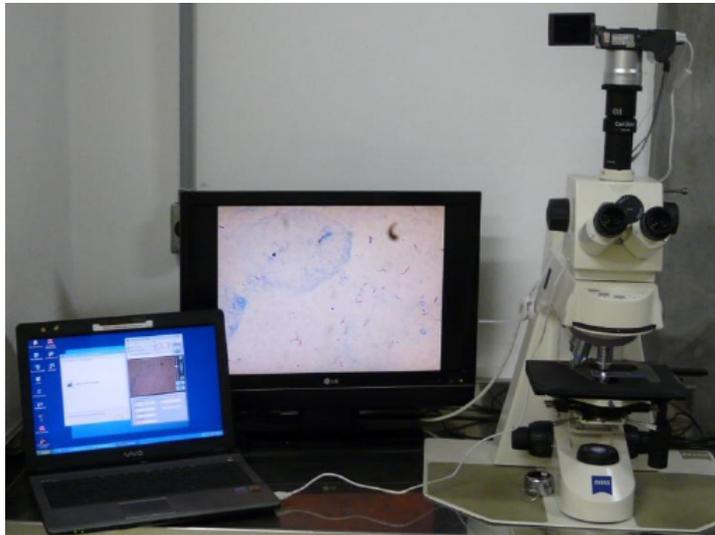


Figura 5.6: Ambiente para a aquisição das imagens.
Fonte: Kimura Junior (2010).

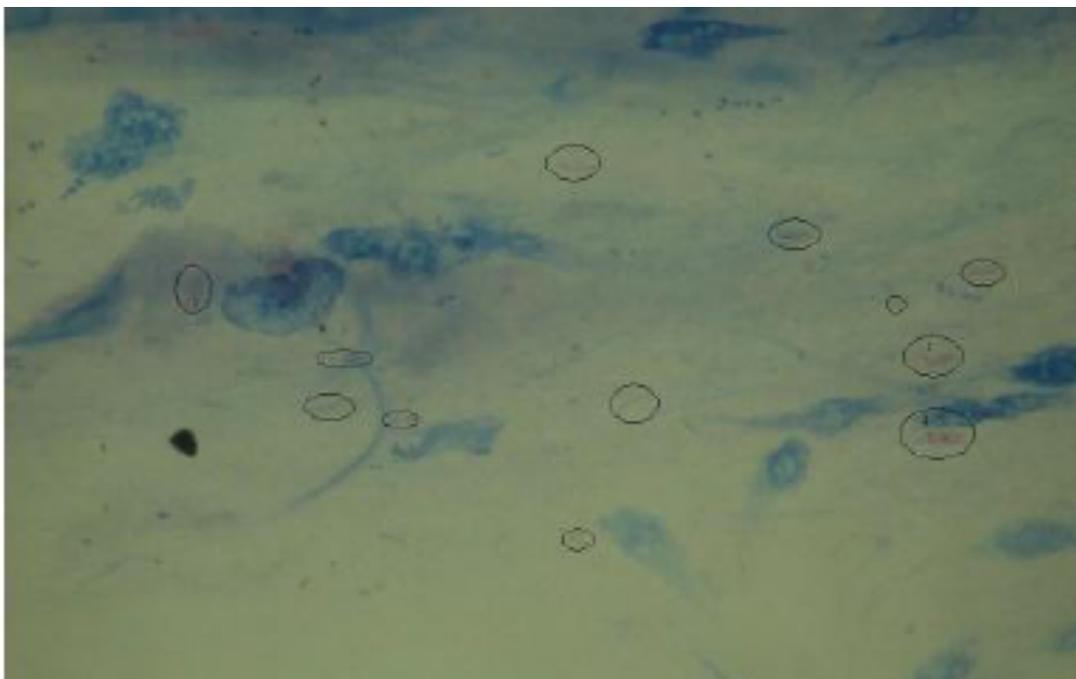


Figura 5.7: Imagem obtida por microscopia com os bacilos marcados por um especialista.
Fonte: Costa *et al.* (2008).

A composição da amostra foi formada por 4.800 pixels (metade representando

bacilos e a outra metade representando o fundo) no formato RGB, obtidos pela extração de 20 pixels de regiões correspondentes a bacilos e 20 pixels de regiões que correspondiam a áreas de fundo de cada uma das 120 imagens disponíveis. Nota-se que as classes estão sobrepostas com ambas transformações dos dados, veja Figura 5.8.

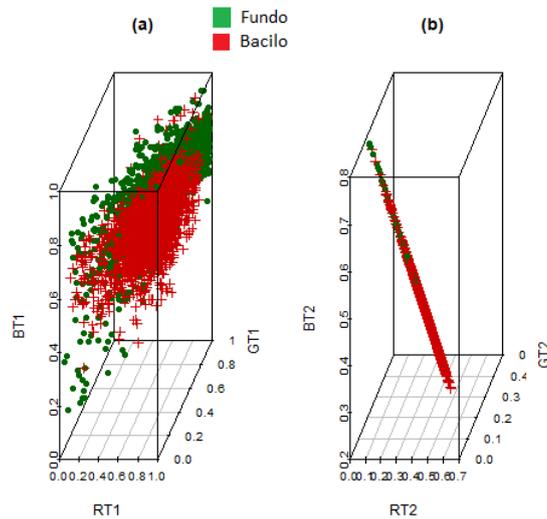


Figura 5.8: Gráfico 3D para Dados de Baciloscopia. (a) Transformação 1 e (b) Transformação 2.

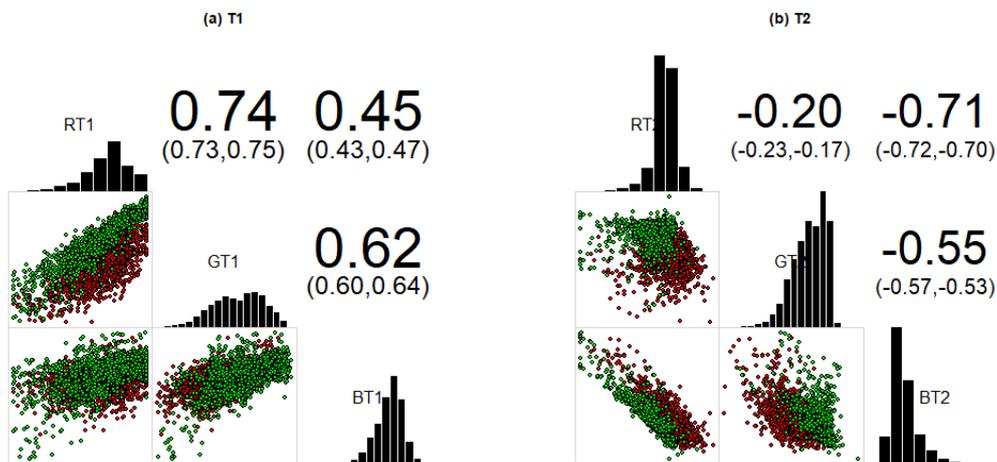


Figura 5.9: Histograma, dispersão e correlação para Dados de Baciloscopia. (a) T1 e (b) T2.

Na Figura 5.9.(a) temos os gráficos de histograma, dispersão e correlação dos Dados Baciloscopia para T1, nota-se muita sobreposição das classes, forte correlação entre algumas variáveis. O histograma nota-se que na variável GT1 os dados seguem um

modelo de mistura. Na Figura 5.9.(b) temos a transformação (T2), onde observamos que também as classes estão sobrepostas, forte correlação negativa entre algumas variáveis e os dados na variável GT2 seguem um modelo de mistura.

A seguir os histograma dos Dados de Baciloscopia para T1 e T2, sendo apresentado os gráficos de cada classe separadamente. De uma maneira geral na classe "Bacilo"(Figura 5.10) as variáveis apresentam comportamento distintos, com assimetrias e caldas alongadas, característica intrínseca aos modelos de mistura finita de densidades, principalmente, para transformação T2.

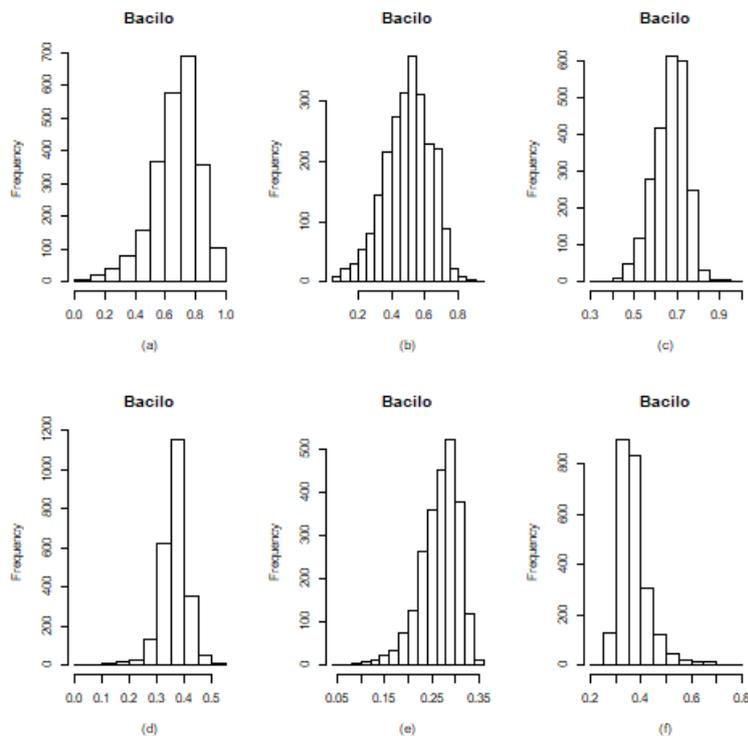


Figura 5.10: Histogramas dos dados da classe Bacilo. (a) RT1, (b) GT1, (c) BT1, (d) RT2, (e) GT2 e (f) BT2.

Na Figura5.11) as variáveis da classe "Fundo"apresentam comportamento semelhante as citadas para o grupo "Bacilo". Os gráficos sugerem que na classe "Bacilo"não apresenta modelo de mistura de densidade.

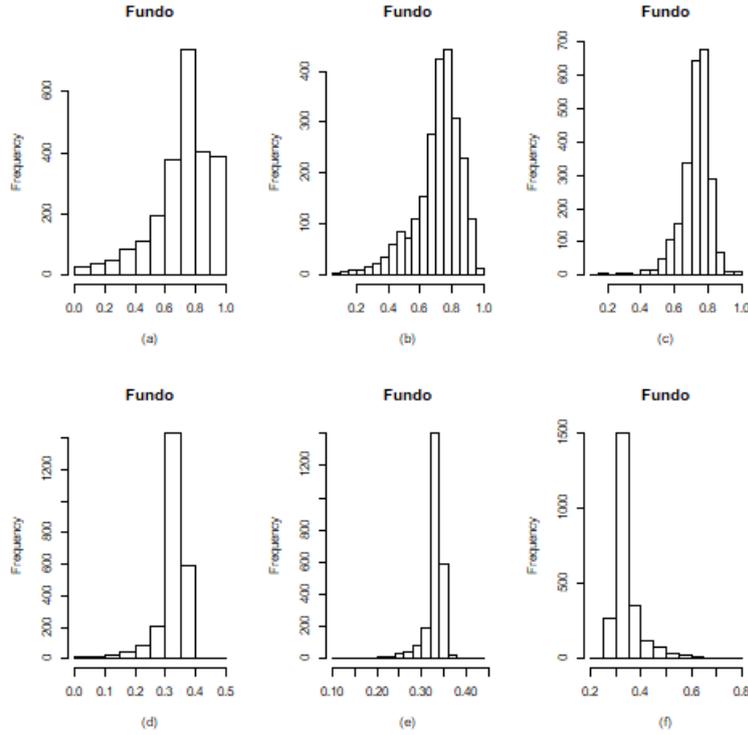


Figura 5.11: Histogramas dos dados da classe Fundo. (a) RT1, (b) GT1, (c) BT1, (d) RT2, (e) GT2 e (f) BT2.

O procedimento para avaliar os classificadores foi baseado no conjunto Treino e Teste, com validação cruzada método "*k-fold*", dividindo o conjunto em $k = 5$ partes. Com amostra de treinamento $n_{treino} = 3.840$ (1.920 pixel de Bacilo e 1.920 pixel de Fundo) Com amostra teste de tamanho $n_{teste} = 960$ (480 pixel de Bacilo e 480 pixel de Fundo). Para a transformação T1 foram avaliados os modelos ADL, ADQ, NBN, NBE, NBMFB e MFPB; para a transformação T2 foram avaliados os modelos NBN, NBE, NBMFB, MFPB e MFD.

Foi obtida as médias da Taxa de Erro (TE), Acurácia (A), Sensibilidade (S) e Especificidade (E). A acurácia nada mais é do que a taxa de acerto, portanto o complementar da taxa de erro, (veja Levy(1999)):

$$A = \frac{VP + VN}{VP + FP + VN + FN} \quad (5.3)$$

$$S = \frac{VP}{VP + FN} \quad (5.4)$$

$$E = \frac{VN}{VN + FP} \quad (5.5)$$

- VP: verdadeiro positivo, classificar o pixel de bacilo como sendo bacilo;
- VN: verdadeiros negativos, classificar o pixel de fundo como sendo fundo;
- FP: falsos positivos, classificar o pixel de fundo como sendo bacilo;
- FN: falsos negativos, classificar o pixel de bacilo como sendo fundo.

A seguir apresentamos os resultados obtidos para classificação dos Dados Baciloscopia com transformação T1 e T2.

Tabela 5.3: Resultado da classificação dos dados da Baciloscopia da T1.

Classificador	TE(%)	A(%)	S(%)	E(%)
NBMFB	21,31	79,27	84,71	72,67
MFPB	15,46	84,79	88,38	80,71
NBN	22,40	78,96	78,67	76,54
NBE	21,17	78,85	85,96	71,71
ADL	11,10	88,75	84,17	93,63
ADQ	9,56	90,31	86,63	94,25

Tabela 5.4: Resultado da classificação dos dados da Baciloscopia da T2.

Classificador	TE(%)	A(%)	S(%)	E(%)
MFD	13,83	86,35	83,00	89,33
NBMFB	13,00	86,67	87,71	86,29
MFPB	50,58	50,10	50,54	48,29
NBN	14,13	87,19	78,25	93,50
NBE	10,83	88,96	90,67	87,67

Os resultados na Tabela 5.3 são referente a classificação dos pixel de imagens do exame de Baciloscopia com transformação dos dados (T1). O ajuste que apresentou a menor média para TE foi com o modelo a ADQ (9,56%), acurácia de 90,31% e especificidade de 94,25%. O modelo MFPB apresentou a melhor sensibilidade (88,38%), indicando ser o melhor classificador par identificar os pixel de bacilo.

Na Tabela 5.4 temos os resultados da classificação dos pixel de imagens do exame de Baciloscopia com transformação dos dados (T2). O ajuste que apresentou a menor média para TE foi o modelo a NBE (10,83%), acurácia de 88,96% e sensibilidade de 90,67%. O modelo NBN apresentou a melhor especificidade (93,50%).

Um fato interessante comparando o erro de classificação nas duas transformações é que nos modelos NBMFB, NBN houve diminuição da TE. Para os modelos NBE e MFPB ocorreu aumento da TE, principalmente, para o modelo MFPB, que com T1 teve TE igual 15,46% com T2 passou a ser 50,58%.

Capítulo 6

Considerações Finais

Neste trabalho abordamos o problema de estimação das densidades condicionais nas classes em Análise Discriminante, considerando o caso em que as observações do vetor de características estejam contidas no intervalo $(0,1)$. Empregamos como modelos para estas distribuições misturas finitas de produtos de densidades Beta, misturas finitas de densidades de Dirichlet e o procedimento Naive Bayes com misturas finitas de densidades Beta.

Os procedimentos de investigação e avaliação dessas modelagens foram desenvolvidas através de estudos de simulação, com o qual foi possível observar e analisar as estimativas dos valores das densidades pontualmente, dos parâmetros e das taxas de erros de classificação.

Nas simulações cujo objetivo foi analisar o comportamento das estimativas das densidades em valores pontuais, notou-se que ambos os modelos conseguiram se ajustar as diferentes situações simuladas, porém, observou-se também que o número de componentes na mistura influenciam na qualidade da estimação da densidade, principalmente, para pequenas amostras.

Os resultados das análises das estimativas dos parâmetros sugerem que para o modelo de mistura finita de produtos de densidades Betas ocorre melhores resultados. Também, para o modelo de misturas finitas de densidades de Dirichlet observou-se maiores

dificuldades na estimação dos parâmetros, principalmente, para tamanho de amostras pequeno. No entanto, verificou-se que em ambos modelos as estimativas tendem a convergir para o verdadeiro valor do parâmetro com o aumento do tamanho da amostra.

Com relação aos experimentos que envolviam classificação, os modelos empregados no classificador de Bayes se ajustaram às diferentes estruturas simuladas, desde as mais simples até as mais complexas em termos de separação das classes, com taxas de erros bastantes razoáveis quando comparadas as dos outros modelos mais usuais que foram considerados para fins de comparação.

Nas aplicações com dados reais, os classificadores conseguiram se ajustar as observações do problema, apresentando dificuldades, em termos das taxas de erro, no caso de poucas observações por classe.

De modo geral, considerando todos os resultados obtidos nos estudos de simulações e na aplicação com dados reais, os classificadores com modelo de mistura finita de produtos de densidades Beta e de mistura finita de densidades de Dirichlet mostraram-se competitivos com os classificadores mais usuais considerados no estudo, sendo uma alternativa nos problemas de Análise Discriminante abordados, em que as observações das variáveis no vetor de característica têm seus valores no intervalo $(0;1)$.

O procedimento de estimação nos modelos investigados foi de máxima verossimilhança com emprego do algoritmo EM (Expectation and Maximization) para obter as estimativas. Uma sugestão para ampliar os estudos realizados, seria empregar procedimentos de estimação bayesiana nos mesmos casos simulados e os mesmos conjuntos de dados reais considerados no estudo realizado.

Apêndice

Os modelos descritos a seguir são referentes ao Estudo de Simulação 3.

A Estrutura 1 é proveniente de MFD tridimensional com $l = 2$:

$$\text{Classe1} : f_1(\mathbf{x}) = 0,5 * \text{Dir}(\mathbf{x}; 1; 200; 500) + 0,5 * \text{Dir}(\mathbf{x}; 1; 100; 200).$$

$$\text{Classe2} : f_2(\mathbf{x}) = 0,5 * \text{Dir}(\mathbf{x}; 10; 2; 5) + 0,5 * \text{Dir}(\mathbf{x}; 40; 2; 5).$$

A Estrutura 2 é proveniente de MFD tridimensional com $l = 2$:

$$\text{Classe1} : f_1(\mathbf{x}) = 0,5 * \text{Dir}(\mathbf{x}; 10; 2; 50) + 0,5 * \text{Dir}(\mathbf{x}; 4; 3; 40).$$

$$\text{Classe2} : f_2(\mathbf{x}) = 0,5 * \text{Dir}(\mathbf{x}; 1; 2; 2) + 0,5 * \text{Dir}(\mathbf{x}; 1; 2; 2).$$

A Estrutura 3 é proveniente de MFD tridimensional com $l = 2$:

$$\text{Classe1} : f_1(\mathbf{x}) = 0,5 * \text{Dir}(\mathbf{x}; 10; 2; 0,5) + 0,5 * \text{Dir}(\mathbf{x}; 1; 1; 0,2).$$

$$\text{Classe2} : f_2(\mathbf{x}) = 0,5 * \text{Dir}(\mathbf{x}; 10; 2; 5) + 0,5 * \text{Dir}(\mathbf{x}; 40; 2; 5).$$

A Estrutura 4 é proveniente de MFPB tridimensional com $l = 2$:

$$\begin{aligned} \text{Classe1} : f_1(\mathbf{x}; \Phi) &= 0,4[\text{Beta}(x_1; 1; 8) * \text{Beta}(x_2; 2; 8) * \text{Beta}(x_3; 3; 3)] \\ &+ 0,6[\text{Beta}(x_1; 1; 8) * \text{Beta}(x_2; 2; 8) * \text{Beta}(x_3; 3; 3)]. \end{aligned}$$

$$\begin{aligned} \text{Classe2} : f_2(\mathbf{x}; \Phi) &= 0,6[Beta(x_1; 50; 10) * Beta(x_2; 50; 50) * Beta(x_3; 30; 30)] \\ &+ 0,4[Beta(x_1; 50; 10) * Beta(x_2; 50; 50) * Beta(x_3; 30; 30)]. \end{aligned}$$

A Estrutura 5 é provenientes de MFPB tridimensional com $l = 2$:

$$\begin{aligned} \text{Classe1} : f_1(\mathbf{x}; \Phi) &= 0,4[Beta(x_1; 1; 8) * Beta(x_2; 2; 8) * Beta(x_3; 3; 3)] \\ &+ 0,6[Beta(x_1; 1; 8) * Beta(x_2; 2; 8) * Beta(x_3; 3; 3)]. \end{aligned}$$

$$\begin{aligned} \text{Classe2} : f_2(\mathbf{x}; \Phi) &= 0,6[Beta(x_1; 5; 1) * Beta(x_2; 5; 5) * Beta(x_3; 3; 3)] \\ &+ 0,4[Beta(x_1; 5; 1) * Beta(x_2; 5; 5) * Beta(x_3; 3; 3)]. \end{aligned}$$

A Estrutura 6 é provenientes de MFPB tridimensional com $l = 2$:

$$\begin{aligned} \text{Classe1} : f_1(\mathbf{x}; \Phi) &= 0,4[Beta(x_1; 1; 8) * Beta(x_2; 2; 8) * Beta(x_3; 3; 3)] \\ &+ 0,6[Beta(x_1; 1; 8) * Beta(x_2; 5; 5) * Beta(x_3; 3; 3)]. \end{aligned}$$

$$\begin{aligned} \text{Classe2} : f_2(\mathbf{x}; \Phi) &= 0,6[Beta(x_1; 1; 8) * Beta(x_2; 2; 8) * Beta(x_3; 3; 3)] \\ &+ 0,4[Beta(x_1; 5; 1) * Beta(x_2; 5; 5) * Beta(x_3; 3; 3)]. \end{aligned}$$

Cada variável da Estrutura 7 é provenientes de MFB, sendo o conjunto de dados tridimensional com $l = 2$:

$$\begin{aligned} \text{Classe1} : f_1(\mathbf{x}; \Phi) &= [0,4 * Beta(x_1; 1; 8) + 0,6 * Beta(x_1; 1; 8)] \\ &* [0,4 * Beta(x_2; 2; 8) + 0,6 * Beta(x_2; 8; 8)] \\ &* [0,4 * Beta(x_3; 3; 3) + 0,6 * Beta(x_3; 3; 3)]. \end{aligned}$$

$$\begin{aligned}
\text{Classe2} : f_2(\mathbf{x}; \Phi) &= [0,6 * \text{Beta}(x_1; 50; 10) + 0,4 * \text{Beta}(x_1; 50; 10)] \\
&* [0,6 * \text{Beta}(x_2; 50; 50) + 0,4 * \text{Beta}(x_2; 50; 50)] \\
&* [0,6 * \text{Beta}(x_3; 30; 30) + 0,4 * \text{Beta}(x_3; 30; 30)].
\end{aligned}$$

Cada variável da Estrutura 8 é provenientes de MFB, sendo o conjunto de dados tridimensional com $l = 2$:

$$\begin{aligned}
\text{Classe1} : f_1(\mathbf{x}; \Phi) &= [0,4 * \text{Beta}(x_1; 1; 8) + 0,6 * \text{Beta}(x_1; 1; 8)] \\
&* [0,4 * \text{Beta}(x_2; 2; 8) + 0,6 * \text{Beta}(x_2; 8; 8)] \\
&* [0,4 * \text{Beta}(x_3; 3; 3) + 0,6 * \text{Beta}(x_3; 3; 3)].
\end{aligned}$$

$$\begin{aligned}
\text{Classe2} : f_2(\mathbf{x}; \Phi) &= [0,6 * \text{Beta}(x_1; 5; 1) + 0,4 * \text{Beta}(x_1; 5; 1)] \\
&* [0,6 * \text{Beta}(x_2; 5; 5) + 0,4 * \text{Beta}(x_2; 5; 5)] \\
&* [0,6 * \text{Beta}(x_3; 3; 3) + 0,4 * \text{Beta}(x_3; 3; 3)].
\end{aligned}$$

Cada variável da Estrutura 9 é provenientes de MFB, sendo o conjunto de dados tridimensional com $l = 2$:

$$\begin{aligned}
\text{Classe1} : f_1(\mathbf{x}; \Phi) &= [0,4 * \text{Beta}(x_1; 1; 8) + 0,6 * \text{Beta}(x_1; 1; 8)] \\
&* [0,4 * \text{Beta}(x_2; 2; 8) + 0,6 * \text{Beta}(x_2; 8; 8)] \\
&* [0,4 * \text{Beta}(x_3; 3; 3) + 0,6 * \text{Beta}(x_3; 3; 3)].
\end{aligned}$$

$$\begin{aligned}
\text{Classe2} : f_2(\mathbf{x}; \Phi) &= [0,6 * \text{Beta}(x_1; 2; 9) + 0,4 * \text{Beta}(x_1; 5; 1)] \\
&* [0,6 * \text{Beta}(x_2; 3; 9) + 0,4 * \text{Beta}(x_2; 5; 5)] \\
&* [0,6 * \text{Beta}(x_3; 5; 5) + 0,4 * \text{Beta}(x_3; 3; 3)].
\end{aligned}$$

Referências Bibliográficas

- Andrews, J. L. & McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing*, **22**(5), 1021–1029.
- Basso, R. M. et al. (2009). *Misturas finitas de misturas de escala skew-normal*. Master's thesis, IMECC/UNICAMP. Dissertação de Mestrado, Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.
- Bouguila, N., Ziou, D. & Monga, E. (2006). Practical bayesian estimation of a finite beta mixture through gibbs sampling and its applications. *Statistics and Computing*, **16**(2), 215–225.
- Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, **16**(5), 1190–1208.
- Cabral, C. R. B., Lachos, V. H. & Prates, M. O. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis*, **56**(1), 126–142.
- Casella, G. (2002). *Statistical Inference*. Duxbury Advanced Series. Duxbury Thomson Learning.
- Coelho, C. F. (2013). *Misturas Finitas de Normais Assimétricas e de t Assimétricas Aplicadas em Análise Discriminante*. Dissertação de Mestrado, ICE/UFAM.
- Costa, M. G., Costa Filho, C. F., Sena, J. F., Salem, J. & de Lima, M. O. (2008). Automatic identification of mycobacterium tuberculosis with conventional light microscopy. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 382–385. IEEE.
- Domingos, P. & Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, **29**(2-3), 103–130.

- Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.
- Frery, A. C. & Perciano, T. (2013). *Introduction to Image Processing Using R: Learning by Examples*. Springer Science & Business Media.
- Gonzales, R. C. & Woods, R. E. (2002). *Digital Image Processing*. Prentice Hall.
- Hastie, T., Tibshirani, R. & Friedman, J. (2009). Unsupervised learning. In *The Elements of Statistical Learning*, pages 485–585. Springer.
- Ji, Y., Wu, C., Liu, P., Wang, J. & Coombes, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics*, **21**(9), 2118–2122.
- Johnson, R. A. & Wichern, D. W. (2012). *Applied multivariate statistical analysis*. Prentice Hall, 6th edition.
- Junior, A. K., Costa, M. G., Costa Filho, C. F., Fujimoto, L. B. & Salem, J. (2010). Evaluation of autofocus functions of conventional sputum smear microscopy for tuberculosis. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 3041–3044. IEEE.
- Kimura Junior, A. (2010). *Avaliação das métricas de autofocus para aplicação em imagens de baciloscopia de tuberculose obtidas utilizando microscopia de campo claro*. Dissertação de Mestrado, FT/UFAM.
- Kotz, S. & Lovelace, C. (1998). Introduction to process capability indices: Theory and practice. *Arnold, London*.
- Ma, Z. & Leijon, A. (2010). Human skin color detection in rgb space with bayesian estimation of beta mixture models. In *Signal Processing Conference, 2010 18th European*, pages 1204–1208. IEEE.
- Ma, Z. & Leijon, A. (2011). Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**(11), 2160–2173.
- McLachlan, G. J. & Peel, D. (2000). *Finite mixture models*. Wiley New York ; Chichester.
- Narayanan, A. (1992). A note on parameter estimation in the multivariate beta distribution. *Computers & Mathematics with Applications*, **24**(10), 11–17.

- Nguyen, T. M. & Wu, Q. J. (2015). A non-parametric bayesian model for bounded data. *Pattern Recognition*, **48**(6), 2084–2095.
- Pereira, J. R. G. et al. (2001). *Misturas Finitas de Densidades com Aplicações em Reconhecimento Estatístico de Padrões*. Tese de Doutorado, UNICAMP.
- Prates, M. O., Lachos, V. H. & Cabral, C. (2013). mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, **54**, 1–20.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.
- Theodoridis, S., Pikrakis, A., Koutroumbas, K. & Cavouras, D. (2010). *Introduction to pattern recognition: a matlab approach*. Academic Press/Elsevier.
- Titterton, D. M., Smith, A. M. F. & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, NewYork.
- Warnes, G. R., Bolker, B. & Lumley, T. (2015). *gtools: Various R Programming Tools*. R package version 3.5.0.
- Webb, G. I. & Ting, K. M. (2005). On the application of roc analysis to predict classification performance under varying class distributions. *Machine learning*, **58**(1), 25–32.