UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA CIÊNCIAS E LETRAS DE RIBEIRÃO
PRETO
DEPARTAMENTO DE FÍSICA E MATEMÁTICA

DIEGO MERIGUE DA CUNHA

CARACTERIZAÇÃO DE TECIDOS MAMÁRIOS ATRAVÉS
DE MODELOS ESTATÍSTICOS UTILIZANDO O
ESPALHAMENTO DE RAIOS-X

RIBEIRÃO PRETO

2006

DIEGO MERIGUE DA CUNHA

**CARACTERIZAÇÃO DE TECIDOS MAMÁRIOS ATRAVÉS
DE MODELOS ESTATÍSTICOS UTILIZANDO O
ESPALHAMENTO DE RAIOS-X**

Dissertação submetida ao Programa de Pós-Graduação em Física Aplicada à Medicina e Biologia da Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, da Universidade de São Paulo, como requisito para a obtenção do título de Mestre em Ciências.

Orientador: Prof. Dr. Martin Eduardo Poletti.

RIBEIRÃO PRETO

2006

AGRADECIMENTOS

A Deus, pelas oportunidades que me são concedidas;

À minha família, pelo apoio incondicional em todos os momentos;

À Alessandra Tomal, pelo companheirismo, pela amizade e por fazer parte da minha vida;

Ao Prof. Dr. Martin Eduardo Poletti, pela orientação deste trabalho;

Ao Prof. Dr. Carlos Júlio Tierra-Criollo, pelos esclarecimentos a respeito da aplicação de técnicas multivariadas de análise de dados;

Ao Prof. Dr. Alfredo Ribeiro-Silva, pelos esclarecimentos e discussões sobre as características histopatológicas das doenças de mama;

Ao Departamento de Patologia da Faculdade de Medicina de Ribeirão Preto, por fornecer os materiais analisados neste trabalho;

Aos professores do curso de pós-graduação em Física Aplicada à Medicina e Biologia, do Departamento de Física e Matemática da Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, que contribuíram para minha formação;

Aos amigos Paulo Petchevist, Herofen Zaias, Otávio Riani e Marina Piacenti, pela convivência ao longo do curso;

Aos amigos André Gonçalves, Denise Assafrão, Cleverson Filgueiras, Carlos Eduardo Galhardo e Mariana Brandão, pela amizade constante, mesmo à distância.

A CAPES, pela concessão da bolsa de pós-graduação;

A FAPESP, pelo apoio financeiro ao projeto Jovem Pesquisador “Caracterização Histopatológica de Tecidos Mamários Através da Radiação Secundária”, nº do processo: 02/00380-3.

SUMÁRIO

LISTA DE FIGURAS.....	v
LISTA DE TABELAS.....	vii
RESUMO.....	viii
ABSTRACT	ix
1 - INTRODUÇÃO	1
2 - FUNDAMENTOS TEÓRICOS	4
2.1- Interação da Radiação com a Matéria	4
2.1.1 – Espalhamento Elástico.....	4
2.1.2 – Espalhamento Inelástico	9
2.2 – Comparação entre Grupos de Amostras	11
2.2.1 - Comparação Univariada.....	12
2.2.2 – Comparação Multivariada	14
2.3 – Reconhecimento de Padrões Utilizando Análise Multivariada	15
2.3.1 - Análise de Componentes Principais.....	15
2.3.2 – Análise de Discriminante	18
2.4 – Critério de Classificação Histológica	21
2.5 – Verificação da Eficácia dos Modelos de Diagnóstico	22
3 - MATERIAIS E MÉTODOS.....	24
3.1 – Coleta das Amostras de Tecido Mamários	24
3.2 – Classificação Histológica dos Tecidos Mamários	24
3.3 - Obtenção dos Perfis de Espalhamento	25
3.3.1 – Preparação das Amostras.....	25
3.3.2 - Arranjo experimental	25
3.4 – Pré-processamento dos Dados Experimentais.....	27
3.4.1 – Suavização do Ruído	28
3.4.2 – Correção por contribuições espúrias	28
3.4.3 – Obtenção da Seção de Choque Diferencial de Espalhamento.....	30
3.5 – Modelos Estatísticos para Diagnóstico através dos Perfis de Espalhamento	32
3.5.1 – Modelo de Diagnóstico utilizando Variáveis Simplificadas	33
3.5.2 – Modelo de Diagnóstico utilizando Componentes Principais.....	35
3.5.3 – Modelo de Diagnóstico utilizando Análise de Discriminante.....	38

4 - RESULTADOS E DISCUSSÕES	41
4.1 – Obtenção dos Perfis de Espalhamento	41
4.1.1 - Amostras de Tecidos Mamários Analisadas	41
4.1.2 – Obtenção Experimental da Distribuição Angular de Fótons Espalhados.....	42
4.2 – Pré-Processamento dos Dados Experimentais.....	42
4.3 – Modelos Estatísticos para Diagnóstico através dos Perfis de Espalhamento	46
4.3.1 - Modelo de Diagnóstico utilizando Variáveis Simplificadas	46
4.3.2 - Modelo de Diagnóstico utilizando Componentes Principais	53
4.3.3 - Modelo de Diagnóstico utilizando Análise de Discriminante.....	64
5 - CONCLUSÕES E PERSPECTIVAS	75
5.1 - Conclusões	75
5.1.1 - Sobre o método de obtenção das seções de choque $d\sigma/d\Omega$ das amostras de tecido	75
5.1.2 - Sobre os modelos de diagnóstico desenvolvidos	76
5.1.3 – Conclusões Finais	77
5.2 – Perspectivas	78
6 - REFERÊNCIAS BIBLIOGRÁFICAS	79

LISTA DE FIGURAS

Figura 2.1 – Fator de forma e função de espalhamento incoerente da molécula de H ₂ O. $F_{mol}(x)$ foi obtido utilizando distribuição de densidade eletrônica. $F_{mol}^{MAI}(x)$ e $S_{mol}^{MAI}(x)$ foram obtidos através do modelo atômico independente (MORIN, 1982).....	7
Figura 2.2 – Seção de choque diferencial molecular em 17,44 keV. As curvas foram obtidas utilizando: (i) $F_{mol}(x)$, (ii) $F_{mol}^{MAI}(x)$ e (iii) $S_{mol}^{MAI}(x)$	8
Figura 2.3 – Seção de choque diferencial elástica para água líquida a 20°C (MORIN,1982), molécula de água (MAI), Tecido mamário normal (50% adiposo e 50% glandular) (POLETTI, GONÇALVES E MAZZARO, 2002a) e seção de choque diferencial inelástica para molécula de água (MAI) obtida para 17,44 keV.	8
Figura 2.4 – Componentes principais CP ₁ e CP ₂ obtidos para o conjunto de dados definidos pelas variáveis X ₁ e X ₂ (JOHNSON E WICHERN, 2002).	17
Figura 2.5 – Função discriminante Z que otimiza a separação entre duas populações bivariadas (JOHNSON E WICHERN, 2002).....	20
Figura 3.1 – Porta-amostra utilizado.	25
Figura 3.2 – Arranjo experimental para as medidas de difração de raios-X.	26
Figura 3.3 – Área do tecido irradiada.....	26
Figura 3.4 - (a) Representação esquemática do arranjo experimental utilizado para medida das amostras de tecido. (b) Contribuição de cada parte do arranjo ao espalhamento.....	29
Figura 3.5 - Arranjo experimental utilizado para medida do espalhamento (a) do ar, (b) do porta-amostra e (c) do filme de PVC.....	30
Figura 3.6 – Variáveis selecionadas a partir das curvas de espalhamento.	33
Figura 4.1 – Quantidade de amostras analisadas. TN = tecidos normais, FA = fibroadenomas, CDIS = carcinoma ductal “in situ”, CDI, CDII, CDIII = carcinomas ductais de grau I, II e III respectivamente e CM = carcinoma mucinoso.	41
Figura 4.2 – Distribuição angular dos fótons espalhados. (a) Água, (b) Tecido Adiposo Normal, (c) Neoplasia benigna (fibroadenoma) e (d) maligna (carcinoma ductal grau III).	42
Figura 4.3 – Seção de choque diferencial de espalhamento para água líquida, comparada com o resultado obtido através do fator de forma tabulado por Morin (MORIN, 1982).	43
Figura 4.4 - Valores médios das seções de choque diferencial de espalhamento dos tipos de tecidos analisados.	44
Figura 4.5 - Seção de Choque para tecidos normais com diferentes componentes histológicos.	44
Figura 4.6 – Perfil de espalhamento de fibroadenomas com presença de células adiposas.	45
Figura 4.7 - Perfil de espalhamento médio dos diferentes tipos de carcinomas analisados.	46
Figura 4.8 – <i>Box-plot</i> das variáveis P ₁ , P ₂ , I ₁ , I ₂ , LMA, e I ₂ /I ₁	47

Figura 4.9 - Intervalos de Confiança construídos para $\alpha=0,05$.	50
Figura 4.10 – Gráfico de $LMA \times I_1$, mostrando as coordenadas de cada grupo.	51
Figura 4.11 – Teste de Kruskal-Wallis comparando TN, FA e C em cada ângulo de espalhamento.	54
Figura 4.12 – (a) Porcentagem de variância p_k correspondente a cada autovalor. (b) Porcentagem de variância acumulada.	55
Figura 4.13 – Componentes Principais obtidas da matriz S total dos grupos TN, FA e C.	56
Figura 4.14 – Posição das amostras no espaço $CP_1 \times CP_4$.	58
Figura 4.15 - (a) Porcentagem de variância p_k correspondente à cada autovalor. (b) Porcentagem de variância acumulada.	60
Figura 4.16 - Componentes Principais obtidas pela matriz R total dos grupos TN, FA e C.	60
Figura 4.17 - Posição das amostras no espaço $CP_1 \times CP_5$.	62
Figura 4.18 – Comparação múltipla entre os grupos nas funções discriminantes Z_1 e Z_2 .	66
Figura 4.19 – Projeções as amostras no espaço $Z_1 \times Z_2$.	67
Figura 4.20 – Comparação múltipla entre os três grupos.	71
Figura 4.21 – Coordenadas das amostras no espaço $Z_1 \times Z_2$.	72

LISTA DE TABELAS

Tabela 2.1 – Resultados possíveis num teste de diagnóstico	22
Tabela 4.1 – Teste de Kruskal-Wallis para as variáveis P_1 , P_2 , I_1 , I_2 , LMA, e I_2/I_1	48
Tabela 4.2 – Classificação das amostras utilizando LMA e I_1	52
Tabela 4.3 – Processo de validação cruzada das variáveis LMA e I_1	52
Tabela 4.4 – Validação e reprodutibilidade do teste de diagnóstico	53
Tabela 4.5 – Teste de Kruskal-Wallis para as primeiras 8 componentes principais da matriz S	57
Tabela 4.6 – Classificação das amostras utilizando CP_1 e CP_4	58
Tabela 4.7 – Validação cruzada do método classificatório.	59
Tabela 4.8 – Validação e reprodutibilidade do modelo de diagnóstico.....	59
Tabela 4.9 - Teste de Kruskal-Wallis para as primeiras 8 componentes principais da matriz R	61
Tabela 4.10 – Classificação das amostras utilizando CP_1 e CP_5	62
Tabela 4.11 – Validação cruzada do método classificatório.	63
Tabela 4.12 – Validação e reprodutibilidade do modelo de diagnóstico.....	63
Tabela 4.13 – Variáveis utilizadas no cálculo de Z_1 e Z_2 no método direto.	64
Tabela 4.14 – Autovalores da matriz $W^{-1}B$	64
Tabela 4.15 – Elementos da matriz de coeficientes a	65
Tabela 4.16 – Teste de Kruskal-Wallis entre TN, FA e C nas variáveis Z_1 e Z_2	66
Tabela 4.17 – Teste λ de Wilks entre os grupos na distribuição bivariada formada por Z_1 e Z_2	66
Tabela 4.18 – Classificação das amostras utilizando Z_1 e Z_2	67
Tabela 4.19 – Validação cruzada do método classificatório.	68
Tabela 4.20 – Validação e reprodutibilidade do modelo de diagnóstico.....	69
Tabela 4.21 – Variáveis selecionadas para o cálculo das funções discriminantes.	70
Tabela 4.22 – Autovalores da matriz $W^{-1}B$	70
Tabela 4.23 – Coeficientes da matriz a	70
Tabela 4.24 – Teste de Kruskal-Wallis entre TN, FA e C.....	71
Tabela 4.25 – Teste λ de Wilks para a distribuição bivariada formada por Z_1 e Z_2	71
Tabela 4.26 – Classificação das amostras utilizando Z_1 e Z_2	72
Tabela 4.27 – Validação cruzada do método classificatório.	73
Tabela 4.28 – Validação e reprodutibilidade do modelo de diagnóstico.....	74

RESUMO

CUNHA, D. M. *Caracterização de Tecidos Mamários através de Modelos Estatísticos Utilizando Espalhamento de Raios-X*. [Dissertação]. Ribeirão Preto: Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, Universidade de São Paulo; 2006. 90 p.

Em um exame mamográfico, quando os fótons de raios-X incidem sobre a mama, uma parte destes fótons é transmitida sem ser desviada da trajetória inicial (radiação transmitida primária), permitindo a formação da imagem mamográfica, e outra é dispersa de sua trajetória inicial pelo tecido (radiação espalhada), atuando de forma deletéria na imagem mamográfica. Entretanto, recentes investigações têm demonstrado que a radiação espalhada pode ser útil na caracterização de tecidos. O objetivo deste trabalho é desenvolver um modelo de diagnóstico de alterações no tecido mamário utilizando as informações presentes na distribuição angular da radiação espalhada (perfil de espalhamento). Os perfis de espalhamento de 40 amostras de tecidos mamários foram obtidos utilizando um difractômetro comercial SIEMENS D5005, operando em modo reflexão na energia de 8,04 keV e variando o detector da posição angular de 5° a 150°, correspondendo a um intervalo de x de $0,03 \text{ \AA}^{-1}$ a $0,62 \text{ \AA}^{-1}$. As amostras de tecido foram previamente classificadas histopatologicamente como tecidos normais, fibroadenomas (neoplasias benignas) e diferentes tipos de carcinomas (neoplasias malignas). Neste trabalho, três modelos de diagnóstico baseados na análise estatística dos perfis de espalhamento foram desenvolvidos. O primeiro, analisa seis parâmetros extraídos dos perfis de espalhamento, já o segundo e o terceiro utilizam análise multivariada (análise de componentes principais e análise de discriminante, respectivamente) para reconhecimento de padrões. Para cada modelo, valores de sensibilidade, especificidade e índice de concordância entre o diagnóstico baseado no modelo utilizado e o diagnóstico histopatológico foram obtidos. Dentre os modelos desenvolvidos, aquele que utiliza análise de discriminante proporciona o melhor diagnóstico das alterações encontradas no tecido, permitindo diferenciar tecidos normais e neoplasias benignas e malignas. Com base nos resultados obtidos conclui-se que modelos baseados na análise estatística dos perfis de espalhamento permitem classificar histologicamente tecidos mamários.

Palavras-chaves: Câncer de mama; Espalhamento de raios-x; Análise estatística multivariada; Materiais amorfos.

ABSTRACT

CUNHA, D. M. *Breast Tissue Characterization by Statistical Models Using X-Ray Scattering*. [Dissertation]. Ribeirão Preto: Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, Universidade de São Paulo; 2006. 91 p.

In mammography, when x-ray photons reach the breast, a fraction of these photons is transmitted without interaction with any tissues (primary transmitted radiation), allowing the formation of the mammographic image, and another fraction of them is deviated by the tissue from its initial trajectories (scattered radiation), reducing the image contrast. However, recent investigations have demonstrated that scattered radiation can be a useful diagnostic tool. The purpose of this work is to develop a diagnostic model for breast tissue characterization using the angular distribution of the scattered radiation (scattering profile). The scattering profiles of 40 breast tissue samples were obtained in a SIEMENS D5005 diffractometer, operating in reflection mode at 8,04keV, and varying the angular position of the detector from 5° to 150°, corresponding to an x interval from 0,03 Å⁻¹ to 0,62 Å⁻¹. All tissue samples were previously classified histopathologically as normal tissues, fibroadenomas (benign alteration) and several types of carcinomas (malignant alteration). Three models of diagnostic based on the statistical analysis of the scattering profiles were developed. The first one was constructed using six parameters extracted from the scattering profiles and the second and third models used the whole information from the scattering profiles. The latter two used multivariate analysis (principal component analysis and discriminant analysis, respectively) for pattern recognition. For each model, values of sensitivity, specificity and rate of agreement between the model diagnostic and histopathological results were obtained. Among the developed models, the discriminant analysis provides the best diagnostic of the lesions present in the tissues (normal tissues, benign and malignant alterations). From the results, it is possible to conclude that models based on the statistical analysis of the scattering profiles allow the histological classification of breast tissues.

Keywords: Breast Cancer; X-ray scattering; Multivariate statistical analysis; Amorphous materials.

CAPÍTULO 1

INTRODUÇÃO

O câncer de mama é o segundo tipo de câncer mais freqüente em todo o mundo, e o primeiro entre as mulheres (INSTITUTO NACIONAL DE CÂNCER, 2005). Embora a epidemiologia da doença ainda não seja totalmente conhecida, sabe-se que certos fatores tendem a aumentar o risco do seu surgimento, como por exemplo, idade e histórico familiar (MCPHERSON, STEEL e DIXON, 2000).

A detecção precoce do câncer de mama é a forma mais eficaz de diminuir a taxa de mortalidade e aumentar a sobrevida dos pacientes (INSTITUTO NACIONAL DE CÂNCER, 2005), sendo a mamografia a técnica mais utilizada para este fim (SEIDMAN E MUSHINSKI, 1983; TABAR *ET AL*, 1985). Entretanto, apesar de ser eficiente em evidenciar a existência de alterações na mama, esta técnica não permite o diagnóstico dessas alterações, sendo incapaz de diferenciar precisamente entre alterações benignas e malignas. A função de classificar estas alterações é desempenhada por médicos histopatologistas através da análise histológica de pequenas amostras de tecido (biópsias) (TABAR *ET AL*, 1985).

Num exame mamográfico, o feixe de raios-X ao incidir sobre a mama vai sendo atenuado à medida que atravessa o tecido, e a fração da radiação que chega ao detector (filme radiográfico) forma a imagem mamográfica. Entretanto, alguns fótons ao interagirem com o tecido, sofrem dispersão (processo de espalhamento). Esta radiação espalhada incidente no filme radiográfico sempre foi considerada um problema a ser minimizado, por ser a responsável pela diminuição da resolução e contraste da imagem. Muitos esforços têm sido feitos na tentativa de reduzir a quantidade de radiação espalhada que chega ao filme (BARNES, 1991).

Entretanto, o conceito de considerar o espalhamento de raios-X como uma desvantagem vem mudando ultimamente, devido a dois fatores importantes (HARDING KOSANETSKY E NEITZEL, 1987; SPELLER E HORROCKS, 1991): (i) os fótons espalhados carregam informação sobre a estrutura presente nos tecidos do paciente (informação que não é considerada na radiografia convencional) e (ii) na maioria das energias

usadas nas aplicações médicas, a interação mais provável é o espalhamento (elástico + inelástico). Estes fatos levaram a se pesquisar as propriedades de espalhamento dos diferentes tecidos, com a obtenção de medidas precisas dos perfis de espalhamento dos diferentes tecidos humanos, e em particular os perfis de tecidos mamários (POLETTI *ET AL*, 2002).

Recentes investigações vêm apontando para a possibilidade de se caracterizar um tecido utilizando informações sobre o espalhamento elástico de raios-X (KIDANE *ET AL*, 1999; SPELLER, 1999; POLETTI, 2001; POLETTI *ET AL*, 2002; POLETTI, GONÇALVES E MAZZARO, 2002a; POLETTI, GONÇALVES E MAZZARO, 2002b; OLIVEIRA, 2006). Nestes trabalhos, fica evidente a diferença entre o perfil de espalhamento de tecidos normais e alterados, mas nota-se que há pouca diferença entre o perfil de alterações benignas e malignas.

A proposta deste trabalho é desenvolver um método de análise que permita a classificação histológica dos tecidos mamários, através de seus perfis de espalhamento. O objetivo é evidenciar diferenças não apenas entre tecidos normais e alterados, mas possibilitar a identificação da alteração como benigna ou maligna.

Para isso, obteve-se experimentalmente o perfil de espalhamento de amostras de tecido mamário de diferentes classificações histológicas, e dentre as várias formas de se buscar a identificação de padrões num conjunto de dados (JAIN *ET AL*, 2000), optou-se pela abordagem estatística, aplicando a estes perfis procedimentos estatísticos para reconhecimento de padrões, e testes de hipóteses para comparar grupos de amostras. As técnicas para reconhecimento de padrões basearam-se em métodos multivariados, em particular a análise de componentes principais e a análise de discriminante.

O trabalho foi organizado da seguinte forma:

O capítulo 2 faz uma breve descrição dos conceitos teóricos em que este trabalho está baseado. São descritos processos de interação da radiação com a matéria, em particular as características do processo de espalhamento elástico e inelástico em átomos e moléculas. Descrevem-se também as formas de comparar grupos de amostras através de testes estatísticos, bem como as técnicas aplicadas a fim de reconhecer padrões de comportamento num conjunto de dados. Além disso, descreve-se também o critério de classificação histológica adotado, e alguns parâmetros utilizados para testar a eficiência e confiabilidade de um teste de diagnóstico.

No capítulo 3 são apresentados os tipos de tecidos mamários analisados, o arranjo experimental utilizado para medida dos perfis de espalhamento, o processo de tratamento dos dados experimentais, a metodologia de construção dos modelos de diagnóstico utilizando

métodos estatísticos de reconhecimento de padrões, e por fim o processo de análise e validação dos modelos desenvolvidos.

O capítulo 4 apresenta os resultados obtidos. Primeiramente são mostrados os perfis de espalhamento obtidos experimentalmente para as amostras de tecidos mamários analisadas. Em seguida, os resultados obtidos ao longo da construção dos modelos de diagnóstico, e a avaliação da eficiência destes modelos na caracterização histológica das amostras de tecido.

Por fim, no capítulo 5, são apresentadas as conclusões sobre os métodos utilizados e resultados obtidos, e as perspectivas para trabalhos futuros.

CAPÍTULO 2

FUNDAMENTOS TEÓRICOS

2.1- Interação da Radiação com a Matéria

Os principais fenômenos de interação da radiação com a matéria, considerando a faixa de energia utilizada em radiodiagnóstico ($E < 100 \text{ keV}$), são o efeito fotoelétrico e o espalhamento elástico e inelástico. Para fótons com energia utilizada em mamografia (17,44 keV) numa amostra de água em particular, 80,7% do total de interações ocorrem por efeito fotoelétrico, 8,5% para o espalhamento elástico e 10,8% para o espalhamento inelástico (JOHNS E CUNNINGHAM, 1983). Tecidos biológicos apresentam valores percentuais semelhantes devido às altas percentagens de água em suas composições.

Neste trabalho está sendo analisada a distribuição angular dos fótons espalhados (perfil de espalhamento) por tecidos mamários, que é governada pela seção de choque do conjunto de moléculas que compõe o tecido. Os tópicos abaixo apresentam a descrição do modelo teórico para a seção de choque diferencial para o espalhamento de fótons em um sistema constituído por um conjunto de moléculas, que é o modelo teórico que mais se aproxima do espalhamento de fótons em tecidos biológicos.

2.1.1 – Espalhamento Elástico

Quando um fóton interage com a matéria se desviando da sua trajetória inicial de modo que ele não sofra variação na sua energia, temos o chamado espalhamento elástico ou coerente. O campo elétrico da onda eletromagnética, com comprimento de onda λ , associado a esse fóton dá origem a uma vibração dos elétrons presentes na matéria. Devido a essa aceleração estes elétrons emitem uma radiação de mesmo comprimento de onda λ que a radiação incidente (JOHNS E CUNNINGHAM, 1983).

Se o espalhamento elástico ocorre devido somente a um elétron livre, ele também recebe o nome de espalhamento Thomson, em homenagem ao primeiro cientista que obteve a seção de choque diferencial de espalhamento para um elétron (JOHNS E CUNNINGHAM, 1983). A seção de choque mede a intensidade da radiação espalhada, ou seja, o número de fótons por unidade de área em função da direção de espalhamento. Thomson considerando um feixe de fótons não polarizado e apenas suposições de física clássica obteve a seguinte expressão

$$\left(\frac{d\sigma}{d\Omega}\right)_{Th} = \frac{r_0^2}{2}(1 + \cos^2 \theta) \quad (2.1)$$

onde $r_0 = \frac{1}{4\pi\epsilon_0} \frac{e^2}{m_0 c^2} = 2.82 \times 10^{-15} m$ é o raio clássico do elétron e θ é o ângulo de espalhamento.

O espalhamento elástico de fótons por um átomo livre, também chamado espalhamento Rayleigh em homenagem ao cientista que o discutiu para a luz visível, Lord Rayleigh (1881), ocorre quando os fótons incidentes têm sua energia conservada e sofrem variação de momento após o espalhamento pelos elétrons ligados ao átomo. As ondas espalhadas por cada elétron sofrem interferência construtiva devido à diferença de caminho ótico percorrido no átomo. A seção de choque diferencial pode ser expressa na forma

$$\left(\frac{d\sigma}{d\Omega}\right)_{Ray} = [F(x,Z)]^2 \left(\frac{d\sigma}{d\Omega}\right)_{Th} \quad (2.2)$$

onde $x = \lambda^{-1} \sin(\theta/2)$ é proporcional ao momento transferido q na interação do fóton com o átomo e a relação entre x e q é dada por $x = 20.607744 \hbar q$, onde $\hbar = 1,054 \times 10^{-34} J.s$. $F(x,Z)$ é chamado fator de forma atômico, onde Z é o número atômico..

A função $F(x,Z)$ é a transformada de Fourier da distribuição de cargas do átomo (HUBBELL *ET AL*, 1975). O fator de forma total do átomo pode ser decomposto na soma dos fatores de forma individuais de cada camada eletrônica do átomo (HUBBELL *ET AL*, 1975).

Para determinar o espalhamento por uma molécula livre podem ser utilizadas três abordagens: A aproximação mais geral utiliza a distribuição de densidade eletrônica dentro da molécula (BLUM, 1971 *apud* NARTEN E LEVY, 1971). Outra aproximação possível requer

o conhecimento das distâncias atômicas dentro da molécula (DEBYE, 1915 *apud* NARTEN E LEVY, 1971). Uma terceira aproximação, mais simples, supõe que o espalhamento por uma molécula livre é composto pelas contribuições independentes de cada átomo que a compõe. Essa última aproximação é também chamada Modelo Atômico Independente (MAI) (JAMES, 1962).

No Modelo Atômico Independente o fator de forma molecular será determinado pela soma dos fatores de forma de cada átomo independentemente, e pode ser obtido de duas maneiras:

(i) Se o número de átomos for conhecido, $F_{mol}^2(x)$ assume a forma:

$$F_{mol}^2(x) = \sum n_i F_i^2(x) \quad (2.3)$$

onde F_i é o fator de forma de cada átomo e n_i o número de átomos de cada elemento.

(ii) Se o número de átomos for desconhecido, $F_{mol}^2(x)$ pode ser obtida pela fração de massa w_i e pela massa atômica A_i do átomo i :

$$F_{mol}^2(x)/M = \sum \left(\frac{w_i}{A_i} \right) F_i^2(x) \quad (2.4)$$

onde M é o peso molecular (CHAN E DOI, 1983).

A seção de choque diferencial elástica para uma molécula livre é da forma:

$$\left(\frac{d\sigma}{d\Omega} \right)_{elas}^{mol} = F_{mol}^2(x) \left(\frac{d\sigma}{d\Omega} \right)_{Th} \quad (2.5)$$

A figura 2.1 mostra o comportamento da função $F_{mol}(x)$ para molécula de água utilizando a distribuição de densidade de carga eletrônica (BLUM *apud* NARTEN E LEVY, 1971) e o modelo atômico independente (HUBBELL *ET AL*, 1975). A seção de choque diferencial elástica para molécula de água está representada na figura 2.2, calculada a partir das duas aproximações de $F_{mol}(x)$. Pode-se notar que as duas aproximações convergem para o mesmo valor quando $x > 0,4 \text{ \AA}^{-1}$.

A descrição do espalhamento por um conjunto de moléculas também pode ser obtida através da aproximação de fator de forma. A expressão para a seção de choque diferencial elástica por molécula é da forma:

$$\left(\frac{d\sigma}{d\Omega}\right)_{elas} = F_{eq}^2(x) \left(\frac{d\sigma}{d\Omega}\right)_{Th} \quad (2.6)$$

onde $F_{eq}(x)$ é o fator de forma equivalente que leva em conta o grau de ordem da estrutura atômica (ou molecular).

A figura 2.3 mostra a seção de choque elástica para água líquida a 20°C obtida a partir do fator de forma equivalente tabulado por Morin (MORIN, 1982). Observa-se que a função de interferência provoca oscilações na seção de choque do conjunto de moléculas. Tais oscilações ocorrem em torno da seção de choque da molécula livre calculada a partir do MAI, observando-se também a existência de interferência destrutiva para valores baixos de x . Este efeito é bem conhecido em estudos de difração por líquidos ou materiais amorfos (JAMES, 1962). Para valores de x maiores as duas curvas convergem para os mesmos valores.

Figura 2.1 – Fator de forma e função de espalhamento incoerente da molécula de H₂O. $F_{mol}(x)$ foi obtido utilizando distribuição de densidade eletrônica. $F_{mol}^{MAI}(x)$ e $S_{mol}^{MAI}(x)$ foram obtidos através do modelo atômico independente (MORIN, 1982).

Figura 2.2 – Seção de choque diferencial molecular em 17,44 keV. As curvas foram obtidas utilizando: (i) $F_{\text{mol}}(x)$, (ii) $F_{\text{mol}}^{\text{MAI}}(x)$ e (iii) $S_{\text{mol}}^{\text{MAI}}(x)$.

Figura 2.3 – Seção de choque diferencial elástica para água líquida a 20°C (MORIN, 1982), molécula de água (MAI), Tecido mamário normal (50% adiposo e 50% glandular) (POLETTI, GONÇALVES E MAZZARO, 2002a) e seção de choque diferencial inelástica para molécula de água (MAI) obtida para 17,44 keV.

2.1.2 – Espalhamento Inelástico

Quando os elétrons do meio absorvem, em forma de energia cinética, parte da energia do fóton incidente, tem-se o chamado espalhamento inelástico ou incoerente. O processo de espalhamento inelástico foi tratado inicialmente por Compton (JOHNS E CUNNINGHAN, 1983) que considerou a radiação como pacotes de onda quantizados com energia $h\nu$. Por isso, esse processo também recebe o nome de espalhamento Compton.

Se um fóton é espalhado por um elétron livre de forma que não ocorra conservação na energia deste fóton, a seção de choque de espalhamento diferencial é dada por:

$$\left(\frac{d\sigma}{d\Omega}\right)_{KN} = \left(\frac{d\sigma}{d\Omega}\right)_{Th} F_{KN} = \left(\frac{d\sigma}{d\Omega}\right)_{Th} [1 + \alpha(1 - \cos\theta)]^{-2} \left\{ 1 + \frac{\alpha^2(1 - \cos\theta)^2}{[1 + \alpha(1 - \cos\theta)][1 + \cos^2\theta]} \right\} \quad (2.7)$$

com $\alpha = \frac{h\nu}{m_0c^2}$.

Essa expressão foi obtida por métodos de eletrodinâmica quântica por Klein e Nishina em 1928 (JOHNS E CUNNINGHAN, 1983).

O espalhamento inelástico num átomo (ou Compton atômico) pode ser descrito como uma aproximação análoga à realizada para o espalhamento Rayleigh. Definindo o fator $S(x,Z)$, também chamado de função do espalhamento inelástico, a seção de choque diferencial do efeito Compton atômico pode ser escrita:

$$\left(\frac{d\sigma}{d\Omega}\right)_{Comp} = S(x,Z) \left(\frac{d\sigma}{d\Omega}\right)_{KN} \quad (2.8)$$

onde $\left(\frac{d\sigma}{d\Omega}\right)_{KN}$ é a seção de choque diferencial obtida para o elétron livre (eq. 2.7). O fator $S(x,Z)$ é análogo ao fator de forma do espalhamento Rayleigh, e é usualmente calculado por tratamento não-relativístico.

No espalhamento inelástico de uma molécula, assim como para um átomo, cada elétron contribui de forma independente para o espalhamento (JOHNS E YAFFE, 1983). A seção de choque pode então ser dada como

$$\left(\frac{d\sigma}{d\Omega}\right)_{inelas}^{mol} = S_{mol}(x) \left(\frac{d\sigma}{d\Omega}\right)_{KN} \quad (2.9)$$

onde $S_{mol}(x)$ é a função de espalhamento incoerente molecular, definida como $S_{mol}(x) = \sum n_i S_i(x)$, se o número de átomos for conhecido. $S_i(x)$ é a função de espalhamento incoerente de cada átomo e n_i o número de cada átomo na molécula (CHAN E DOI, 1983). Os valores de $S_i(x)$ estão tabulados em Hubbell *et al* (HUBBELL ET AL, 1975) corrigidos pelas energias de ligação dos elétrons.

Se o número de átomos for desconhecido, $S_{mol}(x)$ também pode ser obtida a partir da fração de massa w_i e da massa atômica A_i do átomo i , através da expressão

$$S_{mol}(x)/M = \sum \left(\frac{w_i}{A_i}\right) S_i(x), \text{ onde } M \text{ é o peso molecular.}$$

A função de espalhamento incoerente molecular $S_{mol}(x)$ está representada na figura 2.1, e a seção de choque diferencial inelástica molecular na figura 2.2, tendo sido obtidas utilizando o modelo atômico independente para água.

Uma vez que no espalhamento inelástico os comprimentos de onda dos fótons espalhados são diferentes entre si, não existe uma relação de interferência entre as várias moléculas que compõe o conjunto. Pode-se supor que cada átomo contribui de forma independente para o espalhamento (Modelo Atômico Independente). Assim, a seção de choque do conjunto de moléculas, por unidade de molécula, será determinada pela soma das seções de choque de cada átomo do sistema, da mesma forma como realizado anteriormente para a molécula livre:

$$\left(\frac{d\sigma}{d\Omega}\right)_{inelas} = S_{mol}(x) \left(\frac{d\sigma}{d\Omega}\right)_{KN} \quad (2.10)$$

onde $S_{mol}(x)$ é a função de espalhamento incoerente molecular.

A partir da figura 2.3 vê-se que o comportamento da seção de choque, e conseqüentemente do perfil de espalhamento, é uma função do parâmetro x . Para valores pequenos de x a seção de choque apresenta uma dependência da distribuição espacial dos átomos que compõe as moléculas do conjunto, e para valores maiores de x pode ser descrita

pelo Modelo Atômico Independente, ou seja, os átomos que compõe o material espalham os fótons incidentes de forma independente uns dos outros.

No intervalo de x considerado neste trabalho espera-se observar padrões de interferência na região de baixo x nos perfis de espalhamento elástico dos tecidos, como o representado na figura 2.3 para tecidos mamários normais, constituídos de 50% de tecido adiposo e 50% de glandular. Uma vez que estes padrões dependem da estrutura atômica e molecular que compõe o tecido, espera-se encontrar diferenças nos perfis de espalhamento de tecidos mamários normais e alterados, e utilizar estas diferenças como método de diagnóstico.

2.2 – Comparação entre Grupos de Amostras

Neste trabalho, os perfis de espalhamento obtidos experimentalmente foram divididos em grupos, segundo sua classificação histológica (tecido normal, neoplasias benigna e maligna). Para identificar as diferenças existentes em cada grupo, os perfis foram comparados utilizando métodos de inferência estatística.

A estatística inferencial permite obter conclusões a respeito dos dados a partir da formulação de hipóteses. Uma hipótese é uma suposição a ser verificada com base nos dados observados, podendo ser aceita ou rejeitada. A veracidade de uma hipótese é verificada através da aplicação de testes estatísticos, que a determinam probabilisticamente. Esta probabilidade é o nível de significância do teste, definido previamente, e diz respeito à confiabilidade do resultado. Por exemplo, um nível de significância de 0,05 diz que há 95% de chance de a hipótese formulada para o teste estar correta (SPIEGEL, 2004).

Portanto, quando se comparam dois ou mais grupos de indivíduos, pode-se formular a hipótese de que os grupos são iguais (hipótese nula) ou diferentes (hipótese alternativa). O resultado do teste aplicado e o nível de significância adotado permitem tomar a decisão sobre aceitar ou rejeitar a hipótese nula.

O tipo de teste aplicado depende do número de grupos comparados e do comportamento da variável a ser testada. Os tópicos a seguir abordam brevemente os testes aplicados neste trabalho. A seção 2.2.1 mostra os testes univariados utilizados neste trabalho, ou seja, testes aplicados sobre uma única variável. A seção 2.2.2 mostra um teste multivariado utilizado para comparação entre grupos.

2.2.1 - Comparação Univariada

Teste U de Mann-Whitney

O teste de Wilcoxon-Mann-Whitney foi desenvolvido inicialmente por F. Wilcoxon em 1945 com o objetivo de comparar tendências centrais de dois grupos independentes com mesmo número de elementos. Em 1947 EH. B. Mann e D. R. Whitney generalizaram o método para grupos de tamanhos diferentes (CALLEGARI-JACQUES, 2003). O teste baseia-se em procedimentos de soma de postos (*ranks*) para determinar se dois grupos são idênticos, sem a restrição de que tenham distribuições de probabilidades conhecidas, e é análogo ao teste- t de Student para dois grupos independentes que seguem uma distribuição normal.

Seja $G_1 = \{X_1, X_2, \dots, X_{n_1}\}$ e $G_2 = \{Y_1, Y_2, \dots, Y_{n_2}\}$ dois grupos independentes, distribuídos segundo as funções de distribuição contínuas F_1 e F_2 , respectivamente. O objetivo do teste então é verificar se $F_1(x) = F_2(x)$, $\forall x \in \mathfrak{R}$ (hipótese nula H_0), ou se $F_1(x) > F_2(x)$, $F_1(x) < F_2(x)$ ou $F_1(x) \neq F_2(x)$ (hipóteses alternativas).

Combinando os dois grupos, obtém-se um único conjunto $G_1 G_2 = \{X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n\}$ de tamanho $n = n_1 + n_2$. Ordenam-se os elementos de $G_1 G_2$ do menor valor (posto 1) até o maior (posto n), denotando por $R(X_i)$ o posto ocupado por cada valor observado X_i , e define-se a soma:

$$T_{G_1} = \sum_{i=1}^{n_1} R(X_i) \quad (2.11)$$

A soma T_{G_1} é a estatística do teste, e será proporcional ao tamanho do grupo n_1 .

Entretanto, a estatística de Mann-Whitney é colocada freqüentemente em termos do parâmetro $U = U_{G_1}$, em que U_{G_1} pode ser definido como sendo o número de pares de elementos (X_i, Y_j) tais que $X_i > Y_j$, ou seja, o número de X_i 's maiores que Y_j 's (PRATT E GIBBONS, 1981). Este parâmetro fornece resultados equivalentes à U_{G_1} (ROHATGI, 1984).

$$U_{G_1} = \sum_{i=1}^m (\text{número de } X_i \text{'s} > Y_j \text{'s}) \quad (2.12)$$

Para calcular o valor- p do teste, ou seja, encontrar a região crítica para um dado α , deve-se observar a distribuição de probabilidade cumulativa para T_{G_1} para diferentes valores

de n_1 e n_2 . Essa distribuição de probabilidade cumulativa encontra-se tabulada em ROHATGI (ROHATGI, 1984).

Se num mesmo grupo, dois ou mais indivíduos apresentam o mesmo valor observado, então se diz que houve empate (*tie*). Embora para uma distribuição contínua teoricamente não possa haver empates, na prática eles acabam ocorrendo. Neste caso, a atribuição de postos baseados na magnitude dos valores observados não pode ser aplicada da forma convencional a estes indivíduos. Gibbons (GIBBONS, 1971) apresenta alguns métodos para se tratar grupos com empates. O mais freqüentemente utilizado é o método do posto-médio, em que aos indivíduos cujo valor observado se repete, atribui-se um valor para o posto ocupado por eles igual ao valor médio do posto que aquele valor observado teria caso não houvesse empates. Dessa forma, indivíduos que apresentam empates recebem postos iguais.

Análise de Variância Não-Paramétrica: Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é uma generalização do teste de Wilcoxon-Mann-Whitney, e é utilizado para se compararem dois ou mais grupos quanto à tendência central dos dados. É o análogo não paramétrico da análise de variância, ANOVA (*Analysis of Variance*), a ser utilizado quando a suposição de normalidade da distribuição dos dados não é satisfeita.

A estatística do teste baseia-se no parâmetro H , dado por

$$H = \frac{12}{n(n+1)} \sum_{i=1}^k n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2} \right)^2 \quad (2.13)$$

Onde $R_i = \sum_{j=1}^{n_i} R_{ij}$, $i = 1, 2, \dots, k$ é a soma dos postos ocupados pelos elementos do i -ésimo grupo no conjunto formado pela ordenação de todos os elementos de cada grupo.

Quanto maior H , maior a disparidade entre o posto médio amostral e o posto médio do conjunto formado pela união de todos os grupos, tal que a rejeição da hipótese nula acontece sempre que $H \geq H_{\text{crítico}}$.

A distribuição de H para $k = 3$ e $n_i \leq 5$ pode ser encontrada tabulada em CONOVER (CONOVER, 1980). Para valores maiores de n_i a distribuição de H pode ser aproximada pela χ^2 .

Frequentemente, quando o teste é significativo, indicando que há diferenças entre os grupos, deseja-se averiguar quais grupos são diferentes entre si. Este tipo de comparação é denominada comparação múltipla entre os grupos. Dunn (DUNN, 1964) propôs um método de comparação múltipla não-paramétrica, que se baseia na soma dos postos dos elementos de cada grupo. A estatística do teste baseia-se no parâmetro Q , dado por

$$Q = \frac{\frac{R_1}{n_1} - \frac{R_2}{n_2}}{\sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (2.14)$$

onde $N = n_1 + n_2$ é o número total de elementos. A distribuição de $Q_{\alpha, gl}$ pode ser encontrada em Zar (ZAR, 1999).

2.2.2 – Comparação Multivariada

Teste Λ de Wilks

O teste Λ de Wilks é uma forma de comparação entre distribuições multivariadas de um número g de populações. Este teste compara médias populacionais multivariadas, e é um tipo de Análise de Variância Multivariada (MANOVA). A hipótese nula H_0 supõe que as populações tenham médias multivariadas iguais. Este teste supõe que existe normalidade multivariada, homocedasticidade e linearidade entre as variáveis.

A estatística do teste é baseada no parâmetro Λ , dado por

$$\Lambda = \frac{|W|}{|B + W|} \quad (2.15)$$

onde $W = \sum_{i=1}^g \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$ é uma matriz que mede a variabilidade dos valores observados de uma população em relação à média daquela população, e é proporcional à soma das matrizes de covariância de cada população (JOHNSON E WICHERN, 2002). E

$B = \sum_{i=1}^g n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T$ mede a variabilidade das médias populacionais em relação à média total $\bar{X}$ dos dados.

Para $n = \sum_{i=1}^g n_i$ grande, pode-se mostrar (JOHNSON E WICHERN, 2002) que a transformação $-\left(n-1-\frac{p+g}{2}\right)\ln \Lambda$ segue uma distribuição χ^2 com $p(g-1)$ graus de liberdade, onde p é o número de variáveis.

2.3 – Reconhecimento de Padrões Utilizando Análise Multivariada

De forma geral, o reconhecimento de padrões pode ser definido como a categorização dos dados observados em classes identificáveis, através da comparação entre as características existentes em cada indivíduo do conjunto de dados. Vários métodos podem ser aplicados com este objetivo, como por exemplo, o uso redes neurais, métodos estatísticos, ou uma combinação de ambos (TOU E GONZALEZ, 1974).

Neste trabalho adotou-se a abordagem estatística, tendo sido utilizadas duas técnicas capazes de identificar padrões existentes num conjunto de dados (JAIN *ET AL*, 2000): a Análise de Componentes Principais e a Análise de Discriminante.

2.3.1 - Análise de Componentes Principais

A análise de componentes principais visa à obtenção de um conjunto reduzido de novas variáveis, denominadas componentes principais, a partir das variáveis originais sem que haja perda significativa de informação, simplificando assim a interpretação dos resultados. Embora seja uma técnica de redução de variáveis, é freqüentemente utilizada a fim de evidenciar padrões de comportamento que podem estar ocultos pelas variáveis originais (KRZANOWSKI, 1987).

Obtenção das Componentes Principais

Seja $X_1, X_2, \dots, X_m$ um conjunto de m variáveis. As componentes principais $CP_1, CP_2, \dots, CP_m$ são combinações lineares destas variáveis:

$$\begin{aligned} CP_1 &= a_{11}X_1 + a_{12}X_2 + \dots + a_{1m}X_m = a_1X \\ CP_2 &= a_{21}X_1 + a_{22}X_2 + \dots + a_{2m}X_m = a_2X \\ &\vdots \\ CP_m &= a_{m1}X_1 + a_{m2}X_2 + \dots + a_{mm}X_m = a_mX \end{aligned} \quad (2.16)$$

Onde os coeficientes a_{ij} de cada componente são determinados de forma a maximizar a variância de cada componente. A variância de cada CP_i é dada por:

$$\text{Var}(CP_i) = a_i' S a_i, \quad i=1, \dots, m \quad (2.17)$$

onde a_i' é a transposta de a_i , e S é a matriz de covariância.

A primeira componente principal é a combinação linear que maximiza $\text{Var}(CP_1)$. A segunda componente principal é obtida da mesma forma, mas com a restrição de que esta deve ser ortogonal à primeira, ou seja,

$$\text{Cov}(CP_2, CP_1) = 0 \quad (2.18)$$

onde $\text{Cov}(CP_2, CP_1)$ é a covariância entre as combinações lineares.

Assim, a i -ésima componente principal deve maximizar $\text{Var}(a_iX)$ com a restrição de que $\text{Cov}(a_iX, a_kX) = 0, \forall k < i$. Dessa forma obtém-se um conjunto de m combinações lineares independentes entre si.

Se $\lambda_i, i=1, \dots, m$ é o i -ésimo autovalor associado ao autovetor e_i da matriz de covariância S , pode-se mostrar que a variância da i -ésima componente principal CP_i é máxima se $\text{Var}(CP_i) = \lambda_i$, e esta relação mantém a restrição de ortogonalidade entre as componentes principais (JOHNSON e WICHERN, 2002). As componentes podem então ser escritas na forma

$$CP_i = e_i X = e_{i1}X_1 + e_{i2}X_2 + \dots + e_{im}X_m, \quad i=1, \dots, m \quad (2.19)$$

com

$$\begin{aligned} \text{Var}(CP_i) &= e_i' \Sigma e_i = \lambda_i \\ \text{Cov}(CP_i, CP_k) &= e_i' \Sigma e_k = 0, \quad i \neq k \end{aligned} \quad (2.20)$$

Geometricamente, as componentes principais representam uma rotação do eixo das coordenadas originais, na direção da maior variância dos dados, conforme mostra a figura abaixo.

Figura 2.4 – Componentes principais CP_1 e CP_2 obtidos para o conjunto de dados definidos pelas variáveis X_1 e X_2 (JOHNSON E WICHERN, 2002).

No caso em que as variáveis são medidas em escalas diferentes ou caso em que as variáveis apresentam variâncias com ordens de grandeza muito diferentes entre si, é conveniente calcular as componentes principais através da matriz de correlação R . A matriz R corresponde a uma padronização dos valores observados, impedindo que o módulo de cada componente principal seja muito influenciado por uma única variável ou um pequeno grupo de variáveis, cuja variância seja muito maior que a das outras.

Determinação do número de Componentes

Se $\text{Var}(CP_i) = \lambda_i$ é a variância da i -ésima componente principal, então a variância total é dada por

$$\sum_{i=1}^m \text{Var}(CP_i) = \sum_{i=1}^m \lambda_i \quad (2.21)$$

A proporção da variância total devida à k -ésima componente principal pode ser escrita como

$$p_k = \frac{\text{Var}(CP_k)}{\sum_{i=1}^m \text{Var}(CP_i)} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_m} \quad (2.22)$$

Se a maior parte da variância total for devida a um número reduzido de componentes, então as outras componentes podem ser eliminadas sem que haja perda relevante de informação. Desta forma a análise das m variáveis originais se resume à análise de algumas poucas componentes.

Uma forma gráfica de se determinar o número ideal de componentes principais para análise dos dados é através do *scree plot*. O *scree plot* é um gráfico onde os autovalores λ_i aparecem ordenados do maior para o menor no eixo horizontal, e no eixo vertical as respectivas porcentagens da variância total. O número ideal de componentes é dado pelo índice i onde os pontos referentes aos autovalores começam a ficar relativamente pequenos e praticamente constantes.

2.3.2 – Análise de Discriminante

A técnica de Análise de Discriminante visa a encontrar uma combinação linear das variáveis originais, a chamada função discriminante, que permita a melhor separação entre grupos de amostras. É a técnica mais indicada, portanto, quando o objetivo da análise dos dados é tornar evidente diferenças entre grupos.

O caso mais simples supõe que o conjunto de dados pertence a dois grupos distintos, e pode ser generalizado para o caso em que existam mais grupos.

Função discriminante para separação de duas populações

Considerando inicialmente o caso em que se tenham duas populações distintas, contendo respectivamente n_1 e n_2 elementos cada, então pode-se formar dois conjuntos de dados, X_1 e X_2 contendo os valores observados em cada população.

Assim como as componentes principais, a função discriminante Z também é obtida por uma combinação linear dos valores observados X conforme a expressão abaixo.

$$Z = aX \quad (2.23)$$

Onde a matriz a representa os coeficientes da combinação linear.

A combinação linear de X_1 fornece os valores $Z_{11}, Z_{12}, \dots, Z_{1n_1}$ para a primeira população contendo n_1 elementos e X_2 fornece $Z_{21}, Z_{22}, \dots, Z_{2n_2}$ para a segunda. A separação destes grupos em função dessas combinações lineares é dada então pela diferença dos valores médios $\bar{Z}_1$ e $\bar{Z}_2$ padronizadas em unidades de desvio padrão. Assim

$$\text{separação} = \frac{|\bar{Z}_1 - \bar{Z}_2|}{s_Z} \quad (2.24)$$

onde

$$s_Z^2 = \frac{\sum_{j=1}^{n_1} (Z_{1j} - \bar{Z}_1)^2 + \sum_{j=1}^{n_2} (Z_{2j} - \bar{Z}_2)^2}{n_1 + n_2 - 2} \quad (2.25)$$

é a variância combinada (*pooled variance*). O objetivo é então encontrar os coeficientes da matriz a que maximizem a separação entre as médias $\bar{Z}_1$ e $\bar{Z}_2$.

Pode-se mostrar (RENCHE, 1992) que a combinação linear que maximiza a separação entre Z_1 e Z_2 é dada pela função discriminante de Fisher:

$$Z = (\bar{X}_1 - \bar{X}_2)' S_{pooled}^{-1} X \quad (2.26)$$

onde S_{pooled} é a matriz de covariância combinada, formada pela média ponderada das matrizes de covariância S_1 e S_2 de cada população. Esta matriz é da forma

$$S_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} \right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} \right] S_2 \quad (2.27)$$

Também é possível mostrar (JOHNSON E WICHERN, 2002) que o valor máximo de separação entre as populações é dado por

$$D^2 = (\bar{X}_1 - \bar{X}_2) S_{pooled}^{-1} (\bar{X}_1 - \bar{X}_2) \quad (2.28)$$

O fator D^2 é conhecido como distância de Mahalanobis, e no caso acima mede a distância entre as médias X_1 e X_2 .

A figura abaixo exemplifica o efeito da aplicação do método de análise de discriminante. O efeito dessa técnica é projetar os valores observados de cada grupo no eixo definido pela combinação linear, e encontrar a direção em que estas projeções atinjam a separação máxima.

Figura 2.5 – Função discriminante Z que otimiza a separação entre duas populações bivariadas (JOHNSON E WICHERN, 2002).

Funções discriminantes para separação de $g > 2$ populações

Os resultados obtidos anteriormente para duas populações podem ser generalizados para o caso em que se tem um conjunto de g grupos, tal que $g > 2$.

O objetivo agora é encontrar a combinação linear que maximize a soma das diferenças da média de cada grupo em relação à média total (RENCER, 1992):

$$\frac{\sum_{i=1}^g (\bar{Z}_i - \bar{Z})}{s_Y} = \frac{a' Ba}{a' Wa} \quad (2.29)$$

onde as matrizes B e W foram definidas anteriormente (seção 2.2.2).

Pode-se mostrar que o vetor a que maximiza a equação acima pode ser escrito como autovetor da matriz $W^{-1}B$, e que haverá um número de $(g-1)$ autovetores positivos. (JOHNSON E WICHERN, 2002). Conseqüentemente, para um conjunto formado por g populações o número de funções discriminantes que se pode obter é $\min(g-1, p)$, onde p é o número de variáveis.

2.4 – Critério de Classificação Histológica

O critério de classificação dos elementos de cada grupo baseou-se na medida da distância D^2 de Mahalanobis do i -ésimo elemento ao centróide do j -ésimo grupo, dada por

$$D^2 = (X_i - \bar{X}_j)' S_{pooled}^{-1} (X_i - \bar{X}_j) \quad (2.30)$$

Cada elemento era classificado como pertencente ao grupo cujo centróide estava mais próximo.

Além disso, para testar a validade do resultado utilizou-se o processo de validação-cruzada, indicado nos casos em que o número de amostras em cada grupo é relativamente pequeno (HAIR *ET AL*, 1998). Neste procedimento, um elemento é retirado do conjunto e as componentes principais, ou funções discriminantes, são novamente calculadas, obtendo novos centróides em cada grupo. Mede-se então a distância D^2 entre as coordenadas deste elemento e estes centróides, classificando-o quanto à menor distância.

2.5 – Verificação da Eficácia dos Modelos de Diagnóstico

Em geral, quando se realiza um teste de diagnóstico, nem sempre um resultado positivo ou negativo corresponde respectivamente à presença ou ausência de uma alteração. A tabela abaixo mostra os resultados que podem ocorrer quando se realiza um teste diagnóstico.

Resultado do Teste	Existe alteração (Indivíduo doente)	Não existe alteração (Indivíduo normal)	Total
Positivo	Verdadeiro positivo (<i>VP</i>)	Falso positivo (<i>FP</i>)	$VP + FP$
Negativo	Falso negativo (<i>FN</i>)	Verdadeiro negativo (<i>VN</i>)	$FN + VN$
Total	$VP + FN$	$FP + VN$	

Tabela 2.1 – Resultados possíveis num teste de diagnóstico

Conhecendo-se então o número de acertos e erros no teste, os seguintes indicadores podem ser utilizados para averiguar sua eficiência (PEREIRA, 2005):

Sensibilidade (S): Probabilidade de obter um resultado verdadeiro-positivo, ou seja, de o método detectar a alteração quando ela realmente está presente. Se *VP* é o número de verdadeiros-positivos e *FP* o de falsos-positivos, a sensibilidade é dada por

$$S = VP / (VP + FN) \quad (2.31)$$

Especificidade (E): Probabilidade de obter um resultado verdadeiro-negativo, ou seja, a confiabilidade num resultado que não detecta uma anomalia. Se *VN* é o número de verdadeiros-negativos e *FN* o de falsos-negativos, a especificidade é dada por

$$E = VN / (VN + FP) \quad (2.32)$$

Valor Preditivo Positivo (VPP): Proporção de verdadeiros-positivos dentre todos os resultados positivos obtidos, dada por

$$VPP = VP / (VP + FP) \quad (2.33)$$

Valor Preditivo Negativo (VPN): Proporção de verdadeiros-negativos dentre todos os resultados negativos, dada por

$$VPN = VN / (VN + FN) \quad (2.34)$$

Os parâmetros S e E têm importância fundamental na confiabilidade de um teste de diagnóstico. Os valores VPP e VPN , por outro lado, não podem ser aplicados a populações diferentes daquelas utilizadas para sua estimação (MACMAHON, B., TRICHOPOULOS, D., 1996).

Outro parâmetro importante é o valor do parâmetro κ . Este parâmetro mede a reprodutibilidade do teste, isto é, o índice de concordância do resultado, quando comparado por métodos de avaliação independentes. Este índice varia entre -1 e +1, sendo que $\kappa=-1$ indica completo desacordo entre os avaliadores, $\kappa=0$ indica um diagnóstico dado ao acaso e $\kappa=+1$ indica concordância total quanto ao resultado. O valor de κ é obtido da seguinte forma (PEREIRA, 2005):

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (2.35)$$

onde P_o indica a proporção de concordâncias observadas, e P_e a proporção de concordâncias esperadas. Estes valores são dados por

$$P_o = \frac{VP + VN}{VP + VN + FP + FN}$$

$$P_e = \frac{(VP + FP)(VP + FN) + (FN + VN)(FP + VN)}{(VP + VN + FP + FN)^2} \quad (2.36)$$

CAPÍTULO 3

MATERIAIS E MÉTODOS

3.1 – Coleta das Amostras de Tecido Mamários

As amostras de tecido mamário estudadas foram obtidas no Departamento de Patologia do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto. O material colhido para análise correspondia a uma fração dos tecidos retirados em procedimentos cirúrgicos de mastectomia ou mastoplastia, e o volume de tecido colhido era determinado de forma a se ter uma amostra o mais homogênea possível e em quantidade suficiente para ser irradiada.

Depois de colhidas, as amostras eram armazenadas à temperatura ambiente em recipientes plásticos contendo solução de formol tamponado (10%).

Um total de 40 amostras de tecido mamário foram colhidas para análise.

3.2 – Classificação Histológica dos Tecidos Mamários

As amostras de tecidos mamários coletadas foram classificadas como tecidos normais (adiposo e glandular), fibroadenomas e carcinomas mucinoso, intraductal e ductais invasivos de grau I, II e III, conforme o critério de classificação de Bloom e Richardson (BLOOM E RICHARDSON, 1957). A classificação histológica foi realizada através da análise das lâminas de tecido por médicos do Departamento de Patologia do Hospital das Clínicas e posteriormente revisadas por um especialista em patologias mamárias.

3.3 - Obtenção dos Perfis de Espalhamento

3.3.1 – Preparação das Amostras

Os tecidos mamários foram retirados da solução de formol e cortados no tamanho adequado para serem acomodados em um recipiente acrílico cilíndrico (porta-amostra) construído com diâmetro interno de 18 mm e uma altura de 4 mm (figura 3.1). Uma fina camada de policloreto de vinila (filme de PVC), com espessura de $1.5 \pm 0.2 \mu\text{m}$, foi utilizada recobrindo a amostra de forma que esta última apresentasse superfície mais regular possível. Para se minimizar os efeitos do formol na medida do espalhamento, o excesso dessa solução na amostra foi retirado utilizando-se papel absorvedor.

Figura 3.1 – Porta-amostra utilizado.

3.3.2 - Arranjo experimental

Os perfis de espalhamento das amostras de tecido mamário foram obtidos num difractômetro comercial Siemens D-5005, operando no modo reflexão. A figura abaixo mostra um esquema do aparato utilizado.

Figura 3.2 – Arranjo experimental para as medidas de difração de raios-X.

O tubo de raios-X possuía um anodo de Cu ($Z=29$, $K_{\alpha} = 8.04\text{KeV}$ e $K_{\beta} = 8.91\text{KeV}$). A fenda divergente regulava a abertura do feixe emergente, para que a área irradiada na superfície da amostra se mantivesse constante, de dimensões $6\text{mm} \times 12\text{mm}$, e a fenda Söller tornava o feixe paralelo na direção horizontal.

O porta-amostra foi posicionado no centro do goniômetro localizado sobre o eixo central do difractômetro, e girava de forma que o ângulo incidente e o ângulo espalhado fossem iguais em relação à superfície da amostra. A figura abaixo mostra a área de tecido irradiada no porta-amostra.

Figura 3.3 – Área do tecido irradiada

O feixe de raios-X espalhado na amostra era colimado por uma segunda fenda divergente e em seguida atravessava uma fenda Söller, atingindo então o monocromador de Grafite ($2d=0.2708\text{nm}$), que selecionava os fótons espalhados elasticamente. A radiação espalhada atingia finalmente o detector de cintilação, cuja eficiência na detecção de fótons de energia 8,04 keV era de 95%. O detector estava associado à uma eletrônica padrão (fotomultiplicadora, amplificador) e conectado ao micro por uma placa de interface. O software de aquisição dos dados foi desenvolvido pelo próprio fabricante do difractômetro.

Os fótons espalhados eram detectados variando o ângulo de espalhamento no intervalo $5^\circ - 150^\circ$, em passos de $1/3$ de grau, com tempo de contagem de 20s, garantindo assim uma boa contagem de fótons no detector, com uma incerteza estatística de 3%. O tamanho do passo escolhido era suficiente para detectar a existência de picos de espalhamento, uma vez que a largura à meia altura dos picos era de no mínimo 2° para tecidos mamários normais e 9° para tecidos neoplásicos, na energia utilizada. Medidas adicionais foram realizadas para a subtração de contribuições espúrias: (i) com o porta-amostra vazio, mantida a camada de PVC, (ii) com o porta amostras sem o filme de PVC e (iii) sem nenhum tipo de alvo.

Os dados de espalhamento obtidos no difractômetro foram submetidos a um pré-processamento antes da análise, de forma a subtrair as contribuições espúrias, e corrigir efeitos de atenuação da amostra e variação da intensidade do feixe de RX em função do ângulo.

3.4 – Pré-processamento dos Dados Experimentais

O pré-processamento dos perfis de espalhamento obtidos experimentalmente foi realizado em 3 etapas:

- i. Aplicação de filtro para eliminação do ruído dos dados experimentais;
- ii. Subtração das contribuições espúrias, devido à presença de outras fontes de radiação espalhada durante o experimento;
- iii. Aplicação dos fatores de correção aos dados experimentais, com o intuito de obter o valor do coeficiente diferencial linear de espalhamento $\mu_s = n_v \frac{d\sigma}{d\Omega}$ de cada amostra.

Este coeficiente é proporcional à seção de choque diferencial de espalhamento e, portanto, é uma característica própria da amostra de tecido, sendo independente do processo de obtenção dos dados.

3.4.1 – Suavização do Ruído

Aos perfis de espalhamento obtidos experimentalmente, deve-se aplicar um filtro de suavização de ruído, de forma a amenizar a contribuição das frequências altas presentes em todo o espectro.

Optou-se por aplicar um filtro de média móvel. Este tipo de filtro é de fácil aplicação e teoricamente simples. Consiste em dividir o sinal em diversos intervalos, cuja largura é definida inicialmente. Substitui-se então o valor do ponto central do intervalo pelo valor médio do intervalo. Este processo é executado ao longo de todo o sinal, e o resultado é um sinal com menos ruído que o original.

O filtro de média móvel foi aplicado em todos os perfis de tecidos mamários obtidos, utilizando a função *filtfilt* do “*Signal Processing Toolbox*” do software *Matlab*[®]. Esta função aplica o filtro de média móvel duas vezes, primeiro do início do sinal até o último ponto, e em seguida na direção contrária. Dessa forma obtém-se um resultado com deslocamento de fase zero em relação ao sinal original, preservando assim as características dos perfis de espalhamento, como a posição dos picos, por exemplo.

O tamanho do intervalo utilizado para o cálculo da média foi de 3 pontos. Este intervalo permitiu reduzir as altas frequências de forma significativa, sem distorcer o sinal.

3.4.2 – Correção por contribuições espúrias

Considera-se como contribuição espúria ao perfil de espalhamento qualquer contagem de fótons que chegue ao detector, mas que não seja oriunda da amostra de tecido analisada. Devido ao arranjo experimental utilizado podemos ter várias dessas fontes de espalhamento, tais como o espalhamento do feixe de raios-X no ar, sobre o filme de PVC que cobre a amostra, e ainda devido ao fato de o feixe incidente na amostra não ser totalmente atenuado pela mesma e terminar atingindo o porta-amostra de acrílico, o que faz com que este também se torne um fonte de espalhamento de raios-X.

A figura 3.4 mostra um esquema simplificado da disposição de todos os elementos que contribuem para o sinal do espalhamento detectado.

Figura 3.4 - (a) Representação esquemática do arranjo experimental utilizado para medida das amostras de tecido. (b) Contribuição de cada parte do arranjo ao espalhamento.

A partir da figura acima pode-se escrever então a intensidade do espalhamento medida experimentalmente no difractômetro como a soma das seguintes contribuições:

$$I_{M(\theta)} = T_P I_{A(\theta)} + I_{P(\theta)} + T_P T_A I_{PA(\theta)} + I_{AR} \quad (3.1)$$

onde $I_{M(\theta)}$ é a intensidade medida experimentalmente, $I_{A(\theta)}$ é o espalhamento proveniente da amostra, $I_{P(\theta)}$ o espalhamento proveniente do filme de PVC que cobre a amostra, $I_{PA(\theta)}$ o espalhamento do porta-amostra e I_{AR} o espalhamento do ar. T_P e T_A são respectivamente os fatores de transmissão do filme de PVC e da amostra. Através da figura pode-se ver a necessidade de se considerar os fatores de atenuação do filme de PVC e da amostra, T_P e T_A respectivamente, uma vez que o feixe incidente na amostra é antes atenuado pelo filme de PVC, enquanto que o feixe que atinge o porta-amostra atravessa antes o filme de PVC e a amostra.

Para obter a intensidade do espalhamento devido exclusivamente à amostra reescreve-se a equação acima de forma a explicitar quais são os fatores que devem ser eliminados dos dados experimentais:

$$I_{A(\theta)} = T_P^{-1} (I_{M(\theta)} - I_{P(\theta)} - T_P T_A I_{PA(\theta)} - I_{AR}) \quad (3.2)$$

Os fatores de atenuação T_P e T_A são obtidos através da expressão $T = e^{-2\mu l}$, onde μ é o coeficiente de atenuação linear do material, e l é o caminho percorrido pelo feixe de raios-

X. O fator de atenuação da amostra T_A foi obtido utilizando cálculos numéricos e o valor de T_p foi considerado constante ($T_p=0,974$).

Para a obtenção do sinal de espalhamento do ar realizou-se uma medida sem que nenhum objeto estivesse sendo irradiado (figura 3.5a). O sinal de espalhamento do porta-amostra ($I_{PA(\theta)}$) foi obtido medindo-se o espalhamento do conjunto “porta-amostra + ar”, conforme mostra a figura 3.5b, e em seguida subtraindo o sinal do espalhamento devido ao ar $I_{AR(\theta)}$. Por fim para obter o sinal de espalhamento do filme de PVC ($I_{P(\theta)}$) mediu-se o espalhamento do conjunto “porta-amostra + filme de PVC + ar”, como mostra a figura 3.5c, subtraiu-se o sinal encontrado para o porta-amostra $I_{PA(\theta)}$ e também a contribuição do ar I_{AR} .

Figura 3.5 - Arranjo experimental utilizado para medida do espalhamento (a) do ar, (b) do porta-amostra e (c) do filme de PVC.

3.4.3 – Obtenção da Seção de Choque Diferencial de Espalhamento

Uma vez obtido o perfil de espalhamento da amostra $I_A(\theta)$, deve-se obter a seção de choque diferencial de espalhamento $d\sigma/d\Omega$ desta amostra. A obtenção da seção de choque segue basicamente a metodologia proposta por Kane (KANE *ET AL*, 1986). A intensidade $I_A(\theta)$ do espalhamento é igual ao número total de fótons espalhados em um dado ângulo, e se relaciona com a seção de choque pela equação:

$$I_A(\theta) = \int_V I_0(\theta) e^{-\mu l} \frac{d\sigma}{d\Omega}(\theta) n_v dV e^{-\mu l} \Delta\Omega(\theta) \eta_{mon} \eta_{det} \quad (3.3)$$

onde:

$I_0(\theta)$ número de fótons incidentes por unidade de área da amostra,
 μ coeficiente de atenuação linear da amostra,

$e^{-\mu\ell}$ atenuação durante o caminho ℓ percorrido da superfície ao ponto de espalhamento da amostra,

$\frac{d\sigma}{d\Omega}(\theta)$ seção de choque diferencial de espalhamento,

dV elemento de volume,

n_v número de partículas espalhantes por unidade de volume presentes em dV ,

$e^{-\mu\ell}$ atenuação durante o caminho ℓ percorrido pelo fóton desde o ponto de espalhamento até a superfície da amostra,

$\Delta\Omega(\theta)$ ângulo sólido compreendido entre o ponto de espalhamento à abertura do colimador do feixe espalhado,

η_{mon} eficiência do monocromador,

η_{det} eficiência do detector.

A equação (3.3) acima pode ser reescrita como:

$$I_A(\theta) = I_0(\theta) \frac{d\sigma}{d\Omega}(\theta) n_v \Delta\Omega(\theta) \eta_{mon} \eta_{det} \int_V e^{-2\mu\ell} dV \quad (3.4)$$

uma vez que a integral é calculada sobre o volume espalhador irradiado V que é visto pelo detector, e os termos $I_0(\theta)$, $\frac{d\sigma}{d\Omega}(\theta)$, $\Delta\Omega(\theta)$ e η_{mon} dependem somente do ângulo, n_v depende da amostra e η_{det} da energia.

O coeficiente diferencial linear de espalhamento $\frac{d\sigma}{d\Omega}(\theta) n_v$, que é proporcional à seção de choque diferencial de espalhamento $\frac{d\sigma}{d\Omega}(\theta)$, pode então ser escrito como

$$\frac{d\sigma}{d\Omega}(\theta) n_v = (I_{A(\theta)}) / (I_P K(\theta) A(\theta) \eta_{mon} \eta_{det}) \quad (3.5)$$

O termo $A(\theta) = \int_V e^{-2\mu\ell} dV$ na equação (3.5) representa a atenuação na amostra dos feixes incidente e espalhado de todos os diferenciais de volume.

Os termos $I_0(\theta)$ e $\Delta\Omega(\theta)$ na equação (3.4) dizem respeito à variação da intensidade do feixe incidente, que ocorre pelo fato de o número de fótons que atinge a superfície da amostra depender diretamente da abertura da fenda divergente. Estes termos podem ser combinados de forma a constituírem um fator $K(\theta)$, relativo à variação da intensidade do feixe incidente.

A metodologia para obtenção dos fatores de correção $K(\theta)$ e $A(\theta)$ podem ser encontrados em Oliveira (OLIVEIRA, 2006).

3.5 – Modelos Estatísticos para Diagnóstico através dos Perfis de Espalhamento

Foram utilizadas três abordagens para a determinação de um método que permitisse classificar corretamente as amostras de tecido através de suas seções de choque de espalhamento.

A primeira abordagem, mas simples, consistiu de definir algumas grandezas, isto é, variáveis, a partir dos perfis de espalhamento, e compará-las entre os grupos, através de técnicas de análise univariada.

A segunda abordagem consistiu em determinar uma faixa angular dos perfis de espalhamento e utilizar todo o sinal obtido nessa faixa para comparar os grupos. Considerando cada ângulo de espalhamento como uma variável independente, aplicou-se o método de Análise de Componentes Principais para encontrar padrões de comportamento nos perfis de espalhamento.

A terceira abordagem foi realizada sobre a mesma faixa angular definida anteriormente, e consistiu na aplicação da técnica de Análise de Discriminante, com o objetivo de otimizar a diferenciação entre os perfis de espalhamento de diferentes tipos histológicos.

Os procedimentos aplicados em cada abordagem foram realizados utilizando os softwares SPSS 13[®] e MATLAB 6.5[®].

3.5.1 – Modelo de Diagnóstico utilizando Variáveis Simplificadas

Definição de Variáveis a Partir dos Perfis de Espalhamento

Com base nas curvas de espalhamento obtidas para os tecidos mamários foram definidas as seguintes variáveis para análise: posição do primeiro e do segundo pico de espalhamento (P_1 e P_2 , respectivamente), intensidade do primeiro e do segundo pico (I_1 e I_2 , respectivamente), largura à meia altura do primeiro pico (LMA) e a razão entre as intensidades do segundo e do primeiro pico (I_2/I_1). A escolha destas variáveis foi motivada por trabalhos anteriores (EVANS *ET AL*, 1991; KIDANE *ET AL*, 1999; LEWIS *ET AL*, 2000; POLETTI, GONÇALVES, E MAZZARO, 2002a). A figura 3.6 exemplifica estas variáveis extraídas dos perfis de espalhamento dos tecidos.

Figura 3.6 – Variáveis selecionadas a partir das curvas de espalhamento.

Agrupamento dos Dados

As amostras foram divididas em três grupos, um grupo contendo os tecidos sadios (grupo TN), um grupo contendo os tecidos benignos, ou seja, os fibroadenomas (grupo FA), e o terceiro contendo os tecidos malignos, ou seja, os carcinomas (grupo C), sendo que este engloba todos os tipos de carcinomas analisados.

Estatística Descritiva das Variáveis

Um gráfico do tipo *Box-plot* foi feito para cada variável, comparando seus valores entre os diferentes grupos. Este tipo de gráfico permite comparar visualmente a distribuição de uma mesma variável para diferentes grupos, além de detectar pontos discrepantes na distribuição (*outliers*).

A normalidade das distribuições foi averiguada utilizando o teste de Kolmogorov-Smirnov, com a correção de Lilliefors (LILLIEFORS, 1967). Também foi testada a suposição de que as variâncias das variáveis entre os grupos eram homogêneas, utilizando o teste de Levene (JOBSON, 1991).

Comparação entre os grupos

A análise dos dados foi realizada através da aplicação de testes estatísticos para tentar diferenciar os grupos de dados definidos anteriormente.

Inicialmente pretende-se verificar se existe diferença entre tecidos normais (TN), benignos (FA) e malignos (C). Aplicou-se o teste de Kruskal-Wallis, o análogo não-paramétrico da Análise de Variância (ANOVA) para comparação de 3 ou mais grupos. Este teste foi aplicado em todas as variáveis, inclusive nas que apresentavam distribuição normal, visto que nestes casos tanto a ANOVA quanto o teste de Kruskal-Wallis levavam às mesmas conclusões (CALLEGARI-JACQUES, 2003).

Para as variáveis que tenham apresentado resultados significativos no teste de Kruskal-Wallis, realizou-se um teste de comparação múltipla não-paramétrica para averiguar quais grupos diferiam entre si.

Construção do Modelo de Diagnóstico

Foram selecionadas duas variáveis com maior capacidade de diferenciar os três grupos, formando uma distribuição bivariada. Calculou-se então a distância de Mahalanobis de todas as amostras ao centróide de cada grupo, classificando-as de acordo com o centróide mais próximo.

A validação do resultado classificatório foi realizada aplicando o processo de validação-cruzada. Para testar a validade do modelo como teste de diagnóstico foram obtidos os parâmetros sensibilidade, especificidade, valor preditivo positivo e valor preditivo

negativo. A reprodutibilidade do teste foi avaliada através do parâmetro κ , comparando o diagnóstico conferido pela análise histológica das lâminas de cada tecido com o resultado fornecido pelo modelo de diagnóstico.

3.5.2 – Modelo de Diagnóstico utilizando Componentes Principais

Seleção da faixa angular de interesse

Embora o perfil de espalhamento tenha sido obtido variando o detector da posição 5° a 150°, é conveniente analisar em quais sub-intervalos angulares os perfis de espalhamento de tecidos normais, benignos e malignos apresentam maior diferença entre si. Desta forma, seleciona-se qual a faixa angular de maior conteúdo informativo, além de reduzir o número de variáveis de interesse, otimizando futuras coletas de dados.

Um teste estatístico de Kruskal-Wallis foi realizado para todos os ângulos considerando cada amostra como pertencente à um dos grupos TN, FA ou C, de acordo com sua classificação histológica. Considerou-se cada ângulo como uma variável independente, e o teste foi aplicado sobre toda a faixa angular de 5° a 150°.

Estatística Descritiva das Variáveis

Antes de aplicar qualquer técnica de análise estatística, é necessário analisar as características da distribuição das variáveis utilizadas.

Os grupos de amostras utilizadas foram os mesmos definidos anteriormente, ou seja, o grupo TN, contendo os perfis de espalhamento de amostras normais, o grupo FA, contendo os benignos e o grupo C, contendo os malignos.

Num conjunto de variáveis independentes, para analisar a normalidade da distribuição multivariada é suficiente averiguar a normalidade de cada variável separadamente. O teste de Kolmogorov-Smirnov com a correção de Lilliefors foi aplicado às variáveis em cada grupo.

Outro fator importante é o da homogeneidade das variâncias de cada variável. Para testar essa hipótese em cada variável separadamente foi utilizado o teste de Levene.

A linearidade do conjunto de variáveis também deve ser averiguada. Esta característica supõe que exista uma relação de dependência linear entre as variáveis, e é

importante quando se trabalha com o coeficiente de correlação de Pearson. Uma forma de averiguá-la é através da análise dos resíduos de uma regressão linear multivariada. Para isso, considerou-se uma variável como sendo dependente das outras e feito então um ajuste linear, e o mesmo procedimento repetido para cada variável.

Obtenção das Componentes Principais

Uma forma de evidenciar diferenças entre grupos é substituir as variáveis originais, ou seja, os ângulos de espalhamento, por um novo conjunto de variáveis, as componentes principais, e tentar identificar quais destas componentes fornecem uma melhor distinção.

A técnica de componentes principais consiste numa rotação do sistema de coordenadas definido pelo conjunto de variáveis utilizado, de forma a originar um novo sistema de coordenadas orientado na direção de maior variância do conjunto de dados. É na verdade uma técnica de redução de variáveis, embora seja utilizada freqüentemente para observar diferenças entre grupos de indivíduos, e é com esse objetivo que será aplicada aos perfis de espalhamento.

A aplicação da técnica de componentes principais foi feita de duas formas. No primeiro método, as componentes foram extraídas da matriz de covariância S do conjunto formado pelos grupos TN, FA e C, e no segundo, da matriz de correlação R destes grupos.

Método 1: Análise das Componentes da Matriz S

Os dados para análise foram organizados de forma que as linhas representavam as amostras de tecido e as colunas representavam as variáveis, formando assim uma matriz de dados D de dimensões 40×136 .

Para obter as componentes principais o primeiro passo é encontrar os autovalores λ da matriz de covariância, através da equação

$$\det(S - \lambda I) = 0 \quad (3.6)$$

Foi obtido um conjunto de 136 autovalores, que foram dispostos em ordem decrescente de acordo com seu valor. A parcela de variância p_k relacionada a cada autovalor pode ser encontrada dada pela equação 3.7.

As componentes principais são os autovetores associados a cada autovalor e formavam uma matriz P de dimensões 136×136 .

A projeção das componentes sobre a matriz de dados D fornece uma matriz T , que contém as coordenadas de cada amostra no novo sistema de variáveis. Esta matriz é dada pela equação

$$T = D.P \quad (3.7)$$

Para determinar quais componentes diferenciavam os grupos, aplicou-se um teste de hipóteses, primeiramente considerando todos os grupos TN, FA e C, e em seguida apenas nos grupos FA e C.

Método 2 – Análise das Componentes da Matriz R

As componentes principais também podem ser obtidas utilizando a matriz de correlação R ao invés da matriz de covariância S . A única condição que deve ser cumprida para o uso da matriz R é que exista linearidade entre as variáveis, uma vez que os coeficientes de correlação somente identificam a existência de dependência linear entre elas.

Para encontrar as componentes principais, obtiveram-se os autovalores da matriz R , e em seguida seus autovetores, seguindo o mesmo procedimento utilizado na etapa anterior.

Comparação entre os grupos

Para as duas etapas descritas acima, um teste de Kruskal-Wallis foi aplicado às componentes, para descobrir quais eram capazes de diferenciar os grupos. Às componentes com resultado significativamente estatístico foi realizado também um teste de comparação múltipla.

Construção do Modelo de Diagnóstico

Selecionou-se duas componentes capazes de melhor diferenciar cada grupo, e a classificação de cada amostra foi feita comparando-se as distâncias de Mahalanobis dos centróides de cada grupo às amostras, utilizando como critério a menor distância.

A validação cruzada foi utilizada para confirmar a classificação obtida. Neste processo, uma amostra é retirada de seu grupo, e as componentes principais novamente obtidas. Em seguida, este elemento é incluído no conjunto de dados, e identifica-se qual o grupo com centróide mais próximo das coordenadas desta amostra, classificando-a como pertencente a este grupo. Este processo é então repetido para cada amostra.

Os níveis de sensibilidade, especificidade, valor preditivo positivo e valor preditivo negativo foram obtidos para avaliar a validade do método como teste de diagnóstico, e sua reprodutibilidade foi comparada com a classificação histopatológica previamente realizada através do parâmetro κ .

3.5.3 – Modelo de Diagnóstico utilizando Análise de Discriminante

A análise de discriminante consiste em realizar uma combinação linear das variáveis originais, cujo objetivo é separar os grupos da melhor forma possível. Essas combinações são chamadas de funções discriminantes de Fisher.

Esta técnica requer que as variáveis tenham distribuição normal, e que haja uma relação de linearidade entre elas. Além disso, é necessário também que as matrizes de covariância dos grupos sejam homogêneas. Todas estas hipóteses foram testadas anteriormente e aceitas dentro do nível de significância de 0,05.

Obtenção das Funções Discriminantes

Foram utilizados dois métodos para o cálculo das funções discriminantes: o método direto e o método passo-a-passo (*stepwise*).

Método 1: Método Direto

Este método, também chamado método simultâneo, consiste em calcular as funções discriminantes utilizando todas as variáveis do conjunto de dados, independentemente da capacidade discriminatória de cada uma na separação dos grupos.

Entretanto, nem todas as variáveis entram indiscriminadamente no cálculo das funções. Cada variável é submetida a um teste de tolerância, que visa a identificar a existência

de multicolinearidade entre elas, ou seja, se existem variáveis cuja correlação é tão forte a ponto de uma poder ser totalmente explicada por uma combinação das outras. Este tipo de redundância prejudica a capacidade de produzir funções capazes de separar satisfatoriamente os grupos, o que ocorre devido ao fato de uma parcela da variância total dos dados pertencer a variáveis que não acrescentem nenhum tipo de informação relevante no conjunto de dados (HAIR ET AL, 1998). A tolerância de uma variável i é dada pelo parâmetro $TOL_i = 1 - R_i^2$, onde R_i^2 é obtido através de um modelo de regressão linear (SEN E SRIVASTAVA, 1990).

Obtiveram-se então os autovalores e autovetores da matriz $W^{-1}B$, calculados utilizando as variáveis aceitas no teste de tolerância. Como a análise de discriminante está sendo aplicada para tentar diferenciar três grupos de dados, o número de autovalores encontrados é $\min(g-1, p)=2$. Os autovetores de $W^{-1}B$ fornecem o coeficiente da matriz a , que maximiza a separação entre os grupos.

As funções discriminantes foram obtidas pela projeção das variáveis originais na matriz de coeficientes a .

Método 2: Método *Stepwise*

O método *stepwise* é uma alternativa ao método direto, e consiste em selecionar as variáveis que serão utilizadas no cálculo das funções discriminantes baseando-se em algum critério de seleção, e em seguida calculando as funções discriminantes em etapas (*steps*). A cada etapa, a variável que satisfaz o critério de seleção é adicionada ao cálculo. Vários critérios podem ser utilizados, como a distância D^2 de Mahalanobis, Rao's V, Wilk's λ , dentre outros (HAIR ET AL, 1998). Este método é indicado quando se deseja averiguar quais variáveis contribuem significativamente para a separação dos grupos, e quais podem ser descartadas sem comprometer o resultado final.

O critério de seleção de variáveis adotado neste trabalho foi a medida da distância D^2 de Mahalanobis entre os centróides $\bar{Z}_1$ e $\bar{Z}_2$ de cada grupo. Na primeira etapa as funções discriminantes são calculadas com a variável que maximiza a distância de Mahalanobis entre os três grupos. Nas etapas seguintes, as variáveis que também satisfazem a este critério são adicionadas sucessivamente ao cálculo das funções.

Uma vez selecionadas as variáveis com maior capacidade de separar os grupos, o processo de obtenção das funções discriminantes é o mesmo descrito no método direto.

Comparação entre os grupos

Um teste de Kruskal-Wallis foi aplicado a cada função separadamente a fim de verificar se eram capazes de diferenciar os grupos. Além disso, um teste Λ de Wilks foi realizado considerando a distribuição bivariada formada pelas funções.

Construção do modelo de diagnóstico

Como nos outros dois modelos de diagnóstico, o critério de classificação de uma amostra baseou-se na medida de sua distância de Mahalanobis em relação aos centróides de cada grupo.

A validação cruzada também foi aplicada, retirando uma amostra e recalculando as funções discriminantes. Comparou-se então a projeção desta amostra neste novo espaço, classificando-a quanto à sua proximidade aos centróides.

A sensibilidade, especificidade, valor preditivo positivo e negativo também foram obtidas para analisar a validade das funções discriminantes como teste de diagnóstico, bem como o parâmetro κ para analisar a reprodutibilidade do teste.

CAPÍTULO 4

RESULTADOS E DISCUSSÕES

4.1 – Obtenção dos Perfis de Espalhamento

4.1.1 - Amostras de Tecidos Mamários Analisadas

Foram coletadas 40 amostras de tecidos mamários, classificadas como tecido normal, fibroadenoma (neoplasia benigna) e diversos tipos de carcinomas (neoplasias malignas), tais como carcinomas ductais de grau I, II e III, intraductal e mucinoso. A figura 4.1 mostra a proporção do número de tecidos analisados em função de sua classificação histológica.

Figura 4.1 – Quantidade de amostras analisadas. TN = tecidos normais, FA = fibroadenomas, CDIS = carcinoma ductal “in situ”, CDI, CDII, CDIII = carcinomas ductais de grau I, II e III respectivamente e CM = carcinoma mucinoso.

4.1.2 – Obtenção Experimental da Distribuição Angular de Fótons Espalhados

Para cada amostra de tecido foi obtida a distribuição angular do número de fótons espalhados, utilizando o difractômetro Siemens D-5005. Além disso, foi realizada também uma medida do perfil de espalhamento para uma amostra de água líquida. A figura 4.2 mostra o resultado obtido para a amostra de água, tecidos mamários normais adiposos e tecidos alterados.

Figura 4.2 – Distribuição angular dos fótons espalhados. (a) Água, (b) Tecido Adiposo Normal, (c) Neoplasia benigna (fibroadenoma) e (d) maligna (carcinoma ductal grau III).

Os dados experimentais obtidos mostram que o tecido adiposo é o que mais se difere dos demais, com o primeiro pico de espalhamento ocorrendo em 20°. Fibroadenomas e carcinomas apresentam comportamento semelhante à água, variando apenas a intensidade das curvas.

4.2 – Pré-Processamento dos Dados Experimentais

Os perfis de espalhamento obtidos foram submetidos a um pré-processamento antes de serem analisados. Este processo tem como finalidade remover as contribuições espúrias ao

perfil de espalhamento do tecido, e obter a seção de choque diferencial de espalhamento ($d\sigma/d\Omega$) para cada tecido, através da aplicação de fatores de correção.

A figura 4.3 mostra o perfil de espalhamento para a amostra de água, após o pré-processamento, comparada com o valor de referência, obtido a partir do fator de forma tabulado por Morin (MORIN, 1982).

Figura 4.3 – Seção de choque diferencial de espalhamento para água líquida, comparada com o resultado obtido através do fator de forma tabulado por Morin (MORIN, 1982).

A figura 4.3 mostra que a seção de choque de espalhamento obtida experimentalmente para água está de acordo com o resultado prévio, mostrando que a metodologia de obtenção e processamento dos dados experimentais é adequada para determinar a seção de choque diferencial do espalhamento elástico.

A figura 4.4 mostra o perfil de espalhamento das amostras de tecido mamário após serem pré-processadas. Cada curva corresponde ao valor médio dos diferentes tipos de tecido mamário.

Figura 4.4 - Valores médios das seções de choque diferencial de espalhamento dos tipos de tecidos analisados.

Cabe ressaltar que tecidos normais são compostos basicamente por tecido adiposo e fibroglandular, além do estroma. O perfil de espalhamento para um tecido normal pode apresentar basicamente dois comportamentos, dependendo da proporção entre a quantidade dos componentes que o constituem. A figura 4.5 mostra a comparação entre dois casos de tecidos mamários normais encontrados, um com excesso de tecido adiposo, e outro composto por tecido glandular, com pequena quantidade de tecido adiposo.

Figura 4.5 - Seção de Choque para tecidos normais com diferentes componentes histológicos.

Pode-se ver que o tecido mamário composto de células adiposas tem o primeiro pico de espalhamento bem proeminente em 20° . O tecido mamário constituído de tecido glandular apresenta o pico adiposo quase inexistente, mas por outro lado, apresenta um pico de espalhamento característico em 30° , e comportamento bem semelhante ao da água.

As amostras de fibroadenomas analisadas indicam que este tipo de neoplasia apresenta sempre o mesmo comportamento, com o primeiro pico de espalhamento surgindo em $\theta=30^\circ$, variando apenas a intensidade da curva. Somente no caso de haver entre os componentes do fibroadenoma a presença de células adiposas é que este comportamento se altera. A presença destas células faz surgir o pico adiposo característico em 20° , como mostra a figura 4.6.

Figura 4.6 – Perfil de espalhamento de fibroadenomas com presença de células adiposas.

Os carcinomas também apresentam sempre o mesmo comportamento, com o primeiro pico de espalhamento surgindo em 30° . Foram analisados diferentes tipos destes tecidos, tais como carcinomas intraductais, ductais (grau I, II, e III) e mucinoso, e observou-se que os perfis de espalhamento apresentam sempre o mesmo padrão, variando apenas a diferença de intensidade das curvas. Entretanto, como o número de amostras é relativamente pequeno, não é possível afirmar se existe uma relação entre o tipo de carcinoma e a intensidade do espalhamento. A figura 4.7 mostra o perfil de espalhamento médio para cada tipo de carcinoma analisado.

Figura 4.7 - Perfil de espalhamento médio dos diferentes tipos de carcinomas analisados.

Deve-se notar a semelhança entre os perfis de espalhamento de fibroadenomas e carcinomas. Apesar de na média os fibroadenomas terem intensidades menores que os carcinomas, a variação destas intensidades é muito grande, mesmo para os tecidos de mesma classificação histológica, impossibilitando afirmar se realmente existiriam diferenças significativas entre os perfis de tecidos benignos e malignos.

Assim como no caso dos fibroadenomas, a presença de células adiposas entre as células neoplásicas que constituem o carcinoma também dá origem a um pico adiposo em 20°.

4.3 – Modelos Estatísticos para Diagnóstico através dos Perfis de Espalhamento

Estão descritas a seguir as três abordagens utilizadas para classificar as amostras de tecidos mamários.

4.3.1 - Modelo de Diagnóstico utilizando Variáveis Simplificadas

Este modelo baseou-se na definição das variáveis P_1 , P_2 , I_1 , I_2 , LMA, e I_2/I_1 , definidas na seção 3.5.1. A partir delas, buscou-se identificar quais eram as que melhor forneciam uma separação entre os grupos, e que poderiam ser utilizadas para classificá-los.

Gráficos *Box-plot* das variáveis

Para cada variável foi feito um gráfico do tipo *box-plot*, com o objetivo de visualizar diferenças existentes nas distribuições de cada grupo. A figura 4.8 mostra os resultados obtidos.

Figura 4.8 – *Box-plot* das variáveis P_1 , P_2 , I_1 , I_2 , LMA, e I_2/I_1 .

Observa-se que algumas variáveis, como por exemplo, P_1 e LMA permitem discriminar entre tecidos normais (TN) e alterados (FA e C). A variável I_2 também parece

indicar alguma diferença entre eles, embora visualmente não seja possível afirmar se ela é realmente significativa.

Deve-se ressaltar também o fato de nenhuma das variáveis ter evidenciado, ao menos visualmente, alguma diferença entre os tecidos benignos (grupo FA) e malignos (grupo C).

Pode-se notar nos gráficos a existência de alguns pontos discrepantes (*outliers*) em cada grupo. Estes pontos surgem pelo fato de amostras de tecido mamário de indivíduos diferentes apresentarem porcentagens diferentes de algum componente histológico. No caso de tecidos normais, por exemplo, pode haver um excesso de componente glandular ou adiposo. A figura 4.5 exemplifica este tipo de situação. Na figura, nota-se que um tecido mamário composto basicamente por tecido glandular apresenta comportamento bem diferente da média do grupo. Este tecido seria, portanto, considerado como um *outlier* em algumas variáveis, como LMA, por exemplo. Por outro lado, uma mama puramente adiposa seria um *outlier* na variável I_1 . Entretanto, ambos os casos representam um tecido mamário absolutamente normal. Por este motivo, optou-se por não retirar os *outliers*, uma vez que eles acrescentam ao conjunto de dados uma variabilidade que pode ocorrer em casos reais.

Comparação entre os grupos

A comparação entre as distribuições das variáveis em cada grupo foi efetuada através de um teste de hipótese. Uma vez que algumas variáveis apresentam distribuições não normais, além de heterocedasticidade (variâncias não homogêneas) entre grupos, utilizou-se o teste não-paramétrico de Kruskal-Wallis.

A tabela 4.1 mostra o resultado do teste, comparando tecidos normais, benignos e malignos. A coluna χ^2 dá a estatística do teste, e a gl diz respeito aos graus de liberdade referentes ao número g de grupos ($gl=g-1$). Foram considerados significativamente diferentes grupos que apresentassem no teste um nível de significância $\alpha < 0,05$.

Variável	χ^2	gl	α
P ₁	29,110	2	0,000
P ₂	11,077	2	0,004
LMA	28,669	2	0,000
I ₁	7,893	2	0,128
I ₂	4,111	2	0,000
I ₂ /I ₁	22,342	2	0,019

Tabela 4.1 – Teste de Kruskal-Wallis para as variáveis P₁, P₂, I₁, I₂, LMA, e I₂/I₁.

Com exceção da variável I_1 , todas as outras apresentam diferenças entre três os grupos, como fica evidente pelos valores de α obtidos para cada variável.

Entretanto, o resultado deste teste deve ser analisado com cautela, pois é um teste que compara simultaneamente um número g de grupos, e um resultado significativo não necessariamente implica que todos os grupos são diferentes entre si. A existência de um único grupo muito diferente dos demais pode tornar o teste significativo, mesmo que os demais sejam semelhantes entre si (CALLEGARI-JACQUES, 2003).

Como complemento ao teste de Kruskal-Wallis, foi feito um teste de comparação múltipla entre os grupos. Este teste analisa simultaneamente pares de grupos, e permite revelar quais são diferentes entre si. Um intervalo de confiança foi construído com nível de significância $\alpha=0,05$. Se os intervalos referentes aos grupos se sobrepõem, então aceita-se a hipótese de que não há diferença entre eles. Caso os intervalos estejam bem separados, conclui-se que os grupos são diferentes.

A figura 4.9 mostra o resultado da comparação múltipla para cada variável, exceto para I_1 , que não havia identificado nenhuma diferença entre os três grupos.

Figura 4.9 - Intervalos de Confiança construídos para $\alpha=0,05$.

Analisando os intervalos de confiança obtidos entre os pares de grupos, nota-se que o grupo TN mostra-se diferente dos demais para todas as variáveis, exceto para I_2/I_1 , que não diferencia os normais dos fibroadenomas. Os grupos FA e C, por outro lado, não são estatisticamente diferentes em nenhuma das variáveis. Isso mostra que o resultado significativo obtido no teste de Kruskal-Wallis foi devido ao grupo TN, o que está de acordo com a comparação feita pelo *box-plot* destes grupos.

Modelo de Diagnóstico

Mesmo não havendo nenhuma variável capaz de diferenciar entre benignos e malignos, aplicou-se um teste de Mann-Whitney entre FA e C, para averiguar qual delas

apresenta a menor significância. O resultado mostrou que I_1 é capaz de diferenciá-los com $\alpha = 0,47$, indicando que há apenas 53% de chance de a afirmação de que eles sejam diferentes seja verdadeira. Apesar de seu desempenho ruim, optou-se por tentar classificar os elementos utilizando esta variável, combinada com a LMA, que permite diferenciar normais de alterados.

A figura 4.10 mostra o gráfico de $LMA \times I_1$. Os pontos indicados com o símbolo (■) mostram os centróides de cada grupo, isto é, o ponto médio das coordenadas de seus elementos.

Figura 4.10 – Gráfico de $LMA \times I_1$, mostrando as coordenadas de cada grupo.

A figura 4.10 mostra que LMA separa muito bem os tecidos normais dos alterados, mas não diferencia entre benignos e malignos, como mostrado no teste de comparação múltipla. Por outro lado, I_1 não diferencia entre normais e alterados, e nem parece mostrar alguma separação entre benignos e malignos, devido à baixa significância do teste de comparação entre os dois. A classificação das amostras de tecidos foi realizada através do seguinte método: calculou-se o centróide de cada grupo, isto é, a média das coordenadas dos elementos que o constitui, e em seguida mediu-se a distância de Mahalanobis destes

elementos em relação aos três centróides, classificando-os histologicamente de acordo com o centróide mais próximo. A tabela 4.2 mostra o resultado obtido.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	17 (100 %)	0 (0 %)	0 (0 %)	17
FA	1 (11,1%)	3 (33,3 %)	5 (55,6 %)	9
C	0 (0 %)	7 (50,0 %)	7 (50,0 %)	14

Tabela 4.2 – Classificação das amostras utilizando LMA e I_1 .

Na tabela, os valores diagonais representam o número de elementos de cada grupo corretamente classificados. Os valores não-diagonais mostram o número de elementos classificados erroneamente e os números em parênteses mostram os percentuais em relação ao total de cada grupo.

Nota-se que a quantidade de elementos corretamente classificados não é muito expressiva, com um índice de acerto geral de 67,5%. A classificação dos tecidos normais é muito boa, devido à variável LMA que os separa dos demais. Entretanto, a classificação de benignos e malignos é bastante equivocada, como já era esperado, pelo fato de I_1 não diferenciar os dois grupos com um nível de significância satisfatório.

Para confirmar este resultado e avaliar a capacidade destas variáveis em classificar um novo indivíduo, aplicou-se o processo de validação-cruzada. Neste processo, retirou-se um elemento de um grupo e obtiveram-se as coordenadas do centróide daquele grupo. Comparou-se então novamente a distância deste elemento ao novo centróide, e o elemento recebia a classificação de acordo com o centróide mais próximo. Este processo é repetido para todos os elementos de cada grupo. A tabela 4.3 mostra o resultado obtido.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	17 (100 %)	0 (0 %)	0 (0 %)	17
FA	1 (11,1%)	2 (22,2 %)	6 (66,7 %)	9
C	0 (0 %)	9 (64,3 %)	5 (35,7 %)	14

Tabela 4.3 – Processo de validação cruzada das variáveis LMA e I_1 .

A validação cruzada mostra que os tecidos normais são bem classificados, ao contrário dos benignos e malignos, cujas classificações se confundem. O índice geral de acerto na classificação foi de 60%.

Foram obtidos os parâmetros que medem a sensibilidade S , especificidade E , valor preditivo positivo VPP , e valor preditivo negativo VPN , e o índice de concordância κ , para testar a credibilidade destas variáveis como modelo diagnóstico. A fim de analisar a capacidade do modelo em detectar alguma alteração no tecido, considerou-se inicialmente como resultado positivo qualquer diagnóstico de alteração no tecido, seja ela benigna ou maligna. Em seguida, para analisar a capacidade do modelo em diagnosticar as alterações encontradas como benignas ou malignas, considerou-se como resultado positivo apenas o diagnóstico de alteração maligna, e como negativo o de alterações benignas. A tabela 4.4 mostra o resultado encontrado.

Resultado do Diagnóstico	S (%)	E (%)	VPP (%)	VPN (%)	κ
Negativo = TN Positivo = FA ou C	95,6	100,0	100,0	94,4	1,0
Negativo = FA Positivo = C	35,7	25,0	45,4	18,2	-0,3

Tabela 4.4 – Validação e reprodutibilidade do teste de diagnóstico

Os valores obtidos para estes parâmetros mostram que as variáveis utilizadas fornecem resultados bastante confiáveis quando se tenta detectar alguma anomalia no tecido mamário. Entretanto, quando se tenta diagnosticar esta anomalia mais precisamente, como benigna ou maligna, o modelo é incapaz de fornecer resultados confiáveis. O alto valor de κ na classificação entre normais e alterados confirma a alta sensibilidade do método para detectar qualquer alteração. Entretanto, quando se compara o diagnóstico de benignos e malignos, o valor de κ próximo de zero indica a aleatoriedade deste resultado.

4.3.2 - Modelo de Diagnóstico utilizando Componentes Principais

Seleção da faixa angular para análise

Para determinar qual o melhor intervalo angular para análise aplicou-se um teste de Kruskal-Wallis em cada posição angular, comparando os grupos TN, FA e C. O resultado está na figura 4.11.

Figura 4.11 – Teste de Kruskal-Wallis comparando TN, FA e C em cada ângulo de espalhamento.

Da figura 4.11 se observa que para o nível de significância $\alpha_c=0,05$ existe uma região angular, em torno de 20° , onde os grupos são significativamente diferentes. É nessa região onde se situa o pico adiposo dos tecidos normais.

Todos os ângulos que não apresentavam diferenças significativas foram removidos, com exceção da região próxima aos picos de espalhamento de cada tecido, como, por exemplo, de 22° a 25° .

A região de 25° até aproximadamente 60° também apresentou diferenças significativas. É na região em torno de 30° onde estão localizados os picos de espalhamento dos tecidos alterados.

Portanto, de todo o sinal espalhado medido experimentalmente, foi considerado para análise a região compreendida entre 15° e 60° . Neste intervalo angular tem-se um conjunto de 136 variáveis, uma vez que durante a realização do experimento o detector variava a posição angular em passos de $1/3$ de grau.

Obtenção das Componentes Principais

As componentes principais foram obtidas de duas maneiras, a primeira utilizando a matriz de covariância S do conjunto total de dados, e a segunda através da matriz de correlação R .

Método 1: Análise das Componentes da Matriz S

Foram extraídos inicialmente os autovalores da matriz de covariância total S , que incluía as covariâncias dos grupos TN, FA e C. A figura 4.12 mostra a parcela de variância p_k correspondente ao k -ésimo autovalor, e a porcentagem de variância acumulada pelos autovalores.

Figura 4.12 – (a) Porcentagem de variância p_k correspondente a cada autovalor. (b) Porcentagem de variância acumulada.

Do primeiro gráfico observa-se que o primeiro autovalor explica 72,9% da variância total dos dados. O segundo corresponde a 25,5% e o terceiro a 1,3%. Juntos, portanto os três autovalores possuem 99,5% da variância total, como pode ser visto pelo segundo gráfico. Todos os autovalores juntos correspondem a 0,5% da variância.

Portanto, se o objetivo fosse reconstituir todo o sinal de espalhamento das amostras, seria possível fazê-lo de forma satisfatória utilizando somente as três primeiras componentes principais, ao invés de trabalhar com as 136 variáveis originais.

Entretanto, o objetivo deste trabalho é tentar encontrar diferenças entre os perfis dos diferentes grupos que permitam classificá-los quanto à sua estrutura histológica. Apesar de as três primeiras componentes conterem praticamente toda a variância dos dados de espalhamento, se houver alguma diferença sutil entre os grupos, ela poderia aparecer nas componentes subsequentes. Portanto a análise das componentes de menor variância também é importante.

Os gráficos da figura 4.13 mostram o comportamento das 9 primeiras componentes principais.

Figura 4.13 – Componentes Principais obtidas da matriz S total dos grupos TN, FA e C.

Na figura 4.13, o eixo horizontal está relacionado com as variáveis originais θ_i por uma rotação de coordenadas. Entretanto optou-se por representá-lo apenas como assumindo o valor do indicador i da variável, uma vez que seu valor real não é relevante para a análise das componentes. A principal informação que se pode extrair da figura é o comportamento de cada componente.

Pode-se observar que cada componente subsequente carrega menos informação que as anteriores. O comportamento da componente CP₉, por exemplo, já não parece corresponder a nenhum padrão, se assemelhando mais a um sinal aleatório, ou seja, um ruído, devido à parcela de variância quase nula a que ela correspondente.

Calculou-se então a projeção dos dados originais nestas componentes, obtendo as coordenadas dos elementos de cada grupo no novo espaço vetorial.

Comparação entre grupos

Para averiguar quais componentes são úteis em distinguir os grupos realizou-se um teste de hipótese. Foram testadas as primeiras 8 componentes, que juntas correspondiam à 99,9% da variância total dos dados. Os testes de normalidade e homocedasticidade para cada

componente foram realizados, e algumas componentes foram rejeitadas nestes testes. Optou-se então por comparar os grupos através do teste de Kruskal-Wallis. A tabela 4.5 mostra o resultado deste teste.

Variável	χ^2	α
CP ₁	20,042	0,000
CP ₂	8,175	0,017
CP ₃	2,833	0,243
CP ₄	2,936	0,230
CP ₅	0,976	0,614
CP ₆	2,447	0,294
CP ₇	0,147	0,929
CP ₈	0,741	0,691

Tabela 4.5 – Teste de Kruskal-Wallis para as primeiras 8 componentes principais da matriz *S*.

Da tabela 4.5 pode-se ver que apenas as componentes CP₁ e CP₂ são capazes de identificar alguma diferença entre os grupos. Para descobrir quais grupos são diferentes entre si, realizou-se um teste de comparação múltipla destas componentes. O resultado mostrou que estas elas apenas diferenciam entre normais e alterados, e nenhuma das duas é capaz de distinguir entre benignos e malignos.

Para descobrir qual componente melhor diferencia tecidos benignos e malignos, realizou-se um teste de Mann-Whitney. O resultado mostrou que a componente CP₄ fornece o melhor resultado, com $\alpha=0,124$.

Modelo de Diagnóstico

Para classificar as amostras de tecido foram utilizadas as componentes CP₁, que diferencia o grupo TN dos grupos FA e C, e a componente CP₄, que melhor diferencia entre FA e C, mesmo não significativamente.

A figura 4.14 mostra o gráfico de CP₁×CP₄. Pode-se notar que no eixo horizontal é possível distinguir tecidos normais e alterados, enquanto no vertical distinguem-se razoavelmente benignos e malignos, mas não os normais.

Figura 4.14 – Posição das amostras no espaço $CP_1 \times CP_4$.

A tabela 4.6 mostra o resultado classificatório baseado nas distâncias de cada elemento aos centróides dos três grupos.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	14 (82,2%)	2 (11,8%)	1 (6,0%)	17
FA	1 (11,1%)	4 (44,4%)	4 (44,4%)	9
C	2 (14,3%)	3 (21,4%)	9 (64,3%)	14

Tabela 4.6 – Classificação das amostras utilizando CP_1 e CP_4 .

O índice de acerto geral foi de 67,5% dos casos, sendo que a maior contribuição para este índice é a grande quantidade de tecidos normais corretamente classificados. O grupo dos fibroadenomas, por outro lado, apresenta cerca de metade de seus elementos classificados erroneamente.

Para confirmar este resultado, aplicou-se o processo de validação cruzada, adotando o mesmo critério de classificação das amostras.

A tabela 4.7 mostra o resultado obtido. Do total de 40 amostras, apenas 62,5% foram corretamente classificadas.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	13 (76,4%)	2 (11,8%)	2 (11,8%)	17
FA	1 (11,1%)	4 (44,4%)	4 (44,4%)	9
C	2 (14,3%)	4 (22,6%)	8 (63,4%)	14

Tabela 4.7 – Validação cruzada do método classificatório.

A sensibilidade, especificidade, valor preditivo positivo e negativo também foram obtidos com base nas classificações obtidas. O resultado está na tabela 4.8.

Resultado do Diagnóstico	<i>S</i> (%)	<i>E</i> (%)	<i>VPP</i> (%)	<i>VPN</i> (%)	κ
Negativo = TN	86,9	76,4	83,3	81,2	0,64
Positivo = FA ou C					
Negativo = FA	66,7	50,0	66,7	50,0	0,17
Positivo = C					

Tabela 4.8 – Validação e reprodutibilidade do modelo de diagnóstico.

Observa-se na tabela 4.8 que as componentes apresentam uma boa sensibilidade na detecção de anomalias, com $S=86,9\%$, embora a confiabilidade num resultado negativo não seja tão alta (76,4%). Quando se tenta diagnosticar a anomalia como benigna ou maligna, a qualidade do teste diminui bastante.

O nível de concordância κ entre o resultado do teste diagnóstico e o resultado previamente conhecido apresenta valores apenas regulares na detecção de anomalias ($\kappa=0,64$) e o diagnóstico da anomalia detectada apresenta valores bastante baixos ($\kappa=0,17$).

Método 2: Análise das Componentes da Matriz *R*

Obtenção das Componentes Principais

As componentes principais também foram calculadas através da matriz de correlação *R* total, formada pelos grupos TN, FA e C. A figura 4.15 mostra a parcela de variância correspondente a cada autovalor de *R*.

Figura 4.15 - (a) Porcentagem de variância p_k correspondente à cada autovalor. (b) Porcentagem de variância acumulada.

Como no caso dos autovalores da matriz de covariância S , os três primeiros autovalores da matriz R explicam quase toda a variância dos dados, cerca de 99%.

A figura 4.16 mostra o comportamento das 9 primeiras componentes principais.

Figura 4.16 - Componentes Principais obtidas pela matriz R total dos grupos TN, FA e C.

Nota-se que em geral, as componentes principais da matriz R são diferentes daquelas obtidas da matriz de covariância S , embora em algumas pareçam existir comportamentos semelhantes. Mas em ambos os casos as componentes obtidas evidenciam principalmente

características relativas aos picos de espalhamento. Sendo que as primeiras seis componentes são as que parecem apresentar algum tipo de informação a respeito da estrutura dos dados.

Obtiveram-se as projeções destas componentes no conjunto de dados, e um teste de hipótese foi aplicado a estas projeções, com o objetivo de identificar quais componentes são capazes de diferenciar os grupos. Foram consideradas para análise as oito primeiras componentes, responsáveis por 99,9% da variância total dos dados.

Comparação entre os grupos

Para comparar os grupos TN, FA e C nas oito primeiras componentes utilizou-se o teste de Kruskal-Wallis. A tabela 4.9 mostra o resultado do teste.

Variável	χ^2	α
CP ₁	20,070	0,000
CP ₂	6,978	0,031
CP ₃	26,111	0,000
CP ₄	2,562	0,278
CP ₅	1,804	0,406
CP ₆	10,517	0,005
CP ₇	6,741	0,034
CP ₈	3,010	0,222

Tabela 4.9 - Teste de Kruskal-Wallis para as primeiras 8 componentes principais da matriz *R*.

Observa-se que apenas as componentes CP₄, CP₅ e CP₈, não identificaram diferenças entre os três grupos. A comparação múltipla foi realizada para identificar para cada componente quais grupos diferem entre si, mostrando que o resultado significativo no teste de Kruskal-Wallis era devido à diferença entre o grupo TN e os demais, não havendo diferenças significativas entre FA e C.

Para tentar identificar quais componentes melhor separam os grupos FA e C, realizou-se um teste de Mann-Whitney. O resultado mostrou que a componente CP₅ é a que apresenta melhor resultado na diferenciação entre eles, com $\alpha=0,277$.

Modelo de Diagnóstico

Para classificar as amostras optou-se por utilizar a componente CP_1 , que diferencia normais e alterados, e a CP_5 , que melhor diferencia entre benignos e malignos. A figura 4.17 mostra o gráfico de $CP_1 \times CP_5$.

Figura 4.17 - Posição das amostras no espaço $CP_1 \times CP_5$.

Pode-se notar que a separação entre os grupos não é muito boa, exceto para o grupo dos normais. A quantidade de elementos classificados em cada grupo está na tabela 4.10.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	15 (88,2%)	2 (11,8%)	0 (0%)	17
FA	2 (22,2%)	5 (55,6%)	2 (22,2%)	9
C	0 (0%)	5 (35,7%)	9 (64,3%)	14

Tabela 4.10 - Classificação das amostras utilizando CP_1 e CP_5 .

Do total, 11 amostras não foram corretamente classificadas, correspondendo a um índice geral de 72,5% de acerto.

Aplicou-se o processo de validação cruzada para testar a capacidade destas componentes em classificar corretamente novas amostras de tecido. O resultado final mostrou que 67,5% foram corretamente classificadas, conforme mostra a tabela 4.11.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	15 (88,2%)	2 (11,8%)	0 (0%)	17
FA	2 (22,2%)	5 (55,6%)	2 (22,2%)	9
C	0 (0%)	7 (50,0%)	7 (50,0%)	14

Tabela 4.11 - Validação cruzada do método classificatório.

Quando comparados os resultados fornecidos pelas componentes obtidas das matrizes *S* e *R*, conclui-se que ambas fornecem resultados equivalentes, com taxas de acerto não muito diferentes entre si.

A tabela 4.12 mostra os parâmetros indicativos da eficiência do modelo classificatório.

Resultado do Diagnóstico	<i>S</i> (%)	<i>E</i> (%)	<i>VPP</i> (%)	<i>VPN</i> (%)	κ
Negativo = TN Positivo = FA ou C	91,3	82,3	87,5	87,5	0,74
Negativo = FA Positivo = C	50,0	42,8	63,6	30,0	0,06

Tabela 4.12 - Validação e reprodutibilidade do modelo de diagnóstico.

A detecção de alterações é muito boa, ocorrendo em 91,3% dos casos, e a confiabilidade num resultado negativo também apresenta um resultado satisfatório (82,3%). Entretanto, a identificação do tipo de alteração apresenta-se com um percentual bastante baixo, com os tecidos malignos sendo corretamente diagnosticados em apenas 50% dos casos em que realmente ocorrem.

A comparação destes resultados através do índice de concordância κ também leva às mesmas conclusões. A concordância quanto à existência de uma anomalia apresenta-se satisfatória ($\kappa=0,74$), enquanto a concordância quanto ao diagnóstico de malignidade da alteração detectada é bastante próximo de zero, indicando que as duas componentes são incapazes de diferenciar alterações benignas e malignas.

4.3.3 - Modelo de Diagnóstico utilizando Análise de Discriminante

A análise de discriminante foi realizada utilizando dois métodos, o direto e o *stepwise*. As variáveis utilizadas para o cálculo das funções discriminantes foram as mesmas selecionadas para obtenção das componentes principais, compreendendo apenas os ângulos no intervalo angular de 15° a 60°.

Método 1: Método Direto

Obtenção das Funções Discriminantes

O primeiro passo para a obtenção das funções discriminantes pelo método direto foi selecionar, pelo teste de tolerância, as variáveis que entrariam no cálculo destas funções. Este teste verificou a existência de multicolinearidade no conjunto de variáveis, e das 136 variáveis originais, apenas um grupo de 6 foram aceitas considerando o nível mínimo de tolerância de 0,01, e estão representadas na tabela 4.13.

Variável	Ângulo de Espalhamento
θ_1	15,00
θ_2	15,33
θ_3	16,33
θ_4	23,33
θ_5	34,33
θ_6	37,00

Tabela 4.13 – Variáveis utilizadas no cálculo de Z_1 e Z_2 no método direto.

Utilizando estas variáveis, foram obtidos os autovalores positivos da matriz $W^{-1}B$. A tabela 4.14 mostra os autovalores λ , ordenados em ordem decrescente de valor absoluto, bem como a parcela de variância correspondente a cada um e a porcentagem de variância acumulada pela soma dos dois. O número de autovalores é igual à $\min(g-1, p) = \min(2, 150) = 2$.

λ	% de Variância	% de Variância Acumulada
$\lambda_1 = 4,18$	86,3	86,3
$\lambda_2 = 0,31$	13,7	100,0

Tabela 4.14 – Autovalores da matriz $W^{-1}B$.

Assim como na técnica de componentes principais, o primeiro autovalor explica grande parte da variância total dos dados, e os dois juntos correspondem à variância total. Como há dois autovalores, são utilizadas duas funções discriminantes para separar os grupos TN, FA e C.

Obtiveram-se então os dois autovetores correspondentes a cada autovalor. O primeiro fornece os coeficientes da combinação linear para obtenção da primeira função discriminante Z_1 , e o segundo fornece os coeficientes para Z_2 . A tabela 4.15 mostra os valores desses coeficientes.

Variável	Coeficientes da Matriz <i>a</i>	
θ_1	-0,013	0,005
θ_2	0,015	-0,013
θ_3	-0,001	0,008
θ_4	-0,002	-0,001
θ_5	0,001	-0,009
θ_6	0,000	0,011
<i>c</i>	-2,623	-4,504

Tabela 4.15 – Elementos da matriz de coeficientes *a*.

Na tabela, *c* é uma constante. As funções discriminantes adquirem então a forma:

$$Z_1 = -2,623 + 0,013\theta_1 - 0,015\theta_2 + \dots - 0,000\theta_6$$

$$Z_2 = -4,504 + 0,05\theta_1 - 0,013\theta_2 + \dots + 0,011\theta_6$$

Foram então calculados os valores de Z_1 e Z_2 para todos os elementos pertencentes aos grupos, pela projeção das variáveis originais na matriz de coeficientes.

Comparação entre os grupos

Para avaliar o grau de diferenciação entre os grupos, realizou-se um teste de hipótese comparando-os. Como a função Z_2 não tinha distribuição normal em um dos grupos, optou-se por utilizar o teste de Kruskal-Wallis para comparar entre TN, FA e C. A tabela 4.16 mostra o resultado dos testes.

Variável	χ^2	gl	α
Z_1	28,573	2	0,000
Z_2	15,508	2	0,000

Tabela 4.16 – Teste de Kruskal-Wallis entre TN, FA e C nas variáveis Z_1 e Z_2 .

O resultado dos testes de comparação entre os três grupos mostrou-se significativo para as duas funções. Para analisar quais grupos são diferentes entre si realizou-se um teste de comparação múltipla. O resultado está na figura 4.18.

Figura 4.18 – Comparação múltipla entre os grupos nas funções discriminantes Z_1 e Z_2 .

O resultado mostra que a função Z_1 diferencia os tecidos normais dos alterados, mas não os benignos dos malignos. A função Z_2 , por outro lado, é capaz de diferenciar entre benignos e malignos, apesar de não diferenciar este último dos normais. Espera-se então que as duas funções juntas sejam capazes de diferenciar os três grupos. Um teste Λ de Wilks foi então aplicado à distribuição bivariada composta por Z_1 e Z_2 , com o objetivo de averiguar se os grupos eram diferentes. O resultado está na tabela 4.17.

Grupos	Λ	χ^2	gl = $p(g-1)$	α
TN, FA e C	0,1431	70,9733	4	0,000
FA e C	0,5749	11,0718	2	0,004

Tabela 4.17 – Teste Λ de Wilks entre os grupos na distribuição bivariada formada por Z_1 e Z_2 .

Os baixos níveis de significância do teste mostram que as duas funções são capazes de identificar diferenças em cada um dos três grupos.

Modelo de Diagnóstico

A figura 4.19 mostra as coordenadas de cada amostra no plano formado pelas duas funções discriminantes.

Figura 4.19 – Projeções as amostras no espaço $Z_1 \times Z_2$.

Da figura 4.19 é possível visualizar o resultado do teste de comparação múltipla realizado. Analisando os valores do eixo horizontal Z_1 , nota-se que o centróide do grupo TN está bem distante dos grupos FA e C, sendo que estes dois últimos estão bem próximos entre si. No eixo vertical Z_2 nota-se que a maior separação ocorreu entre os grupos FA e C.

Adotando o critério da mínima distância de Mahalanobis aos centróides como classificador das amostras, obtêm-se os resultados da tabela 4.18.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	16 (94,1%)	1 (5,9%)	0 (0%)	17
FA	0 (0%)	9 (100%)	0 (0%)	9
C	0 (0%)	3 (21,4%)	11 (78,6%)	14

Tabela 4.18 - Classificação das amostras utilizando Z_1 e Z_2 .

O grupo TN, dos tecidos normais, teve apenas um de seus elementos classificados erroneamente como fibroadenoma. O grupo FA, dos tecidos benignos, teve todos os seus elementos classificados corretamente. E o grupo C, dos tecidos malignos, teve três de seus elementos classificados como benignos.

Portanto, do total de 40 elementos, obteve-se um percentual geral de acerto de 90%.

Para testar a capacidade destas funções em classificar novos elementos foi utilizado o processo de validação cruzada. Neste processo, um elemento aleatório é retirado de seu grupo, e as funções discriminantes são calculadas novamente para os restantes. Classifica-se então este elemento retirado utilizando as novas funções. Em seguida este elemento volta para o grupo, e então outro é retirado e as funções discriminantes novamente calculadas, e assim por diante, até que todos sejam testados.

A tabela 4.19 mostra o resultado obtido ao fim deste processo.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	14 (82,4%)	2 (11,8%)	1 (5,9%)	17
FA	1 (11,1%)	5 (55,6%)	3 (33,3%)	9
C	0 (0%)	3 (21,4%)	11 (78,6%)	14

Tabela 4.19 – Validação cruzada do método classificatório.

O resultado da validação cruzada mostrou que o percentual geral de acerto foi de 75%.

O grupo FA teve apenas 5 de seus elementos classificados corretamente. O grupo C novamente apresentou três de seus elementos classificados como benignos.

Um dos fatores que influenciou bastante no resultado foi o número reduzido de elementos em cada grupo. Hair *et al* (HAIR *ET AL*, 1998) afirma que isto pode causar a instabilidade das funções discriminantes, ou seja, a cada vez em que se retira um elemento de um grupo para sua validação cruzada, os coeficientes das funções variam muito, adquirindo valores bem diferentes daqueles obtidos para todos os elementos juntos. Isto ficou comprovado quando se aplicou o processo de validação cruzada, em que a retirada de um elemento de um grupo alterava os valores das funções discriminantes para os elementos restantes.

Outro fator que também pode ter influenciado foi o método utilizado para selecionar as variáveis. O teste de tolerância é útil para detectar e remover multicolinearidade, o que influencia positivamente no resultado final, entretanto, muitas variáveis que talvez fossem importantes para a separação entre os grupos podem ter sido rejeitadas pelo teste e excluídas

do cálculo das funções discriminantes. Uma alternativa seria então selecionar as variáveis através de algum critério, o que foi feito posteriormente com a aplicação do método *stepwise*.

A eficiência das funções discriminantes como modelo de diagnóstico foi avaliada através dos parâmetros S , E , VPP , VPN e κ , dispostos na tabela 4.20.

Resultado do Diagnóstico	S (%)	E (%)	VPP (%)	VPN (%)	κ
Negativo = TN Positivo = FA ou C	95,6	82,3	88,0	93,3	0,79
Negativo = FA Positivo = C	78,6	62,5	78,6	62,5	0,41

Tabela 4.20 - Validação e reprodutibilidade do modelo de diagnóstico.

Pode-se notar que as funções discriminantes Z_1 e Z_2 são capazes de detectar alguma alteração no tecido mamário em quase todos os casos, como mostra o valor de $S=95,6\%$, e em $82,3\%$ dos casos o resultado negativo correspondia realmente à ausência de alguma anomalia na mama.

Alem disso, Z_1 e Z_2 são eficientes no diagnóstico de anomalias do tecido, identificando em $78,6\%$ dos casos a presença de neoplasias malignas.

Método 2: Método *Stepwise*

O método *stepwise* para seleção de variáveis é uma alternativa ao método direto, e consiste em determinar quais variáveis entram no cálculo das funções discriminantes em etapas consecutivas, tendo como base algum critério de seleção.

O critério de seleção de variáveis utilizado foi a maximização da medida da distância D^2 de Mahalanobis entre os grupos. Este critério é particularmente útil porque permite utilizar somente variáveis que tenham relevância na separação entre os grupos.

Obtenção das Funções Discriminantes

Foram necessárias 9 etapas para selecionar um conjunto de 7 variáveis. A cada etapa, variáveis eram adicionadas ou removidas, dependendo de sua capacidade em separar os centróides de cada grupo. A tabela 4.21 mostra as variáveis selecionadas no fim deste processo.

Variável	Ângulo de Espalhamento
θ_1	23,00
θ_2	24,66
θ_3	33,00
θ_4	38,00
θ_5	43,66
θ_6	48,66
θ_7	56,00

Tabela 4.21 – Variáveis selecionadas para o cálculo das funções discriminantes.

A partir destas variáveis obtiveram-se os autovalores λ da matriz $W^{-1}B$. A tabela 4.22 mostra os resultados obtidos.

λ	% de Variância	% de Variância Acumulada
$\lambda_1 = 9,04$	81,5	81,5
$\lambda_2 = 1,32$	18,5	100,0

Tabela 4.22 - Autovalores da matriz $W^{-1}B$.

Os autovalores fornecem os autovetores de $W^{-1}B$, que por sua vez fornecem os coeficientes para as combinações lineares Z_1 e Z_2 . A tabela 4.23 mostra os coeficientes obtidos.

Variável	Matriz a	
θ_1	0,003	0,004
θ_2	-0,006	-0,007
θ_3	0,006	0,004
θ_4	0,005	0,011
θ_5	-0,006	-0,015
θ_6	-0,017	-0,011
θ_7	0,016	0,019
c	-0,614	-1,434

Tabela 4.23 – Coeficientes da matriz a .

As coordenadas de Z_1 e Z_2 para todos os elementos de cada grupo foram então obtidas pela projeção das variáveis θ_i , $i=1,\dots,7$ na matriz de coeficientes a .

Comparação entre os grupos

Para verificar se as funções obtidas são capazes de identificar diferenças entre os grupos, aplicou-se o teste de Kruskal-Wallis para cada variável independentemente. A tabela 4.24 mostra os resultados obtidos neste teste.

Variável	χ^2	gl	α
Z_1	32,814	2	0,000
Z_2	20,659	2	0,000

Tabela 4.24 – Teste de Kruskal-Wallis entre TN, FA e C.

Os testes aplicados mostram que existe ao menos um grupo muito diferente dos demais. Para descobrir quais grupos diferem entre si realizou-se uma comparação múltipla. O resultado está na figura 4.20.

Figura 4.20 – Comparação múltipla entre os três grupos.

A comparação múltipla entre os grupos mostrou que Z_1 é capaz de diferenciar os três tipos de tecidos, ao passo que a principal contribuição de Z_2 é diferenciar entre tecidos benignos e malignos. Este resultado é mais expressivo do que o obtido através do método direto, onde apenas uma função permitia a distinção entre benignos e malignos.

Um teste Λ de Wilks foi então aplicado à distribuição bivariada formada por Z_1 e Z_2 , mostrando que os grupos são significativamente diferentes. O resultado está na tabela 4.25.

Grupos	Λ	χ^2	$gl = p(g-1)$	α
TN, FA e C	0.0547	106.0484	4	0,000
FA e C	0.2157	30.6740	2	0,000

Tabela 4.25 – Teste Λ de Wilks para a distribuição bivariada formada por Z_1 e Z_2 .

Modelo de Diagnóstico

A figura 4.21 mostra as coordenadas de todos os elementos no plano $Z_1 \times Z_2$, indicando a localização dos centróides $\bar{Z}_1$ e $\bar{Z}_2$ para cada grupo.

Figura 4.21 – Coordenadas das amostras no espaço $Z_1 \times Z_2$.

Nota-se que os grupos estão bem separados uns dos outros. Os grupos TN e C são os que apresentam melhor resultado, com todos os seus elementos distribuídos homogeneamente ao redor de seus centróides, e não muito distantes deles. Os elementos do grupo FA apresentam-se mais dispersos em relação ao centróide do grupo.

A tabela 4.26 mostra o resultado da comparação entre classificação dos elementos a partir de suas distâncias aos centróides.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	17 (100%)	0 (0%)	0 (0%)	17
FA	1 (11,1%)	8 (88,9%)	0 (0%)	9
C	0 (0%)	0 (0%)	14 (100%)	14

Tabela 4.26 - Classificação das amostras utilizando Z_1 e Z_2 .

A tabela mostra que apenas um elemento do grupo FA foi classificado erroneamente como normal. Comparando este resultado com a tabela de classificação obtida no método direto, nota-se que houve uma melhora expressiva no índice percentual de acerto, que neste caso foi de 97,5%.

Para testar a eficiência das funções discriminantes como critério de classificação de novos elementos, foi realizado o processo de validação cruzada. A tabela 4.27 mostra o resultado da classificação dos elementos de cada grupo.

Grupo	Nº elementos classificados			Total
	TN	FA	C	
TN	15 (88,2%)	2 (11,8%)	0 (0%)	17
FA	2 (22,2%)	5 (55,6%)	2 (22,2%)	9
C	1 (7,1%)	0 (0%)	13 (92,9%)	14

Tabela 4.27 - Validação cruzada do método classificatório.

O processo de validação cruzada mostrou que das 40 amostras apenas 7 não foram classificadas corretamente, correspondendo a um índice de acerto geral de 82,5%.

O grupo TN apresentou dois de seus elementos classificados como FA, e um caso de falso negativo, com um elemento do grupo C classificado como normal.

O grupo FA, por outro lado, teve praticamente metade de seus elementos classificados de forma errada. Um destes elementos classificados como TN possuía o pico de espalhamento adiposo em 20°, que é característico dos tecidos normais, além do pico característico dos tecidos alterados. Esta característica provavelmente foi a responsável por sua localização próxima aos tecidos normais.

Apesar do resultado do grupo FA, pode-se concluir que os resultados obtidos neste método foram em geral bem superiores ao método direto, uma vez que a característica do critério de seleção de variáveis foi escolher aquelas que maximizavam a distância D^2 de Mahalanobis entre os centróides de cada grupo, otimizando o processo de classificação de novos elementos.

Apesar disso, ainda existe o problema da instabilidade das funções discriminantes, devido ao pequeno número de amostras de tecido analisadas, principalmente no grupo FA, que foi o que apresentou maior número de classificações errôneas.

Os parâmetros que avaliam a eficiência e confiabilidade do resultado destas funções como modelo de diagnóstico estão na tabela 4.28.

Resultado do Diagnóstico	<i>S</i> (%)	<i>E</i> (%)	<i>VPP</i> (%)	<i>VPN</i> (%)	κ
Negativo = TN Positivo = FA ou C	86,9	88,2	90,9	83,3	0,74
Negativo = FA Positivo = C	100,0	71,4	86,7	100,0	0,76

Tabela 4.28 - Validação e reprodutibilidade do modelo de diagnóstico.

A sensibilidade para detectar algum tipo de alteração foi um pouco menor do que a obtida no método anterior. Entretanto, o resultado mais expressivo da tabela acima certamente é a alta sensibilidade das funções discriminantes para diagnosticar alterações malignas.

O índice de concordância κ também apresentou níveis satisfatórios, principalmente no diagnóstico de lesões benignas e malignas.

Pode-se notar que as funções discriminantes obtidas no método *stepwise* foram as que melhor identificaram diferenças entre tecidos benignos e malignos. Além disso, mantiveram o nível de sensibilidade aceitável para a detecção de alterações, com o valor de *S* bem próximo da sensibilidade de um exame mamográfico.

CAPÍTULO 5

CONCLUSÕES E PERSPECTIVAS

5.1 - Conclusões

5.1.1 - Sobre o método de obtenção das seções de choque $d\sigma/d\Omega$ das amostras de tecido

O uso de difractômetros comerciais, como o Siemens D5005 utilizado neste trabalho, se mostrou adequado para determinar a distribuição angular da radiação espalhada pelas amostras de tecidos mamários, sendo possível identificar claramente os picos de espalhamento para cada tipo de tecido. Recentemente, Johns e Wismayer (JOHNS e WISMAYER, 2004) também mostraram a viabilidade destes tipos de difractômetros para se obter os perfis de espalhamento de materiais amorfos.

Os fatores de correção aplicados aos dados experimentais neste trabalho foram obtidos baseando-se em correções experimentais e cálculos numéricos, e sua aplicação permite determinar com precisão a seção de choque diferencial de espalhamento $d\sigma/d\Omega$. Para avaliar a adequação dos dados após a aplicação destes fatores, um teste foi realizado numa amostra de água, e o resultado comparado com a seção de choque $d\sigma/d\Omega$ obtida a partir do fator de forma tabulado por Morin (MORIN, 1982), resultado numa concordância aceitável dentro das incertezas experimentais. Além disso, foram comparadas também as seções de choque de tecidos normais e alterados com resultados apresentados em trabalhos anteriores (OLIVEIRA, 2006). Estas comparações permitiram concluir que a metodologia de correção dos dados é adequada para obtenção das seções de choque diferencial de espalhamento das amostras de tecido mamário.

5.1.2 - Sobre os modelos de diagnóstico desenvolvidos

O primeiro modelo construído baseou-se na definição de algumas variáveis a partir dos perfis de espalhamento, tais como posição e intensidade dos picos de espalhamento e largura à meia altura, entre outros (seção 4.3.1).

Algumas destas variáveis se mostraram muito eficientes na distinção entre tecidos normais e alterados, como P_1 e LMA. Esta distinção ocorre pelo fato de a maior parte dos tecidos normais analisados serem ricos em tecido adiposo, cujo perfil é bem característico, enquanto a maioria dos tecidos alterados (benignos e malignos) não apresenta este tipo de tecido em sua composição.

Apesar da alta sensibilidade deste método em classificar tecidos normais e alterados, ele não permite identificar diferenças entre alterações benignas e malignas, pelo fato de as curvas de espalhamento destes dois tipos de tecidos serem muito semelhantes entre si. Além disso, o processo de obtenção destas variáveis no perfil de cada amostra deve ser feito manualmente, o que o torna lento e sujeito a erros sistemáticos.

O segundo método testado para classificar os tipos de tecidos analisados consistiu na análise de um sub-intervalo angular do perfil de espalhamento das amostras, de 15° a 60° , através da análise de componentes principais.

Estas componentes foram obtidas a partir das matrizes de covariância S e correlação R , e em ambos os casos, a análise de componentes principais mostrou-se uma técnica de redução de variáveis altamente eficiente, sendo capaz de condensar praticamente toda a variabilidade dos dados, constituídos inicialmente por 136 variáveis, em algumas poucas componentes (seção 4.3.2).

No que se refere à capacidade destas componentes em identificar diferenças entre tecidos normais, benignos e malignos, observou-se que elas oferecem uma boa sensibilidade e especificidade na detecção de alterações, ou seja, na distinção entre tecidos normais e alterados. Uma vez que os tecidos normais apresentam um perfil de espalhamento bastante diferente dos alterados, era esperado que as primeiras componentes, que correspondem à cerca de 99% da variância total dos dados, permitissem visualizar diferenças entre estes tecidos. No que se refere à caracterização de tecidos alterados, assim como no modelo anterior, o uso das componentes principais não se mostrou um método eficiente.

O terceiro método de classificação consistiu na análise de discriminante, utilizando como variáveis os ângulos contidos no sub-intervalo de 15° a 60° (seção 4.3.3). As funções discriminantes foram obtidas de duas formas, através do método direto, em que todas as variáveis aceitas no teste de tolerância entravam no cálculo das funções, e o método *stepwise*, que selecionava as variáveis que mais contribuíam para a separação entre os grupos.

O método direto permitiu a obtenção de duas funções discriminantes, que juntas eram capazes de distinguir entre os três tipos de tecidos, mostrando ser possível distinguir tecidos normais de alterados e, além disso, caracterizar as alterações como benignas ou malignas, o que não foi possível de ser alcançado nos dois modelos anteriores.

O método *stepwise* otimizou o resultado obtido no método direto, selecionando para o cálculo das funções discriminantes apenas as variáveis que contribuíam para maximizar a distância D^2 de Mahalanobis entre os centróides de cada grupo. Como consequência, cada uma das funções discriminantes era capaz de distinguir entre neoplasias benignas e malignas, além de distinguir também os tecidos normais.

Portanto a análise de discriminante mostrou-se a técnica mais indicada para classificar os tecidos, permitindo detectar alterações com uma sensibilidade equivalente à de um exame mamográfico, e diagnosticá-las também de forma altamente eficiente, utilizando para isso um número reduzido de variáveis angulares.

5.1.3 – Conclusões Finais

Este trabalho mostra que é possível utilizar os perfis de espalhamento dos tecidos mamários como base para a detecção e o diagnóstico de alterações.

A técnica de análise de discriminante se mostrou muito adequada para este fim, sendo bastante sensível em detectar a presença de uma alteração, assim como em distingui-la como benigna ou maligna. Embora o número de amostras analisadas ainda seja insuficiente para consolidar este método de reconhecimento de padrões como modelo de diagnóstico, os resultados obtidos apontam para a viabilidade de sua aplicação como ferramenta no auxílio ao diagnóstico do câncer de mama.

5.2 – Perspectivas

Com os resultados obtidos neste trabalho, podem-se definir as seguintes perspectivas para trabalhos futuros:

- Deve-se aumentar o número de amostras de tecidos mamários analisadas. Este é um quesito fundamental para validar os resultados de classificação histológica baseada nos perfis de espalhamento obtidos neste trabalho. Além disso, uma grande quantidade destes perfis constitui-se num banco de dados para classificação de novas amostras no futuro.
 - Caracterização de neoplasias malignas. Uma vez que os carcinomas de mama são constituídos por alterações de vários tipos, como por exemplo, carcinomas ductais de grau I, II e III, intraductais e lobulares, pode-se estender o método de análise de discriminante, de forma a incluir a classificação de neoplasias malignas.
 - Automatizar o processo de classificação histológica, implementando a metodologia de pré-processamento e de diagnóstico ao processo de obtenção dos perfis de espalhamento.
-

REFERÊNCIAS BIBLIOGRÁFICAS

- BARNES, G. T. Contrast and Scatter in X-Ray Imaging. **Radiographics**. v. 11, n. 2, p.307-323, 1991.
- BLOOM, H. J. G., RICHARDSON, W. W. Histological Grading and Prognosis in Breast Cancer. **British Journal of Cancer**. v. 11, n. 3, p. 359-377, 1957.
- CALLEGARI-JONES, S. M. **Bioestatística: Princípios e Aplicações**. São Paulo: Artmed Editora S. A., 2003.
- CHAN, H. P., DOI, K. Energy and Angular Dependence of X-Ray Absorption and its Effect on Radiographic Response in Screen-Film System. **Physics in Medicine and Biology**. v. 28, n. 5, p. 565-579, 1983.
- CONOVER, W. J., **Practical Nonparametric Statistics**. 2 ed. New York: John Wiley & Sons, Inc, 1980.
- DUNN, O. J. Multiple contrasts using rank sums. **Technometrics**. v. 6 n 3, p. 241-252, 1964.
- EVANS, S. H., BRADLEY, D., DANCE, D.R., BATERMAN, J. E, JONES, C. H. Measurements of Small-Angle Photon Scattering for some Breast Tissues and Tissue Substitute Materials. **Physics in Medicine and Biology**. v. 36, n. 1, p. 7-18, 1991.
- GIBBONS, J. D, **Nonparametric Statistical Inference**, New York: McGraw-Hill, Inc., 1971.
- HAIR, J. F., ANDERSON, R. E., TATHAM, R. L., BLACK, W. C. **Multivariate Data Analysis**. 5ª Ed. New Jersey: Prentice Hall, 1998.
- HARDING, G., KOSANETSKY, J., NEITZEL, U. X-Ray Diffraction Computed Tomography. **Medical Physics**, v. 14, n. 4, p. 515-525, 1987.
- HUBBEL, J. H., VEIGELE, E. A., BRIGGS, E. A. BROWN, D. T. CROMER, D. T., HOWERTON, R. J. Atomic Form Factors, Incoherent Scattering Functions and Photon Scattering Cross Sections. **Journal of Physical and Chemical Reference Data**. v. 4, n. 3, p. 471-538, 1975.
- INSTITUTO NACIONAL DE CÂNCER, Ministério da Saúde. **Estimativa 2006: Incidência de Câncer no Brasil**. Rio de Janeiro: INCA, 2005.
- JAIN, A. K. *et al.* Statistical Pattern Recognition: A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. v. 22, n. 1, p. 4-37, 2000.
- JAMES, R. W., **The Optical Principle of the Diffraction of X-Ray**. London: Bell, 1962.
-

-
- JOBSON, J. D. **Applied Multivariate Data Analysis**. New York: Springer-Verlag New York, Inc., 1991. Vol. 1.
- JOHNS, H. E., CUNNINGHAM, J. R. **The Physics of Radiology**. Springfield: Charles C. Thomas, 1983.
- JOHNS, P.C., YAFFE, M.J., Coherent Scatter in Diagnostic Radiology, **Medical Physics**, v. 10, n. 40, p. 40-50, 1983.
- JOHNS, P. C., WISMAYER, M. P. Measurement of Coherent X-Ray Scatter Form Factors for Amorphous Materials using Diffractometers. **Physics in Medicine and Biology**. v. 49, n. 23, p. 5233-5250, 2004.
- JOHNSON, R.A, WICHERN, D.W., **Applied Multivariate Statistical Analysis**. 5^a Ed. New Jersey: Prentice Hall, 2002.
- KANE, P. P., KISSEL, L., PRATT, R. H., ROY, S. C. Elastic Scattering of X-Rays by Atoms. **Physics Reports**. v. 140, n. 2, p. 75-159, 1986.
- KIDANE, G., SPELLER, R.D., ROYLE, G.J., HANBTY, A.M., X-ray scatter signatures for normal and neoplastic beast tissues. **Physics in Medicine and Biology**, v. 44, n. 7, p.1791-1802, 1999.
- KRZANOWSKI, W. J., Selection of Variables to preserve Multivariate Data Structure using Principal Components. **Applied Statistics**. v. 36, n 1, p.22-33, 1987.
- LEWIS, R. A., ROGERS, K. D., HALL, C. J., TOWNS-ANDREWS, E., SLAWSON, S., EVANS, A., PINDER, S. E., ELLIS, I. O., BOGGIS, C. R. M., HUFTON A. P., DANCE, D. R. Breast Cancer Diagnosis Using Scattered X-Rays. **Journal of Synchrotron Radiation**. v. 7, n. 5, p. 348-352, 2000.
- LILLIEFORS, H. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknow. **Journal of the American Statistical Association**. v. 62, n. 318, p. 399-402, 1967.
- MACMAHON, B., TRICHOPOULOS, D. **Epidemiology – Principles & Methods**. 2^a ed. New York: Little, Brown and Company, 1996.
- MCPHERSON K., STEEL C.M. e DIXON, J. M. ABC of Breast Disease: Breast Cancer-Epidemiology, Risk Factors, and Genetics. **British Medical Journal**, v. 321, n. 7261, p. 624-628, 2000.
- MORIN, L.R.M. Molecular Form Factors and Photon Coherent Scattering Cross Sections of Water. **Journal of Physical and Chemical Reference Data**. v. 11, n. 4, p. 1091-1098, 1982.
- NARTEN, A. H., LEVY. H. A. Liquid Water: Molecular Correlation Functions from W-Ray Diffraction. **The Journal of Chemical Physics**. v. 55, n. 5, p. 2263-2269, 1971.
- OLIVEIRA, O. R. **Determinação Experimental dos Perfis de Espalhamento de Tecidos Mamários (Normais e Alterados): Uma Nova Fonte de Informação Histológica na**
-

Mamografia. Dissertação (Mestrado). Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto. Ribeirão Preto, 2006.

PEREIRA, M. G. **Epidemiologia – Teoria e Prática.** Rio de Janeiro: Guanabara Koogan S.A., 2005.

POLETTI M.E. **Estudo do Espalhamento de Fótons (6,94, 17,44 e 59,54 keV) em Tecidos Mamários, Materiais Equivalentes e sua Influência em Mamografia.** Tese (Doutorado). Universidade Federal do Rio de Janeiro, Instituto de Física. Rio de Janeiro, 2001.

POLETTI, M.E., GONÇALVES, O.D., MAZZARO, I., X-Ray Scattering from Human Breast Tissues and Breast-Equivalent Materials. **Physics in Medicine and Biology**. v. 47, n. 1, p. 47-64, 2002a.

POLETTI, M.E., GONÇALVES, O.D., MAZZARO, I., Coherent and incoherent-scattering of 17.44 and 6.93 keV X-ray photons scattered from biological and biological-equivalent samples: characterization of tissues. **X-ray Spectrometry**. 31: 57, 2002b.

POLETTI, M. E., GONÇALVES, O. D., SCHECHTER, H., MAZZARO, I. Precise Evaluation of Elastic Differential Scattering Cross-Sections and their Uncertainties in X-Ray Scattering Experiments. **Nuclear Instruments and Methods in Physics Research B**. v. 187, n. 4, p. 437-446, 2002.

PRATT, J. W., GIBBONS, J. D. **Concepts of Nonparametric Theory.** New York: Springer-Verlag New York Inc., 1981.

RENCHER, A. C., Interpretation of Canonical Discriminant Functions, Canonical Variates and Principal Components. **American Statistical Association**. v. 46, n 3, p. 217-225, 1992.

ROHATGI, V. K. **Statistical Inference.** New York: John Wiley & Sons, Inc., 1984.

SEIDMAN H. e MUSHINSKI M. H. Breast cancer incidence, mortality, survival and prognosis in Breast Carcinoma: Current Diagnosis and Treatment, editado por Feig S.A. e McLelland R. **American College of Radiology and Masson**, New York, 1983.

SPELLER, R.D. Tissue Analysis Using X-Ray Scattering. **X-Ray Spectrometry**, v. 28 n. 4, p. 244-250, 1999.

SPELLER, R. D., HORROCKS, J. A. Photon Scattering: A “New” Source of Information in Medicine and Biology?. **Physics in Medicine and Biology**. v. 36, n. 1, p. 1-6, 1991.

SPIEGEL, M. R. **Probabilidade e Estatística.** São Paulo: Pearson Education do Brasil Ltda. 2004.

TABAR L. *ET AL.* Reduction in mortality from breast cancer after mass screening with mammography. Randomized trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. **Lancet**, v. 325, n. 8433, p. 829-832, 1985.

TOU, J. T., GONZALEZ, R. C. **Pattern Recognition Principles.** 4ª ed. London: Addison Wesley Publishing Company, 1974.

ZAR, J. H. **Biostatistical Analysis**. 4 ed. New Jersey: Prentice Hall, 1999.