UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE QUÍMICA

FABIANA ALVES DE LIMA RIBEIRO

ANÁLISE DE IMAGENS NÍVEL DE CINZA UTILIZANDO MÉTODOS QUIMIOMÉTRICOS

TESE DE DOUTORADO

Orientadora: Professora Doutora Márcia Miguel Castro Ferreira

CAMPINAS - SÃO PAULO

AGOSTO - 2007

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE QUÍMICA DA UNICAMP

R354a

Ribeiro, Fabiana Alves de Lima.
Análise de imagens nível de cinza utilizando métodos quimiométricos / Fabiana Alves de Lima Ribeiro. -- Campinas, SP: [s.n], 2007.

Orientadora: Márcia Miguel Castro Ferreira.

Tese - Universidade Estadual de Campinas, Instituto de Química.

1. Análise de imagens. 2. Quimiometria. 3. Fibras capilares. 4. Métodos *multi-way*. I. Ferreira, Márcia Miguel Castro. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Título em inglês: Analysis of gray-scale images by using chemometric methods

Palavras-chaves em inglês: Image analysis, Chemometrics, Hair, Multi-way methods

Área de concentração: Físico-Química

Titulação: Doutor em Ciências

Banca examinadora: Márcia Miguel Castro Ferreira (orientadora), Marcelo Martins de Sena (CET-UEG), Gabriela Castellano (IFGW-UNICAMP), Francisco Benedito Teixeira Pessine (IQ-UNICAMP), Ronei Jesus Poppi (IQ-UNICAMP)

Data de defesa: 15/08/2007

-Nos
-Fios
-Ten
Sos
-Da
-Pauta
-De me
Tal
-As
-An/
Do/
Ri/
Nhas
-Gri-
Tam
-Por
-Fal/
-Ta/
-De uma
-Clave
-De
-Sol

(João Ricardo - Cassiano Ricardo, As andorinhas, 1973)

Agradecimentos...

Agradeço sinceramente a todos que contribuiram para a realização deste trabalho.

Fabiana Alves de Lima Ribeiro

Formação acadêmica/Titulação

Doutorado em Química (Área de concentração: Físico-Química)
"Análise de imagens nível de cinza utilizando métodos quimiométricos"
Data da Defesa: 15/08/2007 - LQTA/IQ/UNICAMP
Orientadora: Profa. Dra. Márcia Miguel de Castro Ferreira

Mestrado em Química (Área de concentração: Físico-Química)
"Aplicação de métodos de análise multivariada no estudo de hidrocarbonetos policíclicos aromáticos"
Data da Defesa: 02/02/2001- LQTA/IQ/UNICAMP
Orientadora: Profa. Dra. Márcia Miguel de Castro Ferreira

Licenciatura em Química - IQ/UNICAMP (2003)

Graduação em Química Tecnólogica - IQ/UNICAMP (1997)

Atuação profissional

Instituto Internacional de Pesquisas Farmacêuticas (IIPF)
Pesquisadora/Quimiometrista
12/2005 – atual (42 h/semana)

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto/Universidade de São Paulo (FFCLRP/USP)
Professora Assistente
Cursos: Bacharelado em Química e Licenciatura em Química
Disciplinas minstradas: Físico-Química II (Cinética), Físico-Química Experimental.
03/2005 – 12/2005 (RTP: 12 h/semana)

Laboratório de Biorremediação e Biodegradação
Departamento de Microbiologia – ICB/USP
Pesquisadora
03/2001 – 12/2004

Programa de Estágio Docente – Nível I (docência plena) (IQ/UNICAMP)

QG101 – Química Geral
Curso: Engenharia Mecânica.
02/2003 - 07/2003 (4h/semana)

QF632 - Físico-Química Experimental I
Curso: Bacharelado em Química e em Química Tecnológica.
08/2003 - 12/2003 (8h/semana)

Extensão Universitária

Mini-Curso: Quimiometria - FFCLRP/USP
Minicurso ministrado durante a XXX Semana da Química.
03/2005 - 07/2005 (6 horas)

Cursinho Alternativo da Moradia Estudantil da UNICAMP
Coordenadora de área e professora.
Disciplina: Química.
02/1999 - 02/2000 (trabalho voluntário)

Projeto VEJA - Vivência Educacional para Jovens e Adultos
Professora em curso de ensino supletivo da Moradia Estudantil da UNICAMP.
Disciplina: Ciências.
02/1999 - 02/2000 (trabalho voluntário)

Projeto Universidade Solidária – UNISOL
01/1998 - 02/1998, São Gabriel da Cachoeira (AM).

Projeto Universidade Solidária – UNISOL
01/1996 - 02/1996, Anadia (AL).

Publicações: Periódicos Científicos

F. A. L. Ribeiro, S. Morano, L. R. Silva, R. P. Schneider, M. M. C. Ferreira, "Planilha de validação: uma nova ferramenta para estimar figuras de mérito na validação de métodos analíticos univariados", *Quim Nova*, no prelo, 2007.

F. A. L. Ribeiro & M. M. C. Ferreira, "QSAR model of the phototoxicity of polycyclic aromatic hydrocarbons", *J Mol Struct - THEOCHEM*, 719, 191-200, 2005.

F. A. L. Ribeiro & M. M. C. Ferreira, "Análise de Componentes Principais como Ferramenta para a Investigação de Contaminação Ambiental: Um Estudo de Caso", *Tecno-Lógica*, 9 (1) 35-57, 2005.

F. A. L. Ribeiro & M. M. C. Ferreira, "QSPR models of boiling point, octanol-water partition coefficient and retention time index of polycyclic aromatic hydrocarbons", *J Mol Struct*, 663, 109-126, 2003.

L. M. Aleixo; M. Sitton; F. A. L. Ribeiro, "Estudo Polarográfico sobre a Determinação de Fe(III) Utilizando-se a Técnica de Polarografia de Pulso Diferencial", *Quím Nova*, 24(6), 790-794, 2001.

R. C. R. Figueredo; F. A. L. Ribeiro; E. Sabadini. "Ciências de Espumas Aplicação na Extinção de Incêndios", *Quím Nova*, 22, 126-130, 1999.

Trabalhos apresentados em eventos

15 trabalhos apresentados em congressos e encontros científicos nacionais e internacionais.

RESUMO

Imagens são utilizadas na investigação científica há muito tempo, inicialmente apenas como ferramentas para a representação da prática científica, e após o surgimento das técnicas de microscopia, como instrumento para registro e análise instrumental. Com o aperfeiçoamento das técnicas de obtenção de imagens, cresceu a demanda por técnicas quantitativas e sistematizadas para extrair-lhes informações, e capazes de estabeler critérios estatisticamente confiáveis para, por exemplo, detectar similaridades, padrões, classificar, e até mesmo gerar modelos preditivos. Neste trabalho, métodos quimiométricos foram utilizados como ferramenta quantitativa para a análise de imagens nível de cinza, utilizando como exemplos imagens de microscopia de força atômica de fibras capilares de diferentes classes.

Fibras capilares são frequentemente utilizadas em clínicas, indústrias ambientais e análises forenses, para o diagnóstico de doenças, avaliação da exposição aos agentes tóxicos e poluentes e detecção de abuso de drogas. Para este trabalho foram utilizados dois conjuntos de dados: um deles formado por amostras de fibras caucasianas submetidas ao tratamento de descoloração e outro com fibras de diferentes etnias. O objetivo foi utilizar métodos quimiométricos para construir modelos classificatórios capazes de identificar corretamente novas imagens. Diversas estratégias foram testadas e os melhores resultados foram obtidos utilizando os métodos SIMCA (S*oft Independent Modeling of Class Analogy*), PARAFAC (*Parallel Factor Analysis*), MPCA (*Multi-way Principal Component Analysis*) e NPLS (*Multi-way Partial Least Squares*). Os modelos quantitativos apresentaram erros de calibração abaixo de 10% e erros de predição em torno de 10%. Com exceção de uma das aplicações, que é específica para fibras capilares, os outros métodos de análise de imagens propostos podem ser utilizados na análise quantitativa de qualquer tipo de imagem nível de cinza.

ABSTRACT

Images have been used in the scientific investigation for a long time, initially as a tool for representation of the scientific practice and, after the microscopy development, as an instrument for registration and instrumental analysis. With the improvement of microscopic techniques, there was an increase on the demand for quantitative and systematic tools to extract relevant information from images and for techniques capable to establish reliable statistical approaches, for example, to detect similarities, patterns and for classification. In this work, multivariate methods were used as quantitative tools for the analysis of nível de cinza images of different classes of hair fibers. These images were obtained by atomic force microscopy.

Hair fibers are frequently used in medical clinics, environmental industries and forensic analyses, for the diagnosis of diseases, evaluation of the exposure to toxic agents and pollutants, and detection of abuse of drugs. In this work two data sets were used: the first one contained caucasian hair fibers submitted to peroxide treatment and the second one contained fibers from different ethnic origin (oriental, african and caucasian). The goal was to use chemometric methods to build classification models capable to identify new images correctly. Several strategies were tested and the best results were obtained by using SIMCA (Soft Independent Modeling of Class Analogy), PARAFAC (Parallel Factor Analysis), MPCA (Multi-way Principal Component Analysis) and NPLS (Multi-way Partial Least Squares). The models presented calibration errors below 10% and prediction errors around 10 %. With exception of the descriptor analysis, which is specific for hair fibers images, the proposed methods can be useful for quantitative analysis of any kind of nível de cinza images.

ÍNDICE

LISTA DE TABELAS

LISTA DE FIGURAS

CAPÍTULO 1

Introdução

1.1 Imagens na investigação científica

Imagens são utilizadas na investigação científica há muito tempo, inicialmente apenas como ferramentas para representação da prática científica [Lynch & Woolgar, 1990] e, após o surgimento das técnicas de microscopia, como instrumento para registro e análise instrumental [Cooke, 2000; 1998; 1996; 1994; 1992; 1990; 1988; 1986].

O resultado do registro de uma imagem, seja analógico ou digital, é a representação das intensidades das cores como função de suas coordenadas espaciais *i*, *j*, e em alguns casos, *k*. A análise de imagens digitais existe desde a década de 60 e seus maiores avanços ocorreram nas áreas de sensoriamento remoto, geologia, agricultura, biologia e medicina [Geladi *et al.*, 1992a]. Na área de química, o uso de métodos para análise de imagens tornou-se mais popular após o aperfeiçoamento das técnicas de microscopia ótica [Cooke, 1992; Abramowitz, 2003], microscopia eletrônica [Bozzola & Russell, 1999; Watt, 1997], microscopia de força atômica (AFM[1]) [Alessandrini & Facci, 2005; Hamers, 1996] e espectroscopia de imagens [Lewis *et al.*, 2004; Reich, 2005; De Juan *et al.*, 2004; Burger & Geladi, 2006]. Outro fator que contribuiu muito para a difusão dos métodos de análise de imagens foi a

[1] *Atomic Force Microscopy*.

crescente e rápida sofisticação dos recursos computacionais, que passaram a permitir o gerenciamento de conjuntos de dados cada vez maiores.

Neste contexto, cresceu a demanda por técnicas quantitativas e sistematizadas para extrair informações de imagens, capazes de estabelecer critérios estatisticamente confiáveis para, por exemplo, detectar similaridades, padrões, classificar e até mesmo gerar modelos preditivos.

Vale ressaltar que o uso de técnicas quantitativas para a análise de imagens permite a extração de informações tanto de natureza qualitativa, como por exemplo, o reconhecimento de padrões e classificação, quanto de natureza quantitativa, como por exemplo, a estimativa de propriedades numéricas a partir das imagens. Dois exemplos interessantes da estimativa de parâmetros numéricos a partir de imagens estão descritos nos trabalhos de Gurden *et al.* (2004) e Huang *et al*. (2003). No primeiro caso, os autores estimaram descritores para fibras capilares a partir de imagens AFM e no segundo caso, os autores construiram modelos preditivos para propriedades reológicas de queijos.

O exemplo apresentado na Figura 1.1 ilustra o potencial do uso de técnicas quantitativas para a análise de imagens. Esta figura contém imagens de microscopia de força atômica (AFM) de amostras de fibras capilares de um mesmo indivíduo coletadas em duas regiões distintas da fibra, ponta e raiz, e submetidas ao tratamento de descoloração com peróxido de hidrogênio e persulfato de amônio [Monteiro, 2003].

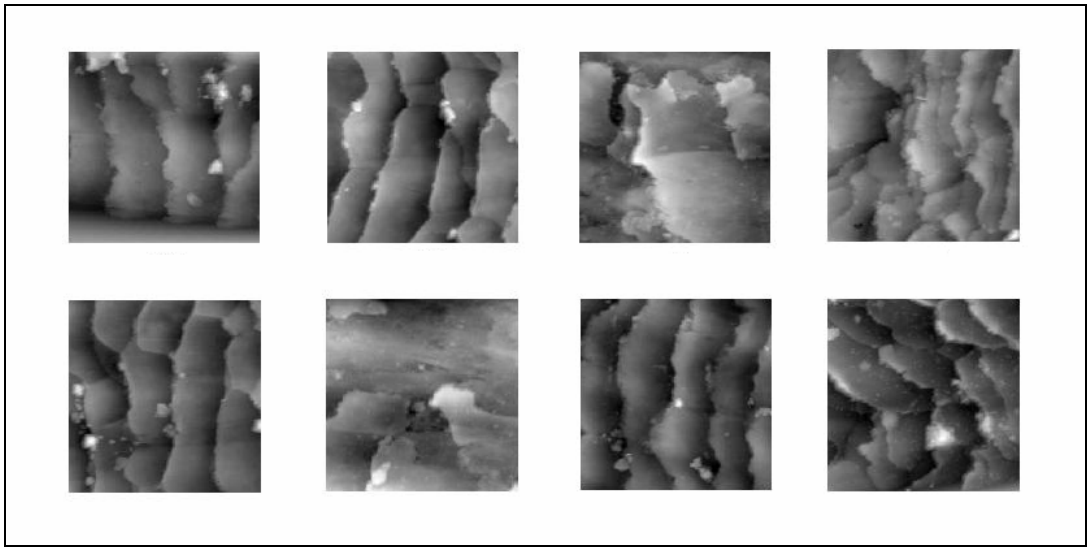


Figura 1. 1: Imagens AFM de fibras capilares.

Para avaliar o efeito do tratamento cosmético na fibra, quatro classes de amostras foram geradas. A Figura 1.1 contém duas imagens de cada uma destas classes:

- ponta controle (sem tratamento);
- ponta descolorida;
- raiz controle (sem tratamento);
- raiz descolorida.

Pela análise visual é possível observar diferenças significativas no estado de degradação da fibra causado pela ação do tratamento cosmético e devido à região em que a amostra foi coletada (ponta/raiz). No entanto, algumas imagens são muito similares e não é possível atribuir todas as amostras às respectivas classes corretamente.

A Figura 1.2(a) contém a identificação das amostras, em que é possível notar que as classes ponta controle e raiz controle são de difícil

distinção entre si. Para finalizar, a Figura 1.2(b) contém uma nova amostra pertencente a uma das classes descritas acima. No entanto, apenas pela comparação visual com as amostras da Figura 1.2(a) não é possível identificar com confiabilidade estatística a qual classe esta nova amostra pertence.

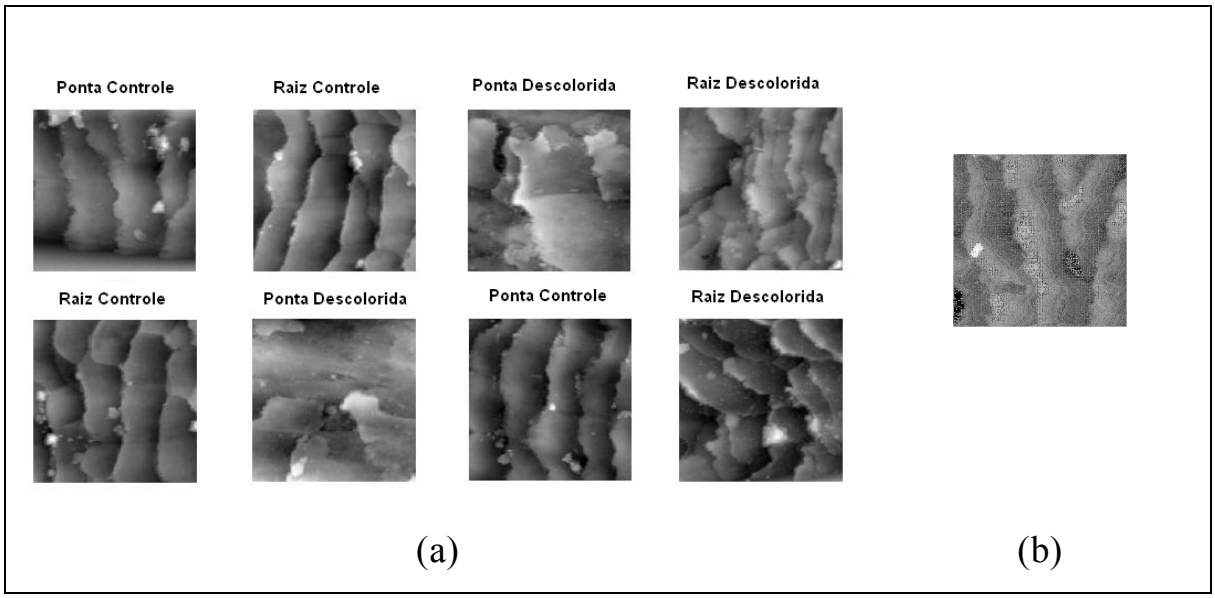


Figura 1. 2: (a) Identificação das classes das imagens AFM de fibras capilares apresentadas na Figura 1.1 e (b) uma nova amostra sem identificação.

Neste trabalho, métodos quimiométricos foram utilizados como ferramenta quantitativa para a análise qualitativa de diversas imagens nível de cinza[2] de microscopia de força atômica de fibras capilares, com propósitos exploratórios e de classificação.

Fibras capilares são frequentemente utilizadas como fonte de amostras para estudos de diversas áreas. Nos estudos de clínica médica elas são

[2] Imagens nível de cinza são também conhecidas como imagens em Preto & Branco (P&B), ou por um dos termos da língua inglesa: *gray-scale* (2867), *grayscale* (634), *gray-level* (1643), *grey-scale* (841), *greyscale* (191) e *grey-level* (822). Os números entre parênteses indicam o número de artigos encontrados com os respectivos vocábulos numa busca recente (Julho/2007) no *Web of Science*.

utilizadas para o diagnóstico de doenças; na área ambiental, a análise de fibras capilares permite avaliar o grau de exposição a agentes tóxicos e poluentes; e em análises forenses são utilizadas na detecção de abuso de drogas, entre outros. Na Indústria Cosmética, estas amostras são ainda alvos de constantes estudos para o desenvolvimento e aperfeiçoamento de produtos [You & Yu, 1997].

Com exceção de uma das aplicações que é específica para fibras capilares, os métodos de análise de imagens propostos neste estudo podem ser utilizados na análise quantitativa de qualquer tipo de imagem nível de cinza.

1.1.1 Imagens Digitais

O uso de técnicas matemáticas e estatísticas para a análise de imagens requer que estas sejam digitalizadas. Imagens digitais são arranjos numéricos em arquivos de dados, e como tal, podem ser submetidas a qualquer técnica numérica, para extração de informações quantitativas ou qualitativas.

Para entender a estrutura deste arranjo numérico, considere uma imagem 2D em Preto e Branco (P&B). Esta imagem possui uma dimensão horizontal x, que pertence ao intervalo contínuo $[0,a]$, e uma dimensão vertical y, que pertence ao intervalo contínuo $[0,b]$, sendo que a e b podem ou não ser iguais, mas usualmente possuem o mesmo valor. A função de intensidade $f(x,y)$ descreve a intensidade em cada ponto no intervalo $[0,a]$ x $[0,b]$, e esta função não existe fora deste intervalo. O domínio desta função pode ser representado pela Figura 1.3, na qual a

origem está demarcada no eixo superior esquerdo [Geladi *et al*., 1992a; Geladi & Grahan, 1996].

O ajuste desta função é extremamente complexo e sua análise não é algo trivial. Por este motivo, para que seja possível a modelagem matemática de imagens, estas devem ser digitalizadas.

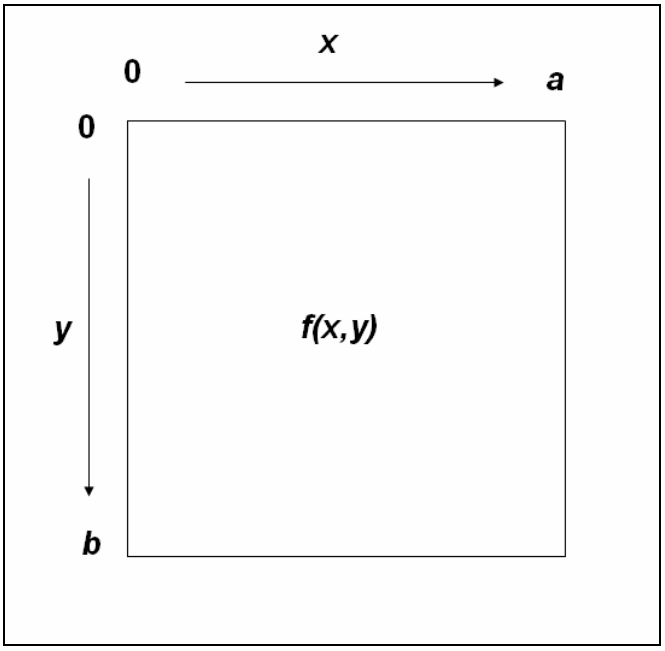


Figura 1. 3: Domínio da função *f(x,y)* para uma imagem P&B em duas dimensões.

Em imagens digitalizadas, os intervalos [0,*a*] e [0,*b*] são transformados em segmentos discretos, e *a* e *b* são substituidos por *i* e *j*, no qual $i = 1, 2, 3, ..., I$, $j = 1, 2, 3, ..., J$. A função que descreve uma imagem digitalizada deve, portanto, ser descrita por $g(i, j)$, em que os valores de *g* representam a intensidade local média da imagem em cada coordenada espacial (i, j), e são representados por um valor inteiro, não negativo e finito. Em imagens em Preto e Branco (P&B) digitalizadas, esta

intensidade *g* é chamada de nível de cinza, e seu domínio pode ser representado pela Figura 1.4.

Cada ponto neste espaço (*i*, *j*) é chamado de *pixel*, do termo em inglês para "*picture element*" [Geladi *et al.*, 1992a; Geladi & Grahan, 1996]. Valores típicos de *I* e *J* para imagens digitalizadas são 32, 64, 128, 256, 512, 1024, 2048, 4096 etc.

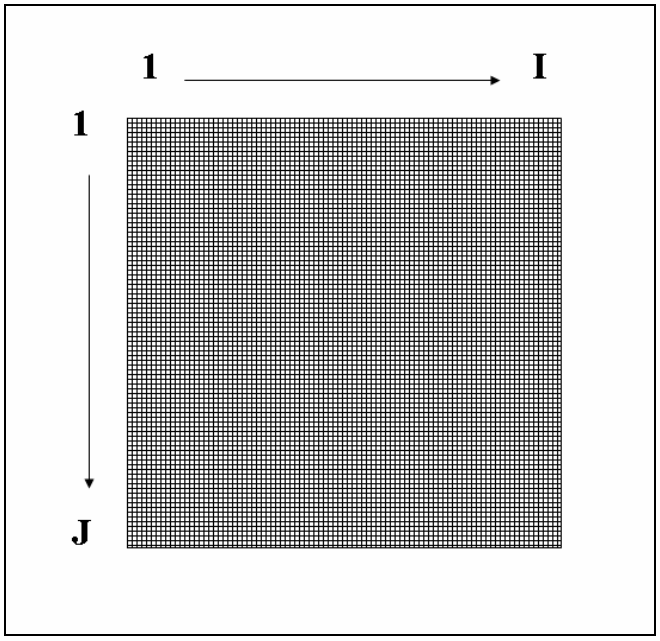


Figura 1. 4: Domínio da função g(i,j) para uma imagem P&B digitalizada.

É possível ainda obter imagens 3D. Neste caso, a função *g* é descrita por *g*(*i*,*j*,*h*), e cada elemento deste arranjo é denominado *voxel*, do termo em inglês para "*volume element*" [Geladi *et al.*, 1992a; Geladi & Grahan, 1996].

1.1.2 Imagens Univariadas x Imagens Multivariadas

As imagens nível de cinza, também conhecidas por imagens *gray-level* ou *nível de cinza*, são imagens univariadas, pois as coordenadas (*i,j*) descrevem apenas a localização espacial dos pixels. Imagens coloridas são formadas pela sobreposição de arranjos do tipo *I* x *J*, um para cada canal de cor, e são consideradas multivariadas, pois qualquer cor em cena será reproduzida pela combinação dos valores de intensidade obtidos nestes três canais. Em televisores e monitores de computador, estes canais são descritos pelas cores vermelho, verde e azul, ou RGB, do inglês para *Red*, *Green* e *Blue* (Figura 1.5) [Geladi *et al.*, 1992a; Geladi *et al.*, 1992b; Geladi & Grahan, 1996].

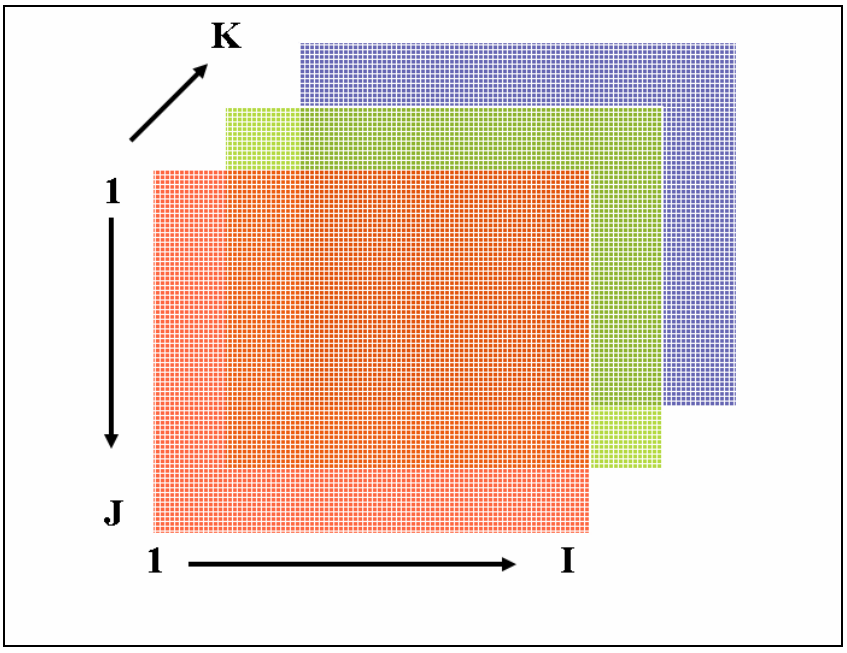


Figura 1. 5: Representação de uma imagem multivariada g(i,j,k) formada pela sobreposição dos três canais RGB.

Este recurso também é utilizado em outros tipos de imagens, como raio-X, ressonância magnética, microscopia eletrônica e técnicas de espectroscopia de imagem [Geladi & Grahan, 1996; Lewis *et al*., 2004; Reich, 2005; De Juan *et al.*, 2004].

1.1.3 Resolução Espacial

A resolução espacial é a capacidade de detalhamento em meios de captação, registro ou reprodução de imagens, e é determinada pelo grau de distinguibilidade conferida aos seus elementos constitutivos[3]. Sendo assim, a digitalização de uma imagem é fortemente influenciada pelo limite da sua resolução espacial. Quanto maior a resolução de uma imagem, maior é a capacidade de registrar os seus detalhes. Em baixas resoluções, alguns objetos são capazes de preservar a sua definição visual melhor do que outros.

No entanto, o aumento da resolução espacial poderá ser um entrave durante o processamento e análise, pois à medida que aumenta a resolução, aumenta também o tamanho do arquivo digital de dados, e cresce a limitação computacional para realização dos cálculos. Este tem sido um dos principais fatores limitantes para a análise e processamento de imagens.

[3] *Novo Dicionário Eletrônico Aurélio versão 5.11©, 2004 by Regis Ltda*.

1.1.4 Processamento de Imagens X Análise de Imagens

Formalmente, há dois principais tipos de operações que podem ser realizadas em imagens: *processamento de imagens* e *análise de imagens*. Estas duas operações podem ser mais bem entendidas se considerarmos as relações existentes entre dados, imagens e técnicas de processamento descritas na Figura 1.7 [Geladi *et al.*, 1992a].

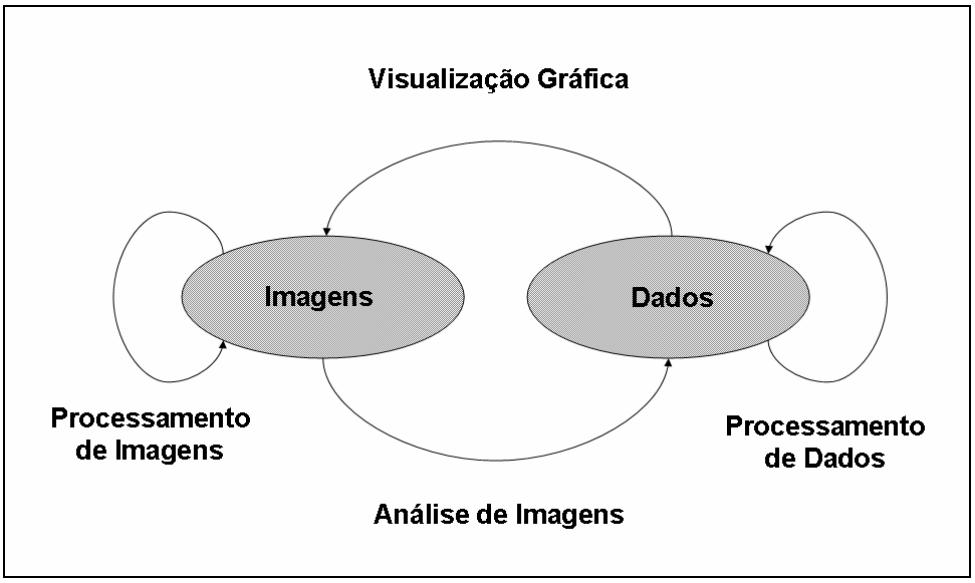


Figura 1. 6: Relações entre dados, imagens e técnicas de processamento (adaptado de Geladi *et al.*, 1992a).

O termo *processamento de imagens* é utilizado para descrever qualquer operação na qual a imagem sofre uma transformação, gerando como resultado uma outra imagem. O objetivo neste caso é melhorar a definição da imagem, ou ressaltar determinados aspectos ou características. Esta transformação pode ser representada pela Equação (1), na qual a imagem $\mathbf{I}_a$ é transformada na imagem $\mathbf{I}_b$, pela função $\mathcal{T}$,

sendo que $\mathbf{I}_a$ e $\mathbf{I}_b$ contêm a forma $g(i,j)$ descrita anteriormente [Geladi *et al.*, 1992a].

$$\mathbf{I_b} = \mathscr{T}[\mathbf{I_a}] \tag{1}$$

Em contrapartida, a *análise de imagens* pode ser entendida como o conjunto de operações matemáticas ($\mathscr{T}$) aplicadas na imagem que permite extrair-lhes informações na forma de dados, cujas estruturas podem ser escalares (s), vetores ($\mathbf{v}$) ou matrizes ($\mathbf{M}$). Este tipo de operação pode ser descrita pela Equação (2).

$$s = \mathscr{T}_1[\mathbf{I_a}] \quad \text{ou} \quad \mathbf{v} = \mathscr{T}_2\,[\mathbf{I_a}] \quad \text{ou} \quad \mathbf{M} = \mathscr{T}_3\,[\mathbf{I_a}] \tag{2}$$

A Figura 1.7 descreve ainda a *visualização gráfica*, que nada mais é do que a representação de uma estrutura de dados seja ela vetorial ou matricial, por um dispositivo de saída de dados (monitor de vídeo, impressora, projetor multimídia etc.).

Por fim, o *processamento de dados* representa qualquer operação lógica ou aritmética que pode ser realizada sobre uma informação, na forma de dados, a fim de obter um resultado desejado ao usuário. Um exemplo de *processamento de dados* é a simples utilização da calculadora, no qual digitamos os dados de entrada, que são os valores numéricos, processamos os dados ao selecionar alguma operação matemática e obtemos os dados de saída, que são os resultados da operação selecionada.

Outro exemplo dessa relação entre dados e imagens pode ser o uso de um simples editor de textos. Quando digitamos um texto, via teclado, os

dados são enviados para o computador, quando utilizamos comandos de formatação, exigimos o *processamento de dados* pelo editor, que em geral já gera a *visualização gráfica* da imagem do documento, para que possamos verificar o resultado do processamento no monitor de vídeo. Outra forma de *visualização gráfica* é a impressão desse documento. Ao inserimos uma imagem matricial neste documento, estamos mais uma vez exigindo a *visualização gráfica* dos dados desta imagem, se no documento alterarmos qualquer característica da imagem, como: dimensão, cor, brilho, contraste, transparência e etc., estaremos exigindo *processamento da imagem*, e caso seja necessário obter alguma informação desta imagem, como: dimensão, área, linhas de tendência, limites entre cores e etc., estaremos realizando a *análise da imagem*, embora estas funcionalidades de análise não sejam comuns na maioria dos editores de texto, mais sim em *softwares* de produção e tratamento de imagens matriciais e/ou vetoriais, como os produtos: *Corel Draw*, *Photoshop* e *AutoCad*.

Processamento de Imagens

As técnicas de *processamento de imagens* permitem aumentar e melhorar a definição da imagem, com o intuito de facilitar a sua visualização [Van Espen *et al.*, 1992]. Os métodos de processamento de imagens são geralmente classificados em pontuais, locais, geométricos e globais, dependendo da quantidade de dados que são utilizados para gerar a imagem de saída. Nos métodos denominados pontuais, cada pixel de entrada gera um único pixel de saída. Nos métodos locais, os pixels de saída dependem dos valores dos pixels vizinhos ao pixel de entrada.

Nos métodos geométricos, as coordenadas dos pixels são consideradas nos cálculos, por exemplo, em operações tais como a rotação de imagens. Em métodos globais, os valores de todos os pixels de entrada afetam os pixels de saída. Um exemplo deste tipo de operação é a aplicação de filtros nos domínios de Fourier e wavelet, que são utilizados em geral para melhorar a imagem de entrada e ressaltar diferenças no gradiente de cor [Geladi *et al*. 1992a; Huang *et al.*, 2003; Bonnet, 2004].

Análise de Imagens

As técnicas de *análise de imagens* permitem a extração de informação quantitativa ou qualitativa de imagens, isto é, permitem gerar dados a partir de imagens.

Diversos métodos quimiométricos têm sido empregados na análise de imagens multivariadas, uma área da quimiometria conhecida pela sigla MIA, e que tem origem no nome em inglês *Multivariate Image Analysis* [Esbensen & Geladi, 1989; Geladi *et al.*, 1989; Geladi *et al.*, 1992a; Geladi *et al.*, 1992b; Geladi & Grahan, 1996]. A maioria dos métodos de MIA são baseados no uso da análise de componentes principais (PCA[4]) das imagens após estas serem submetidas a um processo de desdobramento ao longo do modo apropriado. Este tipo de abordagem em geral é aplicado a uma única imagem multivariada formada por diversos canais, como a imagem apresentada na Figura 1.5. Estes canais podem ser espectrais ou temporais, e a análise utilizando PCA será útil

[4] *Principal Component Analysis*.

na detecção de variações ao longo da imagem e na identificação de suas causas por meio da análise dos *escores* e dos *pesos*[5].

Métodos de regressão baseados no desdobramento de imagens também são muito utilizados e são conhecidos como regressão de imagens multivariadas (MIR[6]) [Lied *et al.*, 2000; Lied & Esbensen, 2001]. Exemplos interessantes de MIA e MIR são encontrados na área farmacêutica, na avaliação da homogeneidade de misturas [Lyon *et al.*, 2002], identificação de impurezas [De Juan *et al.*, 2004; Roggo *et al.*, 2005] e caracterização de efeitos de processos [Roggo *et al.*, 2005; Lewis *et al*., 2004]. Outras aplicações podem ser encontradas nas áreas de ciências de alimentos [Antonelli *et al.*, 2004; Yu *et al.*, 2003; Courcoux et al., 2002; Chevallier et al., 2006; Borin et al., 2007], ciência de materiais [Bharati et a., 2004; Artyushkova & Fulghum, 2002; van den Broek et al., 1996; Huang & Esbensen, 2000 e 2001], biomedicina [Nattkemper, 2004], espectroscopia de imagens [Liu *et al.*, 2005; Burger & Geladi, 2006] e na investigação de fenômenos de difusão em líquidos [Gurden *et al*., 2003].

Huang *et al.* (2003) demonstraram que métodos baseados no desdobramento de imagens são apropriados para imagens caracterizadas por arranjos do tipo **OOV** (Objeto x Objeto x Variável). As imagens multivariadas são exemplos clássicos deste tipo de arranjo, pois apresentam a estrutura pixels (**O**) x pixels (**O**) x canal de cor (**V**). Imagens univariadas por outro lado, como é o caso das imagens nível de cinza, são caracterizadas por arranjos do tipo **OO** (pixel x pixel). Estas imagens podem ser empilhadas formando um arranjo do tipo **OOO**

[5] Cf Apêndice para detalhes a respeito dos métodos quimiométricos citados neste capítulo.
[6] *Multivariate Image Regression*.

(amostras x pixels x pixels) e que quando submetido à utlização de filtros apropriados fornecerá um arranjo **OVV** trilinear, que é capaz de ser modelado por métodos *multi-way* que não exigem desdobramento, como PARAFAC[7], TUCKER e NPLS[8].

Os métodos *multi-way* apresentam grandes vantagens para a análise de imagens nível de cinza quando comparados com os métodos baseados em desdobramento, pois permitem a análise de diferentes imagens não congruentes [Huang *et al.*, 2003; Geladi *et al.*, 2000; Gurden *et al.*, 2003].

Em geometria, congruência significa que dois polígonos podem ser empilhados de tal maneira que cada um poderá encobrir o outro completamente [Geladi & Grahn, 1996]. Assim, duas ou mais imagens são congruentes se elas são perfeitamente empilháveis, de tal forma que cada pixel em uma imagem possuirá um pixel espacialmente correspondente nas outras imagens, como é o caso das imagens multivariadas.

Arranjos formados por imagens de diferentes amostras não são congruentes, uma vez que os pixels de uma imagem não são espacialmente equivalentes aos das outras imagens.

1.2 Fibras Capilares

O cabelo humano é um tecido complexo rico em queratina, que nasce a partir de cavidades subcutâneas denominadas folículos. Os folículos capilares apresentam-se em toda a extensão da pele, desde a superfície

[7] *Parallel Factor Analysis*.

até o extrato córneo, a epiderme e a derme. As fibras capilares são características apenas de mamíferos e apresentam funções protetoras, sensoriais e também como atrativo sexual [Robbins, 1994].

As fibras são formadas por quatro estruturas distintas: as cutículas, o córtex, a medula e o complexo da membrana celular (CMC) (Figura 1.8), e seu diâmetro pode variar de 15 a 120 μm dependendo de fatores como genética, idade, origem e estado de degradação [Robbins, 1994].

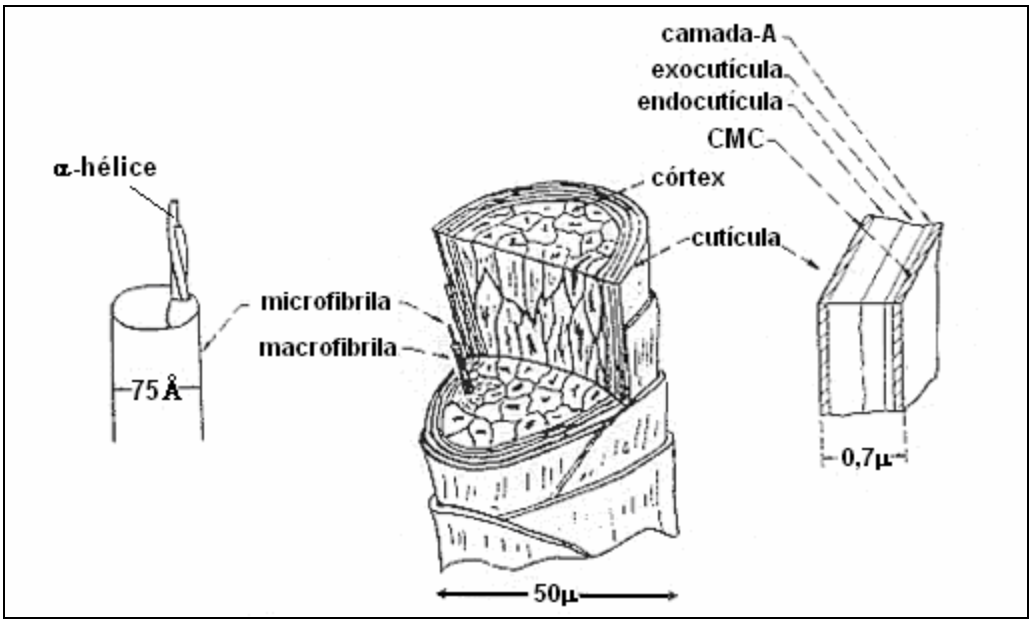


Figura 1. 7: Estrutura da fibra capilar e seus componentes [Adapt. Robbins, 1994].

A superfície externa do cabelo humano, assim como as fibras capilares de mamíferos em geral, é uma imbricada construção em que o córtex central é envolto por largas e finas camadas celulares, as cutículas, que se sobrepõem na direção da raiz até a ponta. Cada cutícula tem aproximadamente de 0,5 a 1.0 μm de espessura, 45 μm de comprimento

[8] *Multilinear Parcial Least Squares*.

e em torno de 60 μm^2 de área. Possuem bordas arredondadas e são separadas entre si por finas camadas do complexo da membrana celular [Robbins, 1994]. Na direção longitudinal, as cutículas apresentam um ângulo médio de aproximadamente 5° em relação ao eixo da fibra, e a sobreposição de suas extremidades forma uma série de degraus irregulares com aproximadamente 5 μm de extensão. A Figura 1.9 sumariza estas estruturas [Swift, 1999; Swift & Smith, 2000].

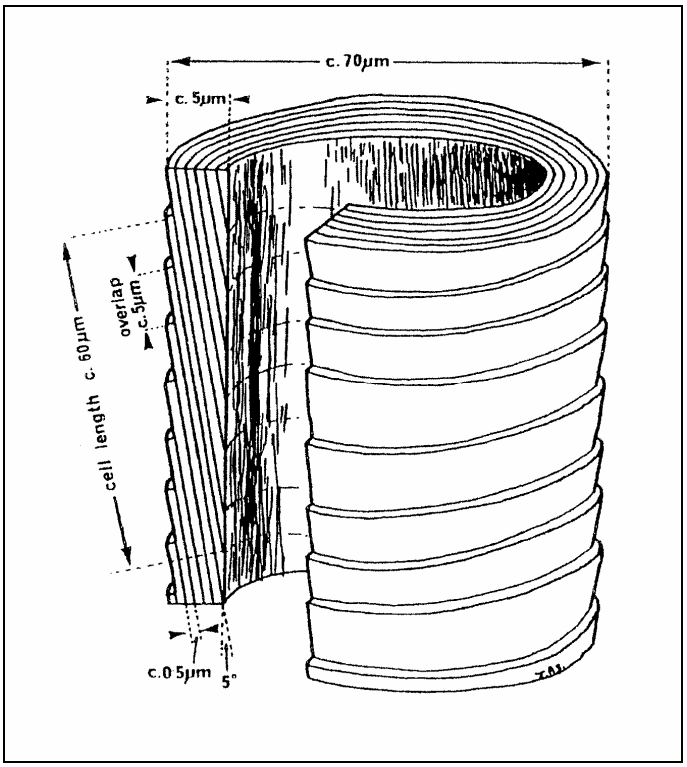


Figura 1. 8: Diagrama esquemático ilustrando a forma e dimensões médias das cutículas em uma fibra capilar [Fonte: Swift, 1999].

Muito do nosso conhecimento da estrutura e constituição química interna destes componentes derivam das observações realizadas

inicialmente com Microscopia Eletrônica de Transmissão (TEM[9]), Microscopia Eletrônica de Varredura (MEV[10]) de fibras seccionadas [Robbins, 1994; Swift & Brown, 1972; Swift, 1999; Smith, 1998; Swift & Smith, 2000], e mais recentemente com o uso de Microscopia de Força Atômica (AFM) [You & Yu, 1997; Swift, 1999; Swift & Smith, 2000; Hadjur *et al.*, 2002; McMullen *et al.*, 2000; McMullen & Kelty, 2001; Monteiro, 2003; Gurden *et al.*, 2004].

1.2.1 As Cutículas

As cutículas têm função protetora e ocorrem na superfície da fibra capilar, apresentando-se na forma de camadas concêntricas sobrepostas umas às outras, semelhante às telhas em um telhado, e distribuídas ao redor do córtex. Estas camadas não possuem pigmentos e são totalmente queratinizadas, sendo o componente da fibra capilar que possui o maior teor de cistina [Robbins, 1994].

As cutículas constituem a parte mais externa da fibra e, portanto, sofrem diretamente os efeitos dos agentes agressores, sendo muito utilizadas como indicadores da influência de fatores externos sobre a estrutura da fibra capilar [You & Yu, 1997; Monteiro, 2003; Monteiro *et al.* 2005; Gurden *et al.*, 2004; Smith, 1998].

Na região próxima à raiz, o cabelo humano apresenta cutículas uniformes, intactas e com bordas arredondadas. Ao longo do fio e em direção às pontas, as cutículas apresentam danos como desbotamento, bordas quebradas e irregulares, e até mesmo regiões com ausência total

[9] *Transmission Electron Microscope*.
[10] Também conhecido pela sigla do seu nome em inglês, SEM, *Scanning Electron Microscopy*.

de cutículas. Estes danos são causados pela escovação, lavagem com xampu, ação de cosméticos, processos de coloração e exposição ao sol. Cada cutícula é formada por diversas estruturas: uma fina camada externa denominada epicutícula, a camada A, e duas outras camadas internas, a exocutícula e a endocutícula. Estas camadas são separadas pelo complexo da membrana celular (CMC), constituido basicamente pela camada-β e pela camada-δ internas, e que permitem a aglutinação das cutículas e das células corticais [Robbins, 1994; Swift, 1999]. Detalhes da sua constituição podem ser observados na Figura 1.10.

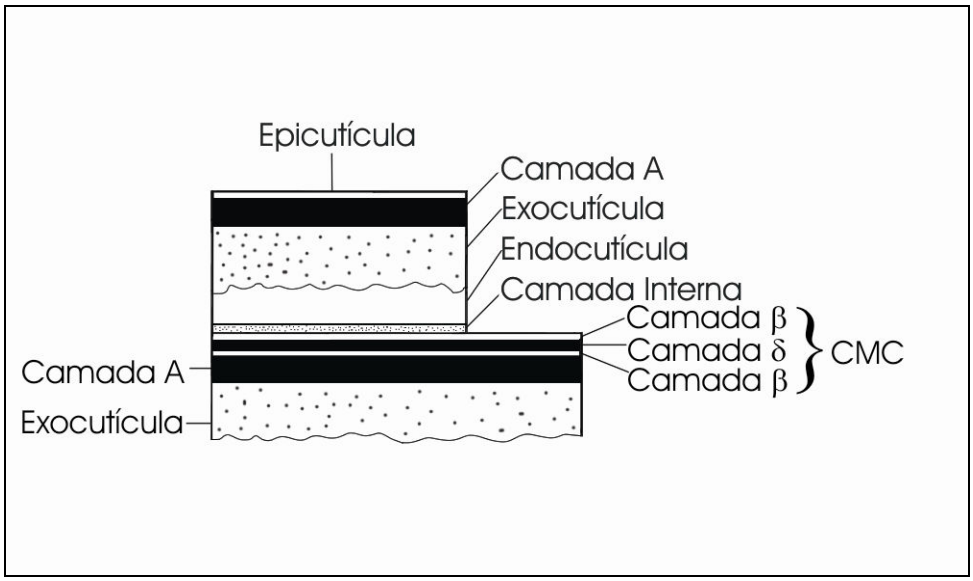


Figura 1. 9: Diagrama da estrutura interna da cutícula perto da borda, mostrando a cutícula adjacente, em que é possível observar o material intercelular [Adaptado de Robbins, 1994 e Swift, 1999].

A epicutícula é uma membrana protéica quimicamente resistente, e com aproximadamente 250 nm de espessura. É a camada mais externa e, portanto, a que mais influencia as propriedades da superfície, sendo resistente a ácidos, agentes oxidantes e ataques enzimáticos [Robbins,

1994; Monteiro, 2003]. A camada-A localiza-se na região mais externa de cada cutícula e apresenta espessura constante, com valor médio de 110 nm. Possui constituição protéica com alto teor de cistina[11] (>30%), que contém enxofre (Figura 1.11) e é altamente resistente à degradação por agentes químicos e físicos [Robbins, 1994; Swift, 1999].

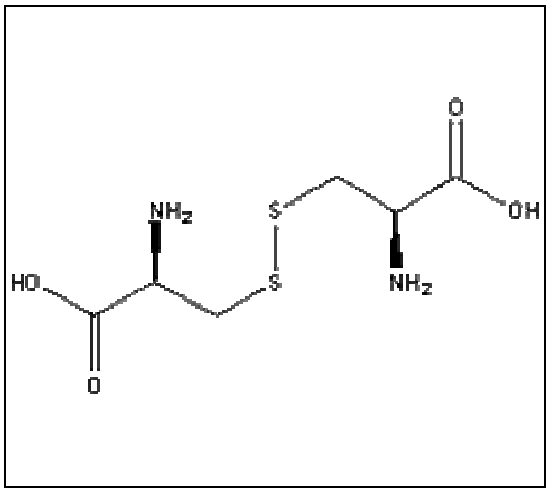


Figura 1. 10: Estrutura da cistina [Fonte: Chemfinder].

A exocutícula, também conhecida como camada-B, localiza-se logo abaixo da camada-A e também possui constituição protéica com alto teor de cistina (~15%) em relação ao córtex, mas menor do que aquele encontrado na camada-A [Robbins, 1994; Swift, 1999]. Sua espessura varia entre 100 e 300 nm e, juntamente com a epicutícula, forma uma capa protéica que protege a fibra capilar contra danos mecânicos. Estas estruturas com maior teor de cistina (epicutícula, camada A e

[11] Cistina e Cisteína: cistina é a denominação dada à formação estável do aminoácido sulfurado cisteína. Dos 20 aminoácidos existentes, apenas dois (metionina e cistina) apresentam enxofre em sua constituição [Lehninger, 1989].

exocutícula) são altamente reticuladas por ligações dissulfeto (-SS-), o que lhes confere natureza hidrofóbica.

A endocutícula apresenta espessura altamente variável, que compreende uma faixa de 50 a 300 nm. É constituída por pelo menos três componentes protéicos química e morfologicamente distintos, com altos níveis de aminonácidos de caráter ácido e básico[12] quando comparados aos outros constituintes das cutículas. Outro fator distinto em relação às duas camadas anteriores é o seu baixo teor de cistina (3~%), que lhe confere natureza mais hidrofílica, quando comparada às outras estruturas [Swift, 1999].

O alto conteúdo de aminoácidos ácidos e básicos da endocutícula, juntamente com a relativa ausência de ligações intermoleculares na forma de cistina e isodipeptídeos tornam esta camada menos rígida e relativamente susceptível à permeação por água [Swift, 1999]. Estudos utilizando AFM detectaram um considerável aumento no *step height*[13] de fibras capilares e lã expostas à umidade [You & Yu, 1997].

1.2.2 Complexo da Membrana Celular

O complexo da membrana celular (CMC) consiste de uma série de membranas celulares de material adesivo que permitem a aglutinação das cutículas e das células corticais. Considera-se que a CMC engloba duas subunidades com características bem distintas, camadas-β e a camada-δ interposta, mas na maioria das vezes pode-se considerar

[12] Dos 20 aminoácidos existentes, dois contêm grupos R carregados negativamente (ácido aspártico e ácido glutâmico) e apresentam caráter ácido, e três deles contêm grupos R carregados negativamente (lisina, arginina e histidina) e apresentam caráter básico [Lehninger, 1989].

apropriado incluir outras finas camadas protéicas periféricas [Robbins, 1994; Swift, 1999].

A camada-δ, conhecida como cimento intercelular, compreende uma fina lâmina de aproximadamente 5 nm de espessura e sua constituição química ainda é incerta, devido principalmente à dificuldade em isolá-la para estudos sem que sofra alterações ou contaminação dos outros constituintes da fibra, mas há evidências de que seja constituída de proteínas e polissacarídeos, e de que apresenta ausência de cistina [Swift, 1999].

A camada-β localiza-se na região contígua à superfície da cutícula e é constituída principalmente por lipídios saturados. Uma característica importante desta camada é a presença do ácido 18-metil-eicosanóico (18-MEA), um ácido graxo covalentemente ancorado na camada superficial das fibras e que está presente no CMC tanto de cutículas quanto do córtex. No CMC da cutícula, 18-MEA é o ácido graxo mais abundante (~50% em massa), seguido dos ácidos palmítico e oléico. No CMC do córtex ele aparece somente em nível de traços [Swift, 1999].

De modo geral, o CMC contém baixa proporção de aminoácidos contendo enxofre quando comparado com outros constituintes da fibra, e junto com a endocutícula este complexo algumas vezes é chamado de região "não queratinosa". Estas regiões ainda foram pouco estudadas, mas acredita-se que sejam o caminho para a entrada ou difusão de cosméticos dentro da fibra e o entendimento de como isto acontece tem sido alvo dos estudos na área cosmética [Silva & Joekes, 2005; Colombera, 2004; Robbins, 1994].

---

[13] Cf. seção 2.1, para detalhes sobre *Step Height*.

Durante as operações rotineiras de cuidados com os cabelos, como escovação, lavagem com xampu, uso do secador etc., o ato de estirar o fio provoca o lento desprendimento e quebra das cutículas. Estes estragos ocorrem nesta região não queratinosa, inicialmente na ponta do fio, e posteriormente ao longo de toda a sua extensão [Robbins, 1994]. O estiramento do fio produz tensão na região não queratinosa (complexo da membrana celular e endocutícula), provocando rachaduras. Se ocorrer cisão da endocutícula, esta se eleva e sofre corrosão por ação mecânica, expondo uma superfície áspera e granular. O CMC é extremamente suscetível ao ataque por ácidos, soluções aquosas quentes e agentes redutores [Robbins, 1994].

1.2.3 O Córtex

O córtex constitui a parte principal da massa da fibra capilar e representa 70% da sua massa total. É constituído por macrofibrilas de queratina, estruturas alongadas e retorcidas, que se alinham ao longo do eixo da fibra capilar (Figura 1.8) [Robbins, 1994; Monteiro *et al.*, 2005]. Estas macrofibrilas possuem duas estruturas principais, as microfibrilas e a matriz, que diferem em estrutura e composição de aminoácidos. As microfibrilas são estruturas protéicas fibrosas e cristalinas, compostas principalmente por proteínas helicoidais com baixo teor de cistina [Kuzuhara, 2006]. Estas estruturas estão alinhadas ao longo do eixo da fibra e imersas em uma matriz amorfa com alto teor de cistina, que consiste de grupos dissulfeto (-SS-), que formam ligação cruzada nas fibras de queratina, contribuindo para suas propriedades físicas e

mecânicas e para sua estabilidade estrutural [Kuzuhara, 2006; Robbins, 1994].

As células do córtex apresentam ainda pequenos grânulos de melanina (~3% da massa total da fibra), que podem apresentar formato oval ou esférico, variando entre 0,2 a 0,8 μm de diâmetro, e que se dispersam ao longo das células corticais. A melanina existe em duas formas quimicamente distintas: as eumelaninas, que são responsáveis pela cor em cabelos castanhos e pretos, e as feomelaninas, que são responsáveis pela cor de cabelos loiros e vermelhos. As diferentes tonalidades da cor dos cabelos são definidas pelo formato, tipo e quantidade de melanina presente na fibra, sendo que em fibras humanas estes pigmentos não ocorrem nas cutículas [Ozeki, 1996; Robbins, 1994].

O centro do córtex contém a medula, que pode estar ausente ou presente ao longo de todo comprimento do fio.

1.2.4 A Medula

A medula está contida no centro do córtex e apresenta-se em pequena quantidade e irregularmente distribuída. Pode estar presente ou ausente ao longo de todo o comprimento do fio e em geral ocorre nas fibras mais espessas. A medula representa uma pequena porcentagem da massa total da fibra e é muito difícil de ser isolada, portanto, tem sido pouco estudada [Robbins, 1994; Monteiro, 2003].

1.3 A constituição química de fibras capilares

A fibra capilar é formada principalmente por α-queratina, uma proteína insolúvel e que corresponde a 85% da massa total do cabelo. Outros constituintes não menos importantes são os lipídios (3%), pigmentos (2%), algumas substâncias hidrossolúveis (pentoses, fenóis, ácido úrico, glicogênio), água e metais. A α-queratina é um biopolímero constituído por um conjunto de cadeias polipeptídicas formadas pela condensação de diferentes aminoácidos, sendo que a proporção de aminoácidos em geral é diferente na cutícula e no córtex, como foi apresentado anteriormente. A arquitetura da fibra é formada por diversos tipos de ligações e que contribuem para a estabilização da estrutura protéica da fibra: ligações peptídicas e ligações de hidrogênio, que ocorrem entre os grupos NH e CO das cadeias polipeptídicas paralelas e adjacentes e estão presentes em grande número; ligações salinas, ou coulômbicas, que são resultantes de interações eletrostáticas entre grupos básicos de cadeias laterais de aminoácidos (íons amônio positivamente carregados) e grupos ácidos de cadeias laterais de aminoácidos (grupos carboxila negativamente carregados) da mesma cadeia helicoidal ou de outra cadeia e, principalmente, ligações dissulfeto (-SS-), que são formadas pela ligação entre dois átomos de enxofre no aminoácido cistina. Este tipo de ligação é particularmente importante e característica de fibras capilares, pois são extremamente fortes e determinam suas propriedades físicas e mecânicas [Robbins, 1994; Monteiro, 2003].

A cistina é o aminoácido presente em maior abundância na queratina, sendo responsável pela sua extrema resistência e insolubilidade. A

cistina possui dois grupos NH_2 e dois grupos COOH podendo, portanto, participar de duas cadeias pepitídicas interligadas por ligações de dissulfeto do aminoácido. Estas ligações podem ser intercadeias ou intracadeias e são responsáveis pelo grau de ondulação da fibra capilar [Robbins, 1994; Monteiro,2003].

1.4. O efeito do crescimento da fibra: ponta vs. raiz

A fibra capilar apresenta uma variação natural e gradual em toda a sua extensão, da raiz até a ponta, e que tem sido largamente estudada utilizando técnicas de microscopia eletrônica de varredura [Swift & Brown, 1972; Swift, 1999; Smith, 1998; Swift & Smith, 2000] e AFM [Swift, 1999; Swift & Smith, 2000; Hadjur *et al.*, 2002; McMullen *et al.*, 2000; McMullen & Kelty, 2001]. Na região em que a fibra emerge do couro cabeludo, próxima à raiz, a superfície da cutícula é relativamente lisa e uniforme, com cutículas regularmente imbricadas e bordas com contorno liso e uniforme [Swift & Brown, 1972, Swift & Smith, 2000].
À medida que se aproxima da região da ponta, a superfície da cutícula torna-se cada vez mais irregular e o conteúdo de ácido 18-MEA diminui gradativamente devido ao atrito mecânico das cutículas na superfície da fibra [Swift, 1999]. O contorno liso das bordas das cutículas começa a ser progressivamente eliminado, tornando-se mais angular e serrilhado [Swift & Brown, 1972, Swift & Smith, 2000]. As bordas das cutículas superiores tendem a se romper e frequentemente há marcas e estrias paralelas ao eixo da fibra na região onde as cutículas foram arrancadas.

Na extremidade da ponta, é possível ainda encontrar regiões com ausência total de cutículas e com o córtex totalmente exposto, sendo que nestes casos pode ocorrer a divisão da fibra em dois ou mais componentes semi-cilíndricos conhecidos popularmente como "pontas duplas" [Hadjur *et al.*, 2002; Swift & Brown, 1972, Swift & Smith, 2000]. A completa remoção da cobertura de cutícula leva a maior vulnerabilidade do fio, que pode levar à sua ruptura [Swift & Smith, 2000].

Cabelos normais sofrem deterioração gradual com o tempo e a atrofia mecânica pela escovação, penteabilidade e manipulação contribuem consideravelmente para estas mudanças. Estima-se que a taxa de perda de cutícula em um cabelo normal seja de 0,5 μm a cada 2000 escovadas. O comprimento médio de uma cutícula ao longo do eixo da fibra é 40μm, então cerca de 160.000 escovadas seriam capazes de remover todas as cutículas. A taxa média de crescimento do cabelo humano é de 0,35 mm dia^{-1} [Swift & Brown, 1972, Swift & Smith, 2000] e então para que a remoção de toda a cutícula a uma distância de 40 cm a partir do couro cabeludo de um indivíduo fosse possível seria necessária uma taxa de 150 escovadas dia^{-1}, o que é incomum. A taxa de degradação das fibras muda muito em função do estilo de vida do indivíduo e da exposição da fibra a tratamentos cosméticos como descoloração, permanente, tinturas, secador e xampu, além de outros processos naturais como exposição ao sol, chuva, poeira etc. [Swift & Brown, 1972, Swift & Smith, 2000].

De forma geral, o aspecto das fibras não é homogêneo. Em um mesmo indivíduo, por exemplo, as fibras do topo da cabeça apresentam diferenças em relação às fibras das têmporas e da nuca etc. [Swift &

Brown, 1972; Swift & Smith, 2000]. Ouro aspecto importante é a ondulação presente na superfície das cutículas e que gradativamente se intensifica da raiz até a ponta. Este aspecto não tinha sido detectado inicialmente nos estudos utilizando técnicas de MEV, que mostravam uma superfície mais lisa. Swift & Smith, 2000 conseguiram detectar esta ondulação utilizando técnicas de AFM.

Com técnicas de AFM é possível notar ainda algumas descontinuidades na superfície das cutículas de cabelos mais deteriorados. Estas descontinuidades são regiões de trauma, que ocorrem nas cutículas subjacentes na região onde originalmente existiam as bordas de cutículas sobrejacentes e que foram quebradas e removidas. Estas descontinuidades frequentemente apresentam-se na forma de um resíduo granular, e alguns estudos indicam que se devem à fratura através da endocutícula, com perda da camada mais externa e mais dura (exocutícula e camada-A) [Swift & Smith, 2000].

Utilizando técnicas de AFM, McMullen *et al.*, (2000) e McMullen & Kelty (2001) observaram que tanto fibras virgens quanto fibras submetidas a diferentes tratamentos (descoloração e extração com solvente) apresentam grande quantidade de microporos na superfície das cutículas. Este aspecto pode ser interpretado como uma degradação da superfície, e que se intensifica com a idade da fibra e o tipo de tratamento submetido devido aos efeitos dos agentes alcalinos/oxidativos, ou extração por solventes, que dissolvem ou solubilizam os lipídios da superfície da fibra, revelando uma superfície com microporos.

1.5. O Efeito do Descoloramento

O objetivo do tratamento de descoloração é clarear as fibras, e este efeito é alcançado pela oxidação dos grupos cromóforos dos pigmentos do cabelo (as melaninas). No entanto, devido às severas condições necessárias para destruir estes grupos cromóforos, efeitos colaterais ocorrem simultaneamente nas proteínas das fibras capilares, que possuem grupos altamente oxidáveis (por exemplo, as ligações dissulfeto). A oxidação de proteínas da fibra ocorre tanto na matriz do córtex quanto nas cutículas [Robbins, 1994].

A descoloração de fibras capilares é usualmente realizada com uso de peróxido de hidrogênio, que destrói alguns dos pigmentos naturais da fibra, provoca oxidação da cisteína a ácido cistéico, e produção em menor quantidade de intermediários como o monóxido ou dióxido de cistina [Robbins, 1994; Kuzuhara, 2006; Monteiro, 2003]. Ainda em menor quantidade, outros resíduos de aminoácidos como tirosina, treonina e metionina também sofrem degradação [Robbins, 1994]. Para acelerar a descoloração, em geral são utilizados persulfatos de potássio ou amônio e meio básico (entre pH 9 e pH 11) ajustado com hidróxidos de sódio ou amônio [Robbins, 1994; Monteiro, 2003; Kuzuhara, 2006].

Dois tipos de mecanismos foram propostos para a degradação oxidativa das ligações dissulfeto: uma fissão –S–S– e uma fissão –S–C– (Equações (3) e (4)), sendo que a primeira é predominante [Robbins, 1994].

Fissão –S–S–

$$R–S–S–R \rightarrow R–SO–S–R \rightarrow R–SO_2–S–R \rightarrow [R–SO_2–SO–R\,] \rightarrow R–SO_2–SO_2–R \rightarrow 2R–SO_3H \qquad (3)$$

Fissão –S–C–

$$R–S–S–R \rightarrow R–S–S–OH + R–OH \rightarrow R–S–SO_2H \rightarrow R–SO_3H \rightarrow R–SO_3H + H_2SO_4 \quad (4)$$

Os efeitos da descoloração são mais intensos nas regiões morfológicas ricas em cistina, como a camada A e a exocutícula da cutícula, e na matriz do córtex. A superfície da fibra é mais afetada pelo tratamento que a região central, e a região da ponta, por ser mais exposta, é mais afetada que a região da raiz [Robbins, 1994].

Os efeitos do cuidado diário (escovação, xampu, secador etc.), exposição ao sol e uso de tratamentos cosméticos sobre fibras humanas são amplamente visuais e tácteis e estão intimamente relacionados à arquitetura superficial das fibras individuais. Swift & Brown (1972) utilizaram MEV para comprovar as mudanças na superfície das fibras decorrentes da aplicação de *sprays*, permanente, descoloração e remoção de depósitos de sujeira com xampu. No entanto, fibras capilares apresentam grande variabilidade estrutural e somente os efeitos mais acentuados destes tratamentos foram detectados pela técnica MEV. Swift & Brown (1972) compararam fibras controle com fibras submetidas a 30 min de descoloração com uma solução de 9% de peróxido de hidrogênio[14]. Os autores detectaram pequenas variações na estrutura superficial da fibra descolorida, além da ocasional quebra de pequenos fragmentos provenientes das bordas das cutículas, indicando que após o tratamento com peróxido as cutículas tornam-se mais

[14] Neste caso, os autores utilizaram apenas peróxido, que possui um poder oxidante menor do que quando associado a persulfato de amônio. As imagens utilizadas nesta tese são de fibras submetidas a uma solução contendo peróxido (6%), persulfato de amônio e hidróxido de amônio na proporção 2:1:0,5.

suscetíveis às perdas devido ao atrito mecânico entre as fibras. Posteriormente, Swift & Smith (2000) utilizando AFM revelaram que a superfície da cutícula não é tão lisa quanto se supunha anteriormente nos estudos com MEV. Também utilizando MEV e AFM, Monteiro (2003) observou a completa deformação da região do córtex e a formação de microporos em fibras submetidas ao tratamento de descoloração por 30 minutos em uma solução (2: 1: 0,5) contendo peróxido de hidrogênio 6%, persulfato de amônio e hidróxido de amônio. Estes resultados estão em concordância com aqueles obtidos por Sant'Anna (2000) utilizando microscopia eletrônica de transmissão, que observou que após 30 minutos em contato com uma solução a 32°C contendo H_2O_2, persulfato de potássio e pH 10,5 (NH4OH), grande parte do material cuticular foi removido.

Recentemente, Monteiro *et al.* (2005) utilizando técnicas de análise termogravimétrica (TGA) e calorimetria diferencial de varredura (DSC) obervaram diferenças no comportamento térmico de fibras controle e fibras submetidas ao tratamento de descoloração com peróxido, persulfato de amônio e hidróxido de amônio. Nas análises com TGA, as fibras descoloridas apresentaram número menor de etapas na perda de massa devido à degradação da queratina, e estas etapas ocorreram em temperaturas maiores quando comparadas com as fibras controle. Isto indica que este tratamento de descoloração influencia o processo de perda de massa, sugerindo que a estrutura de queratina da fibra torna-se mais desorganizada após este tratamento apresentando, portanto, um menor número de estágios de perda de massa. As curvas de DSC indicam aumento na temperatura máxima em que ocorre vaporização da água e desnaturação da queratina nos cabelos descoloridos, quando

comparados ao cabelo controle. Isto indica que cabelos submetidos à descoloração retém menor quantidade de água, que é responsável pelo enfraquecimento das ligações intercadeias e intracadeias das ligações de queratina, e que mantêm a estrutura conformacional do cabelo [Monteiro *et al.*, 2005; Monteiro, 2003]. O processo de descoloração também promove aumento das interações iônicas devido à formação de ácido cistéico, aumentando desta maneira a estabilidade da estrutura da queratina e elevando as temperaturas máximas de desnaturação, uma vez que mais energia será necessária para desestabilizar esta estrutura. Os autores confirmaram ainda a degradação da estrutura cuticular da fibra e as modificações químicas no córtex pela observação de fibras controle e fibras descoloridas utilizando AFM.

Kuzuhara (2006) investigou as modificações químicas nas regiões da cutícula e do córtex de fibras orientais brancas e pretas submetidas à descoloração (5,9% de peróxido, persulfato de amônio e hidróxido de amônio) utilizando espectroscopia Raman. As observações foram realizadas diretamente em diversas profundidades das fibras seccionadas, sem isolar a cutícula do córtex. O autor observou diminuição significativa na intensidade das bandas S-S (~506 cm^{-1}) e C-S (604 cm^{-1}) e aumento na intensidade da banda S-O (1040 cm^{-1}), indicando a clivagem dos grupos (-SS-) e sua conversão em ácido cistéico, após submeter as fibras ao tratamento de descoloração. Estes efeitos foram drasticamente mais acentuados na região da cutícula, que possui a maior porcentagem de cistina da fibra [Robbins, 1994], em comparação com as modificações observadas na região do córtex.

A análise de aminoácidos das fibras revelou que o conteúdo de ½-cistina de ambas as fibras (preta e branca) diminuiu consideravelmente,

enquanto o de ácido cistéico aumentou após as fibras serem submetidas ao tratamento descolorante. Adicionalmente, os conteúdos de metionina, lisina e histidina nas fibras pretas diminuíram após o tratamento, mas não apresentaram mudanças significativas nas fibras brancas, sugerindo que a quantidade de outros aminoácidos também foi afetada [Kuzuhara, 2006]. Esta variação na composição dos aminoácidos após o tratamento com peróxido também foi descrita por Robbins (1994).

A descoloração aumenta ainda a fricção interfibras consideravelmente [Robbins, 1994; McMullen & Kelty, 2001]. A fricção é a resistência ao movimento provocado pelo deslizamento de um corpo em outro, e no caso das fibras capilares, o aumento da fricção afetará propriedades como penteabilidade, volume e capacidade de reter a forma do penteado.

1.6. Características Étnicas

As fibras capilares são classificadas em três etnias majoritárias: negróide (etnias negras da África, Melanésia e Papua), caucasiano (ancestrais europeus) e mongol (asiáticos, índios americanos e esquimós) cujas características podem ser observadas na Tabela 1 [Robbins, 1994].

Das três etnias, os cabelos caucasianos apresentam o menor diâmeto (50-90 μm, em adultos), com fibras lisas ou levemente onduladas e com formato levemente elíptico.

Tabela 1: Características de fibras das três etnias majoritárias [Robbins, 1994].

Etnia	Características da Fibra			
	Espessura	Curvatura	Formato da Seção Transversal	Cor
Caucasiana	Fino	liso a enrolado	quase redondo a levemente oval	loiro a castanho
Negróide	Espesso	ondulado a crespo	quase oval a levemente elíptico	castanho a preto
Mongol	Espesso	liso a ondulado	quase redondo a levemente oval	castanho a preto

As fibras do tipo negróide são os que mais se desviam da circularidade, com um formato totalmente elíptico, responsável pelo crescimento ondulado e encarapinhado da fibra. A maneira como o cabelo torna-se encaracolado ou crespo ainda não está completamente entendida, mas alguns estudos indicam que esta característica é determinada pela forma do folículo na zona de queratinização, que molda a fibra durante as etapas iniciais do seu crescimento. A frequência desta característica é determinada pela proporção relativa de três tipos de células corticais (ortocórtex, paracórtex e mesocórtex), sua localização no córtex da fibra e sua composição protéica [Robbins, 1994].

As fibras do tipo mongol são as mais circulares, e este fator resulta no aspecto totalmente liso dos cabelos. Estas fibras também apresentam maiores raios de curvatura quando comparadas às outras fibras [Robbins, 1994; McMullen *et al.*, 2000].

Estudos utilizando imagens AFM das três classes de cabelos revelaram que fibras do tipo oriental apresentam as bordas mais serrilhadas, quando comparadas às fibras caucasianas e africanas, além de possuirem maior número de cutículas por área analisada. Neste

parâmetro, os cabelos do tipo negróide são os que apresentam menor número de cutículas por área [Monteiro, 2003].

1.7. O uso de AFM e descritores para estudar características de fibras capilares

Microscopia de Força Atômica (AFM) tem se mostrado uma técnica extremamente útil para o exame não destrutivo da superfície da fibra [You & Yu, 1997; Monteiro *et al.*, 2005; Monteiro, 2003; Smith, 1998; Swift & Smith, 2000, Hadjur *et al.*, 2002]. Esta técnica fornece uma riqueza de informações estruturais usualmente não revelada por outras técnicas como microscopia eletrônica de varredura (MEV) ou de transmissão (TEM) [Smith, 1998; Swift & Brown, 1972, Swift & Smith, 2000]. No entanto, a amostragem deve ser cuidadosa e representativa, pois fibras capilares apresentam grande diversidade de aspecto, variando muito com a localização no corpo e a distância da pele [Smith, 1998].

AFM fornece imagens com alta resolução, em escala nanométrica e com geração de imagens tridimensionais digitalizadas que permitem uma avaliação mais detalhada da superfície da fibra capilar. Esta técnica preserva as características originais da amostra, pois não precisa depositar material nenhum como é o caso de MEV, que requer o recobrimento da amostra com material condutor [You & Yu, 1997; Hadjur *et al.*, 2002].

Recentemente, imagens obtidas com AFM têm sido utilizadas para o cálculo de descritores de fibras capilares, como por exemplo, na

estimativa da altura da cutícula (*step height*[15]) de cabelo humano para caracterização da rugosidade da superfície das cutículas de fibras virgens e fibras submetidas a tratamentos químicos como descoloração, extração por solventes para retirada de lipídios e recobrimento com polímeros [You & Yu, 1997; McMullen *et al.*, 2000; Monteiro, 2003].

Smith (1998) usou análise de variância (ANOVA) para avaliar a altura da cutícula, sua posição e quantidade, em fibras capilares a partir de imagens AFM. You & Yu (1997) utilizaram os descritores *step height* e rugosidade para uma avaliação quantitativa de fibras capilares submetidas a diferentes valores de pH em soluções de tampão fosfato e em diferentes tempos. Estes autores também investigaram o efeito do aquecimento das fibras.

Monteiro (2003) utilizou o cálculo de parâmetros como rugosidade, distância superficial e altura da cutícula[16] para caracterizar fibras submetidas a diferentes tratamentos cosméticos e fibras de diferentes etnias. Feughelman & Willis (2001) utilizaram a estimativa do ângulo entre a superfície e o eixo da fibra para avaliar os efeitos da temperatura e umidade, e observaram que ocorre maior abertura das cutículas com o aumento da temperatura e da umidade. Estes efeitos da umidade foram também observados por O´Connor *et al.* (1995), que acompanharam a cinética de hidratação de fibras capilares por meio da variação da altura das cutículas estimada a partir de imagens AFM e observaram que a saturação da hidratação da fibra ocorre nos primeiros 2 minutos.

---

[15] Cf. seção 2.1.

[16] Para uma descrição mais detalhada sobre a distância superficial e altura da cutícula, confira o Capítulo 2.

1.8 Descrição dos Dados

Neste trabalho foram utilizados dois conjuntos de dados formados por imagens de microscopia de força atômica de amostras de fibras capilares.

Fibras submetidas ao tratamento cosmético com peróxido: este conjunto é formado por imagens de fibras capilares do tipo caucasiano coletadas nas regiões da ponta e da raiz e submetidas a tratamento com H_2O_2, persulfato de amônio e hidróxido de amônio. Imagens de fibras não submetidas ao tratamento foram utilizadas como controle (Monteiro, 2003). O conjunto contém fibras de 4 classes diferentes: 14 amostras de cabelo retiradas da ponta, sem tratamento (controle); 09 amostras de cabelo retiradas da ponta e descoloridas; 06 amostras de cabelo retiradas da raiz, sem tratamento (controle); e 11 amostras de cabelo retiradas da raiz e descoloridas. Cada imagem de cabelo apresenta-se na forma de uma matriz **X** de 256 x 256 pixels. Estas imagens foram empilhadas, gerando um arranjo cúbico I x J x K, em que I = 40 amostras, $J = K = 256$ pixels.

A Figura 1.12 apresenta imagens de duas amostras de cada classe das fibras estudadas, onde é possível observar o aspecto característico de cada uma delas. As amostras da classe raiz controle apresentam cutículas mais fechadas e preservadas, ao contrário das amostras ponta descolorida, cujas cutículas são mais degradadas e abertas. As amostras ponta controle e raiz descolorida apresentam um aspecto intermediário entre as outras duas classes. O tratamento com peróxido está descrito em detalhes no item 1.9 (Metodologias Analíticas).

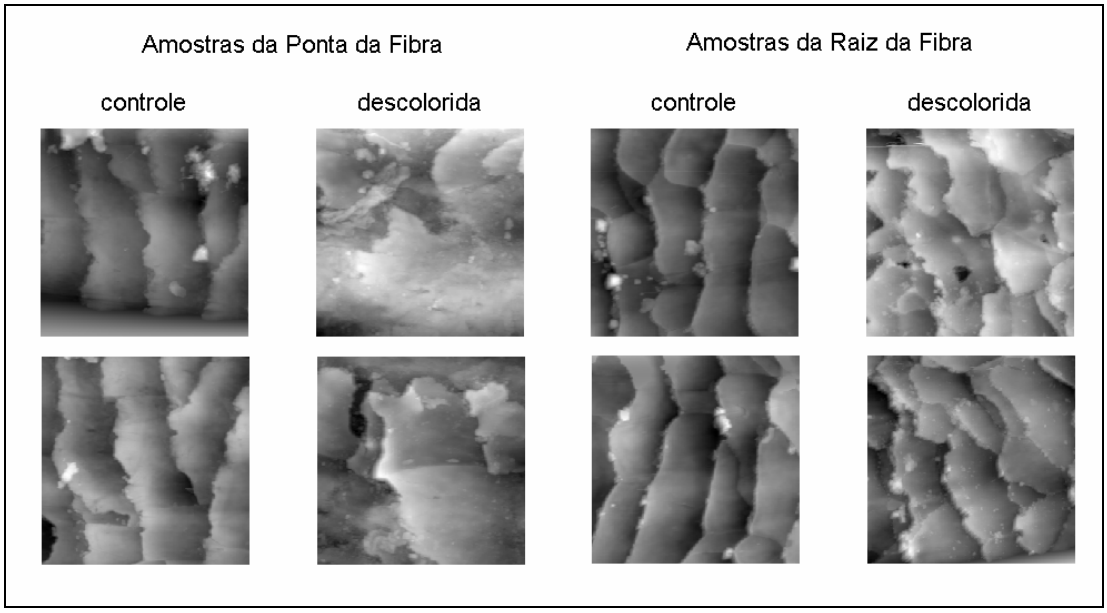


Figura 1. 11: Imagens AFM de Fibras capilares submetidas ao tratamento cosmético com H_2O_2 e perssulfato de amônio.

Fibras provenientes de diferentes etnias: foram utilizadas fibras caucasianas, fibras de afrodescendentes para representar a classe do tipo negróide e fibras orientais, para representar a classe do tipo mongol. Este conjunto é formado por imagens de fibras capilares de três etnias, sendo: 11 amostras de fibra caucasiana, 13 amostras de fibra de afrosdescendente e 12 amostras de fibra de cabelo oriental, coletadas nas regiões próximas à raiz das fibras. Cada imagem apresenta-se na forma de uma matriz **X** de 256 x 256 pixels. Estas imagens foram empilhadas, gerando arranjo cúbico I x J x K, em que I = 36 amostras, J = K = 256 pixels.

A Figura 1.13 apresenta imagens de duas amostras de cada classe das fibras estudadas, onde é possível observar o aspecto característico de cada uma delas.

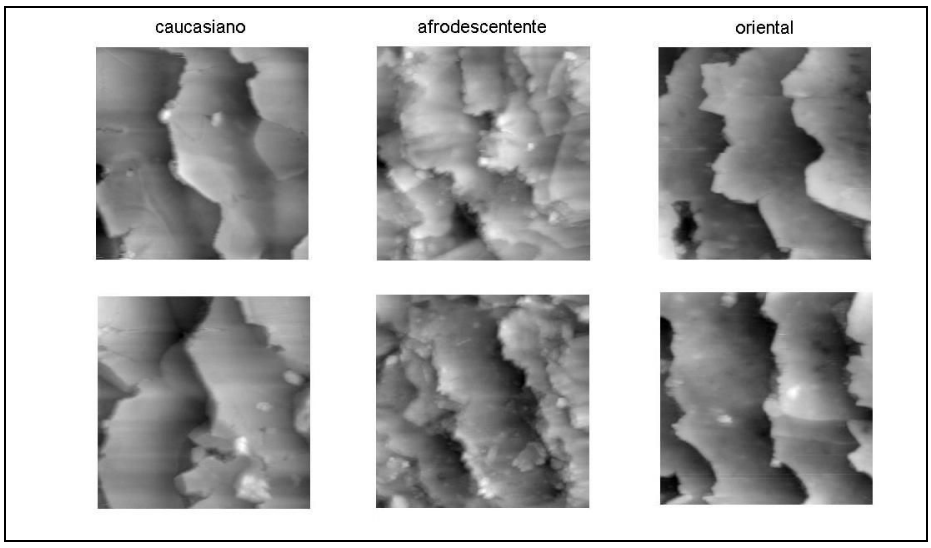


Figura 1. 12: Imagens AFM de fibras capilares de diferentes etnias.

1.9 Metodologia Analítica[17]

Neste estudo foram utilizadas amostras de três tipos de fibras capilares obtidos de De Meo Brothers, New York, USA, e acondicionadas em condições ambiente: (i) cabelo caucasiano na cor preta, (ii) cabelo oriental e (iii) cabelo de afrodescendentes. Com o intuito de simular as condições em que os cabelos são normalmente manipulados, seja em casa ou nos salões de beleza, as amostras foram submetidas aos seguintes tratamentos:

Lavagem: todas as amostras foram inicialmente lavadas com uma solução de LESS (lauril eter sulfato de sódio) a 10%, que é semelhante à formulação encontrada em xampus, e sem agitação. Na sequência as

[17] A parte analítica foi realizada pela Dr[a.] Valéria F. Monteiro e faz parte da pesquisa desenvolvida durante seu doutorado, sob orientação do Prof. Dr. Élson Longo do Laboratório Interdisciplinar de Eletroquímica e Cerâmica (LIEC), do Departamento de Química da Universidade Federal de São Carlos (UFSCAR). Para detalhes, consultar referência MONTEIRO, 2003.

amostras foram submetidas ao enxágue com água destilada, e secas à temperatura ambiente. Esta etapa é extremamente importante, pois ao manusear o cabelo para produzir as mechas para os testes, estas podem ser contaminadas por gordura e poeira, comprometendo os resultados [Monteiro, 2003; Gurden *et al.*, 2004].

Descoloração: para as amostras submetidas à oxidação, foi utilizada uma mistura na proporção de 2:1:0,5 em massa de uma solução descolorante contendo 6% de peróxido de hidrogênio (H_2O_2), uma solução concentrada de hidróxido de amônio e persulfato de amônio em pó. As mechas foram colocadas na solução descrita acima por 30 minutos, sem agitação e à temperatura ambiente, e posteriormente submetidas aos mesmos procedimentos de lavagem e secagem a qual as amostras não tratadas foram submetidas [Monteiro, 2003; Gurden *et al.*, 2004].

Análises AFM: segmentos de fibras de aproximadamente 1,00 cm de comprimento retirados das regiões da ponta ou raiz, conforme a região que se desejava amostrar, foram cuidadosamente cortados e imobilizados sobre o suporte amostrador com auxílio de fita adesiva dupla-face. Estas fibras foram examinadas por microscopia de força atômica (AFM), utilizando um instrumento Digital Instruments NanoScope IIIa, sob condições atmosféricas a 25°C utilizando uma força de carregamento de 3.6nN (Monteiro, 2003; Gurden *et al*., 2004). As imagens foram obtidas na região próxima ao centro do eixo da fibra, para evitar efeitos indesejáveis devido à sua curvatura, com dimensões de 30μm x 30μm (256 x 256 pixels) e de 20μm x 20μm (512 x 512 pixels) para os conjuntos de dados de tratamento com peróxido e etnia, respectivamente.

As imagens geradas encontram-se disponíveis como arquivos do NanoScope versão 4.43, e estes arquivos foram lidos no ambiente computacional MatLab 6.5 (Mathworks Inc., 2002), onde as análises quimiométricas foram realizadas.

1.10. Organização dos Dados e Estratégias de Análise

O conjunto de dados é formado por imagens nível de cinza de microscopia de força atômica de amostras de fibras capilares de diferentes classes. Para a análise destas imagens foram adotadas quatro principais estratégias, duas delas baseadas na construção de descritores para as fibras capilares, e outras duas pela análise quimiométrica direta das imagens.

O cálculo de descritores para fibras capilares a partir de imagens de microscopia tem sido amplamente detalhado na literatura científica [Sauermann *et al.*, 1988; O´Conner *et al.*, 1995; Smith, 1998]; McMullen & Kelty, 2001; Feughelman & Willis, 2001]. Recentemente, Gurden *et al.* (2004) demonstraram o potencial do uso destes descritores em conjunto com o método dos mínimos quadrados parciais (PLS[18]), na discriminação de diferentes classes de fibras capilares. O método PLS é adequado para a análise de arranjos de dados do tipo matricial **X** (*I*, *J*), em que *I* são as amostras do conjunto, e *J* são as variáveis. O presente trabalho pretende investigar o desempenho de outros métodos quimiométricos em conjunto com o cálculo dos descritores, em

[18] *Partial Least Squares* (Cf. Apêndice).

particular os métodos de ordem superior, PARAFAC, MPCA[19] e NPLS, que permitem a análise de arranjos de dados tridimensionais do tipo **X** (*I*, *J*, *K*), em que *I* são as amostras do conjunto, e *J* e *K* são as variáveis. Os descritores utilizados no presente trabalho foram calculados com o algoritmo desenvolvido por Gurden *et al.* (2004).

Ainda fazendo uso dos métodos de ordem superior, o enfoque principal deste estudo de caso é desenvolver uma metodologia para a discriminação das amostras de fibras capilares diretamente a partir das imagens AFM, sem a necessidade do cálculo dos descritores. Uma vez desenvolvida, esta metodologia poderá ser aplicada para o estudo de qualquer tipo de imagem nível de cinza.

Todos os cálculos foram realizados no ambiente computacional MatLab 6.5 (Mathworks Inc., 2002) com auxílio dos pacotes PLS Toolbox 3.0 (Eingenvector Inc., 2003) e Nway Toolbox 2.11 (Andersson & Bro, 2000).

Segue abaixo uma breve descrição das estratégias de análise que foram utilizadas neste trabalho.

1.10.1 Cálculo dos Descritores & Análise com Método de Primeira Ordem: SIMCA

A Figura 1.14 representa o córtex central da fibra capilar recoberto pelas camadas de cutículas, que se sobrepõem desde a raiz até a ponta [Robbins, 1994]. A partir das imagens de AFM, foram gerados 10 descritores[20] que sumarizam as características mais importantes da fibra capilar (Gurden *et al.*, 2005). Estes descritores são estimados ao longo

[19] *Multilinear Principal Component Analysis* (Cf. Apêndice).

de toda a superfície da imagem e para esta análise, foram utilizadas a média e o desvio padrão de cada descritor, fornecendo um total de 20 variáveis. O desvio padrão foi utilizado com o intuito de inserir na descrição dos dados a dispersão dos valores dos descritores para cada classe de amostra.

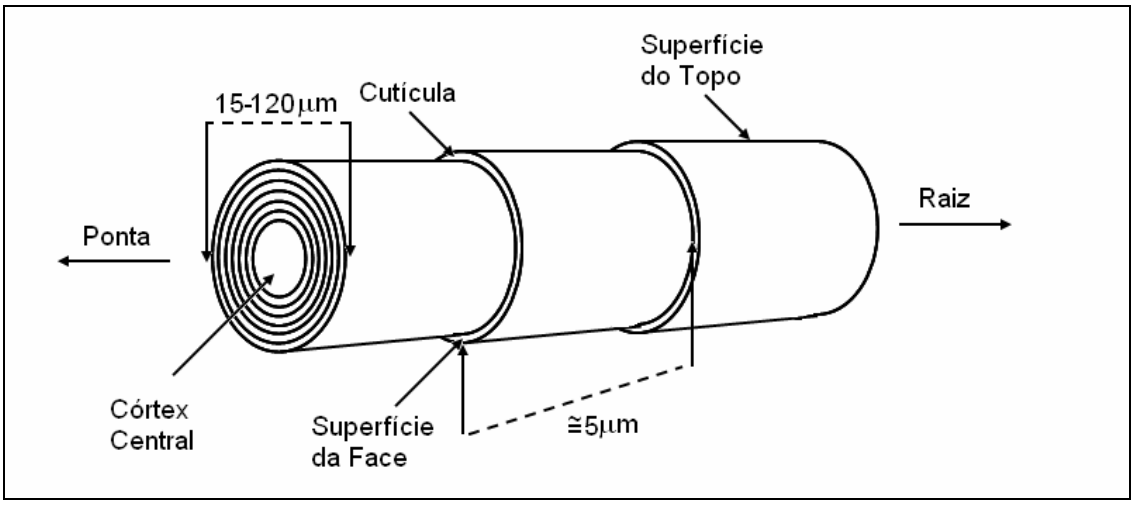


Figura 1. 13: Esquema de uma fibra capilar mostrando a disposição das cutículas.

Estes dados foram ajustados numa matriz de dados **X** (*I*, *J*), com *I* = número de amostras de cabelo e *J* = 20 descritores calculados a partir das imagens. Este arranjo foi analisado utilizando o método de reconhecimento de padrões SIMCA[21] [Wold, 1976].

[20] A estimativa dos descritores será apresentada detalhadamente no Capítulo 2.
[21] A sigla tem origem no nome em inglês *Soft Independent Modeling of Class Analogy* e ainda não há tradução oficial para o Portugês (Cf. Apêndice).

1.10.2 Cálculo dos Descritores & Análise com Método Multi-way: NPLS

O cálculo de descritores é baseado no número de cutículas identificadas na imagem. Fibras capilares muito degradadas apresentam poucas cutículas e em alguns casos isto pode levar a uma estimativa do valor de alguns descritores que não representa realmente as características da fibra, pois somente a média e o desvio padrão são considerados para cada imagem. Como alternativa a este problema, foram geradas matrizes de descritores para cada imagem de cabelo, do tipo **X** (*J*, *K*) em que *J* = 256, o número de linhas na matriz-imagem original, e *K* = 16, correspondendo aos 10 descritores, mais o desvio padrão dos descritores de 1 a 6 para cada linha *I*, num total de 16 variáveis[22]. As matrizes de descritores de 256 x 16 foram empilhadas gerando um arranjo cúbico $\underline{\mathbf{X}}$ (*I*, *J*, *K*), com *I* = número de amostras, *J* = 256 linhas (correspondendo aos 256 pixels da imagem original) e *K* = 16 descritores. Desta maneira, espera-se obter uma maior representatividade das características da fibra ao longo de toda a imagem. O método NPLS foi utilizado para discriminar as classes de fibras capilares a partir das informações contidas nos descritores [Bro, 1996, 1997, 1998].

1.10.3 Imagem & Método Multi-way: NPLS

As imagens nível de cinza podem ser arranjadas na forma matricial **X** (*J*, *K*), em que *J* = *K* = número de pixels da imagem, que pode ser 128,

[22] Os detalhes da estimativa destes descritores serão apresentados no Capítulo 2.

256, 512 etc. Um arranjo cúbico do tipo $\underline{\mathbf{X}}$ (I, J, K) foi montado, em que I = número de amostras e J = K = 256 pixels e o método NPLS foi utilizado para identificar as classes de amostras [Bro, 1996, 1997, 1998].

1.10.4 Imagem & Método Multi-way: PARAFAC e MPCA

Neste caso, o arranjo de imagens $\underline{\mathbf{X}}$ (I, J, K) com I = número de amostras e J = K = 256 pixels foi modelado utilizando análise de fatores paralelos, conhecida pela sigla PARAFAC, para identificar as classes de amostras. A análise de componentes principais multilinear, ou MPCA, foi utilizada para classificar novas amostras [Bro, 1997 & 1998, Henrion *et al.*, 1992].

CAPÍTULO 2

Trabalhando com os Descritores: – Métodos de Primeira Ordem e Métodos de Ordem Superior

Efeito de Tratamento Cosmético: Discriminação Quantitativa de Imagens de Fibras Capilares Submetidas ao Tratamento de Descoloração

2.1 Cálculo dos Descritores

A classificação das imagens AFM das amostras de cabelo foi primeiramente executada pela caracterização da superfície capilar utilizando descritores que sumarizam as informações mais importantes da estrutura da cobertura de cutículas. A estimativa destes descritores foi realizada após uma etapa prévia de *planificação* da superfície da fibra e depois pelo ajuste de funções adequadas ao seu *perfil superficial longitudinal*.

2.1.1 Planificação

A Figura 2.1(a) apresenta uma das imagens utilizadas neste estudo e a Figura 2.1(b) mostra a superfície desta fibra, onde é possível notar a existência de um ângulo lateral ao longo do eixo central da fibra capilar. Cada fibra apresentará inclinação característica, e para que a estimativa dos descritores seja realizada de maneira uniforme para as diferentes

fibras, é necessário que este ângulo seja corrigido, o que é feito por meio da subtração da superfície de fundo estimada, numa etapa prévia denominada *planificação*. Esta superfície de fundo é estimada pelo ajuste de um polinômio de duas variáveis, no qual um polinômio de primeira ordem é utilizado na direção do eixo *x* e um polinômio de ordem 3 é utilizado na direção do eixo *y* [Gurden *et al.*, 2004]. A Figura 2.1(c) apresenta a superfície de fundo da fibra, e a Figura 2.1(d) mostra a superfície da fibra após a subtração desta superfície de fundo.

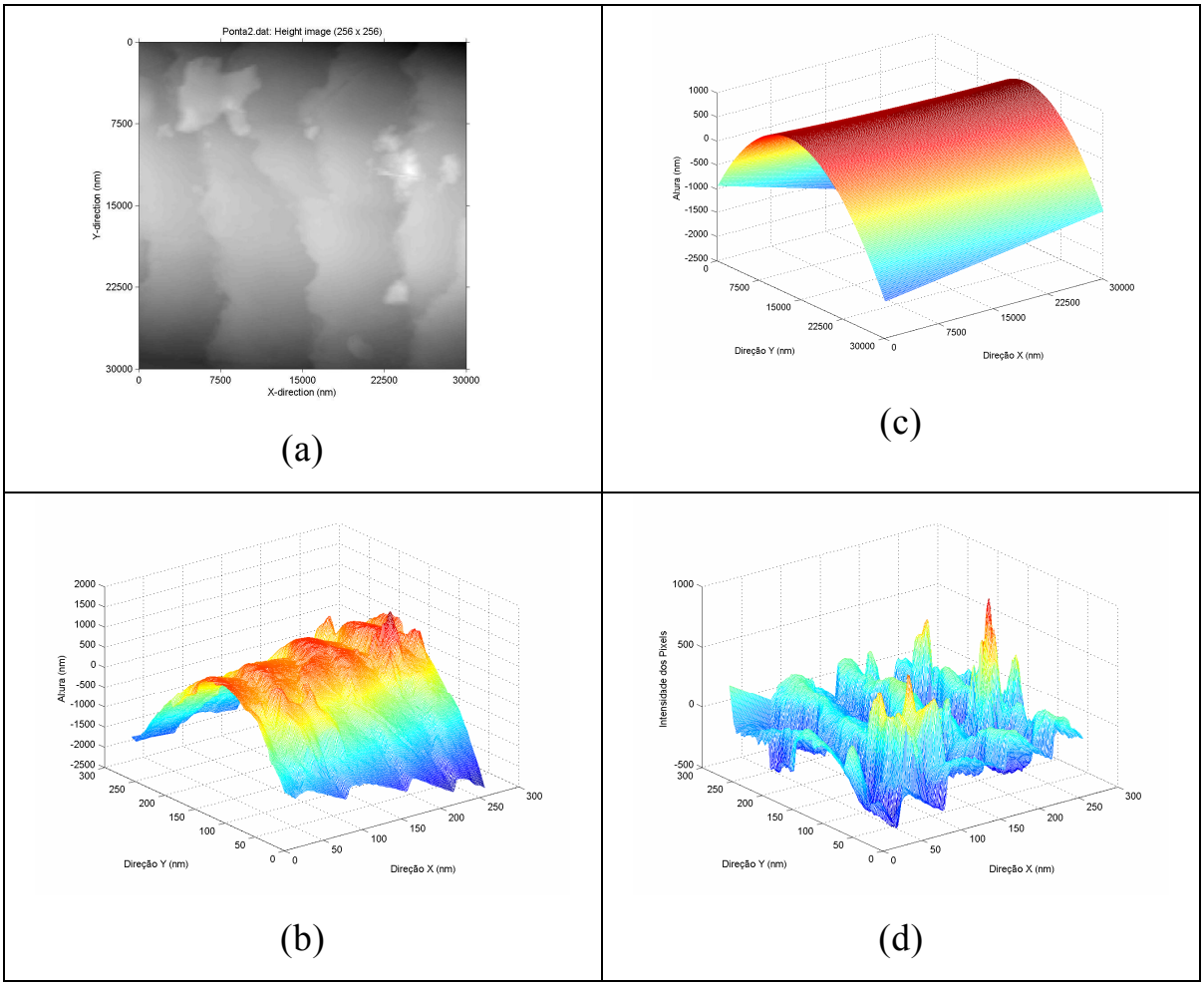


Figura 2. 1: (a) imagem AFM de uma das fibras capilares utilizadas no estudo, (b) superfície da fibra antes da planificação, (c) superfície de fundo e (d) superfície da fibra após a etapa de planificação.

2.1.2 O Perfil da Superfície da Fibra e a Estimativa dos Descritores

O algoritmo desenvolvido por Gurden *et al.* (2004) estima os descritores capilares baseando-se no *perfil superficial longitudinal* da fibra. Cada imagem é formada por uma matriz **X** (*I*, *J*), cujas coordenadas *i* e *j* representam as coordenadas geométricas dos pixels. Cada linha desta matriz apresenta o *perfil superficial longitudinal* da fibra naquela posição da matriz. Um exemplo deste perfil pode ser observado na Figura 2.2, em que é possível reconhecer o padrão da estrutura das cutículas, com a sobreposição das cutículas adjacentes. É possível observar claramente neste perfil o ponto em que uma cutícula termina e a cutícula subjacente começa a surgir.

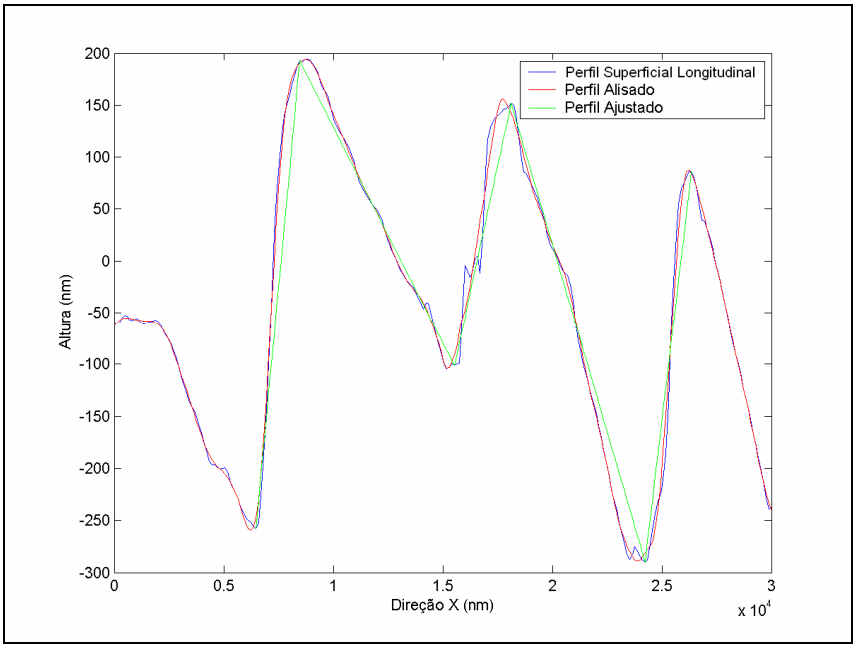


Figura 2. 2: Perfil longitudinal da superfície da fibra e ajustes realizados para a estimativa dos descritores.

O ponto de partida para o cálculo dos descritores utilizados neste estudo é a estimativa da altura das cutículas, que é realizada com o uso da primeira derivada do *perfil superficial longitudinal* (Figura 2.2) para localização dos pontos de mínimo e de máximo. O ponto de mínimo representa a posição em que uma cutícula termina, e a cutícula adjacente começa a emergir na superfície da fibra (pontos A e C na Figura 2.3), e o ponto de máximo representa o topo da cutícula (ponto B na Figura 2.3). Após localizar estes pontos, é possível estimar a altura da cutícula, e todos os outros descritores para a fibra capilar.

No entanto, a superfície das cutículas apresenta irregularidades que poderão ser identificadas erroneamente como pontos de mínimo e máximo. Para evitar a influência destas irregularidades e garantir a correta localização destes pontos, foi realizado previamente um alisamento no *perfil da superfície longitudinal* pelo uso do algoritmo de Savitsky-Golay [Savitzky & Golay, 1964] com polinômio de segunda ordem e tamanho de janela igual a 2000 nm. Este algoritmo é frequentemente utilizado como filtro em dados espectroscópicos para redução do ruído da linha de base causado por pequenas variações do sinal analítico [Beebe, 1998]. O uso do filtro de Savitsky-Golay gera um *perfil alisado* (Figura 2.2), e o uso da primeira derivada neste *perfil alisado* permitirá encontrar os pontos de mínimo e de máximo, e do *perfil ajustado* (Figura 2.2).

A partir destes perfis é possível calcular 10 descritores para as fibras capilares: *step height*, *tilt*, *backtilt*, *layer spacing*, *face distance*, *top*

distance, *fit error*, *cuticle density*, *roughness*, *fitability*[23] cujos detalhes podem ser observados na Figura 2.3.

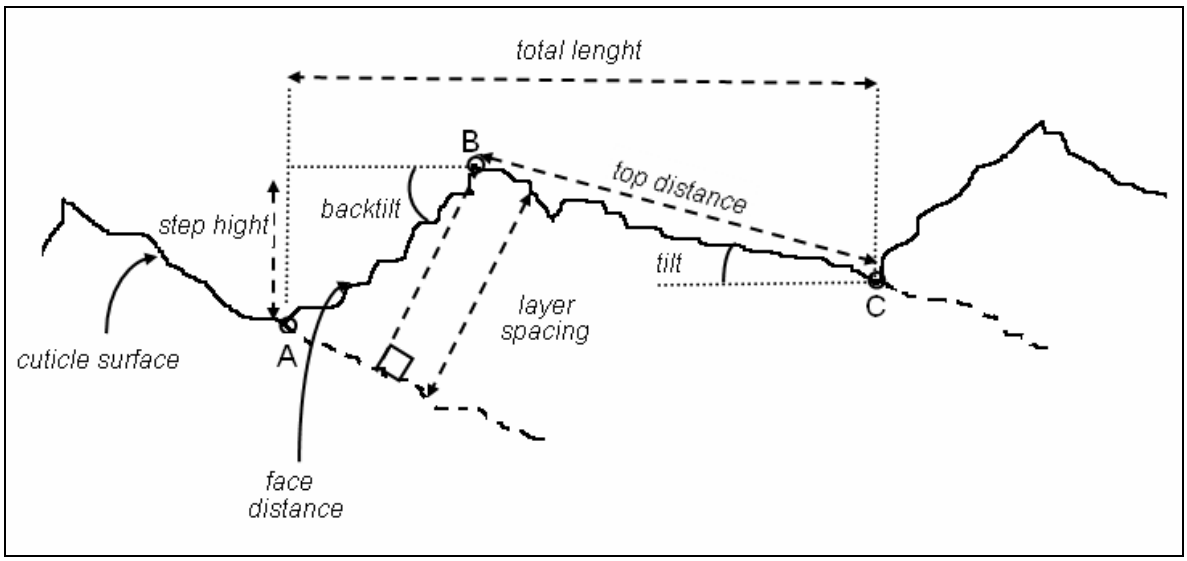


Figura 2. 3: Representação do perfil longitudinal da superfície capilar mostrando alguns dos descritores utilizados na caracterização da fibra [Adaptado de Gurden *et al.*, 2004].

1) Step Height: este descritor representa a distância vertical entre o topo da cutícula (ponto B na Figura 2.3) e o ponto onde a cutícula subjacente começa a surgir (ponto A). É um parâmetro quantitativo importante para avaliar os efeitos de tratamentos químicos e físicos na superfície das fibras capilares, e é a base para a estimativa de outros descritores utilizados neste trabalho [O´Connor *et al.*, 1995; You & Yu, 1997; Swift & Brown, 1972; Smith, 1998; Swift & Smith, 2000; Monteiro, 2003; Gurden *et al.*, 2004]. Valores típicos de *Step Height* encontram-se entre 300 e 500 nm, mas podem variar devido aos efeitos de doenças, meio

[23] Não foi encontrada na literatura científica brasileira a nomenclatura em português para os descritores capilares. Com o intuito de evitar traduções livres que possam gerar confusões futuras na interpretação dos descritores, decidiu-se neste trabalho manter os nomes em inglês encontrados na literatura especializada.

ambiente, tratamentos cosméticos e etnia [You & Yu, 1997; Smith, 1998].

2) Tilt: ângulo entre a superfície do topo da cutícula e o eixo da fibra [Swift & Smith, 2000; Feughelman & Willis, 2001; Gurden *et al.*, 2004].

3) Backtilt: ângulo entre a face da cutícula (entre os pontos A e B) e o eixo da fibra [Gurden *et al.*, 2004].

4) Layer spacing: é a distância mínima entre o topo da cutícula no seu ponto final e o topo da cutícula a ela adjacente. Fibras deterioradas costumam apresentar cutículas soltas ou quebradas na região do topo (neste ponto final), permitindo a existência de uma camada de ar ou água entre uma cutícula e outra. Neste caso, o cálculo deste descritor deve consider a distância total entre as duas cutículas subjacentes, e não apenas a espessura absoluta da cutícula [Gurden *et al.*, 2004].

5) Face distance: é a distância entre a superfície do topo da cutícula e o ponto onde a cutícula adjacente começa a emergir (distância entre os pontos A e B na Figura 2.3) [Gurden *et al.*, 2004].

6) Top distance: é a distância entre o topo de uma cutícula e o ponto onde ocorre o final da face da cutícula sobreposta a ela (distância entre B e C na Figura 2.3) [Monteiro, 2003; Gurden *et al.*, 2004].

7) Fit error: é uma medida da adequação do *perfil ajustado* (perfil A-B-C) ao *perfil da superfície longitudinal*, ou seja, é o erro de ajuste dado pela raiz quadrada do quadrado das somatórias das diferenças entre o *perfil ajustado* e *o perfil da superfície longitudinal* (Equação (5), em que N = número de pixels na direção do eixo x). Um alto valor deste descritor significa um alto grau de irregularidades na superfície da

cutícula devido, por exemplo, à formação de estrias e rugosidades [Swift & Smith, 2000; Gurden *et al.*, 2004].

$$Fit\ error = \sqrt{\frac{\sum (x_i - \hat{x}_i)^2}{N}} \qquad (5)$$

8) Cuticle density: número de cutículas por milímetro [O´Connor *et al.*, 1995; Gurden *et al.*, 2004].

9) Roughness: é uma medida de aspereza total do *perfil superficial longitudinal* da fibra. A rugosidade é estimada pela divisão do comprimento total do perfil da fibra pelo comprimento da imagem. Desta forma, um valor de rugosidade igual a 1 indica uma superfície completamente lisa. Este valor aumentará com o aumento do número e da altura das cutículas por fibra [You & Yu, 1997; Monteiro, 2003; Gurden *et al.*, 2004].

10) Fitability: é uma medida do quanto o padrão apresentado pelo *perfil superficial longitudinal* da fibra se assemelha ao padrão de uma ou mais cutículas. A maioria das imagens consiste de 3 a 6 cutículas sobrepostas, mas fibras muito deterioradas apresentam cutículas quebradas e soltas, o que torna difícil o reconhecimento do padrão da cutícula no perfil da superfície da fibra, e consequentemente resulta em um mau ajuste do algoritmo para o cálculo dos descritores. A *Fitability* é dada pela porcentagem do perfil da superfície da fibra que apresenta um bom ajuste pelo algoritmo [Gurden *et al.*, 2004].

A caracterização da estrutura da fibra foi feita pela estimativa do valor médio e desvio padrão encontrado para cada um destes descritores em cada imagem. A Tabela 2 contém os valores encontrados para os

descritores estimados para as fibras capilares apresentadas na Figura 1.12. É possível notar como os desvios são grandes, demonstrando a heterogeneidade característica das fibras capilares.

Tabela 2: Descritores calculados para as fibras capilares mostradas na Figura 1.12.

(No.) Descritor	Ponta Controle**	Ponta Descolorida**	Raiz Controle**	Raiz Descolorida**
(1) Step height	382.1 ± 162.10nm 383.06 ± 101.74nm	369.57 ± 105.41nm 539.7 ± 232.87nm	369.62 ± 118.40nm 386.27 ± 103.89nm	363.94 ± 101.23nm 476.43 ± 174.76nm
(11) DP*				
(2) Tilt	4.26 ± 3.54º	3.57 ± 2.00º	3.31 ± 1.64º	3.22 ± 2.35º
(12) DP*	3.70 ± 1.17º	2.74 ± 1.84º	3.72 ± 1.17º	4.36 ± 1.98º
(3) Backtilt	13.49 ± 7.83º	12.48 ± 8.09º	19.01 ± 8.01º	16.34 ± 7.66º
(13) DP*	14.89 ± 6.92º	15.82 ± 10.20º	19.83 ± 6.55º	16.87 ± 8.96º
(4) Layer spacing	515.03 ± 191.26nm 520.00 ± 153.86nm	521.64 ± 218.23nm 570.05 ± 223.40nm	417.83 ± 100.80nm 473.43 ± 147.40nm	438.53 ± 168.75nm 551.11 ± 183.99nm
(14) DP*				
(5) Face distance	1937.6 ± 746.4nm 1862.1 ± 826.2nm	2412.0 ± 1401.3nm 2716.10 ± 1628.0nm	1365.7 ± 461.8nm 1451.7 ± 708.7nm	1784.0 ± 1035.4nm 2128.6 ± 1270.4nm
(15) DP*				
(6) Top distance	5741.9 ± 1933.7nm 6176.5 ± 1729.3nm	6720.6 ± 3144.3nm 7087.1 ± 4586.8nm	6253.1 ± 2020.0nm 5955.1 ± 1471.2nm	6627.0 ± 3033.0nm 6212.6 ± 2520.1nm
(16) DP*				
(7) Fit error	39.07 ± 20.29nm 37.89 ± 10.86nm	48.58 ± 22.24nm 65.08 ± 31.22nm	34.11 ± 19.89nm 31.78 ± 10.12nm	44.69 ± 14.57nm 53.64 ± 16.12nm
(17) DP*				
(8) Cuticle density	112.86 ± 28.14/mm 127.73 ± 15.31/mm	69.17 ± 32.94/mm 66.67 ± 24.68/mm	121.35 ± 23.89/mm 129.43 ± 20.93/mm	105.60 ± 25.38/mm 102.08 ± 21.79/mm
(18) DP*				
(9) Roughness	1.02 ± 0.01 1.01 ± 0.003	1.012 ± 0.006 1.016 ± 0.008	1.015 ± 0.006 1.015 ± 0.005	1.01 ± 0.004 1.02 ± 0.010
(19) DP*				
(10) Fitability	66.88 ± 15.08% 81.44 ± 8.02%	40.69 ± 24.61% 41.42 ± 21.55%	70.88 ± 12.02% 75.17 ± 13.05%	66.28 ± 16.74% 63.70 ± 13.57%
(20) DP*				

* DP = desvio padrão dos valores estimados do descritor para toda a imagem; ** cada célula da tabela contém os valores para as duas amostras de cada classe apresentadas na Figura 1.12.

2.2 Análise dos Dados com Método de Primeira Ordem

2.2.1 Arranjo de Dados

O conjunto de dados utilizado é formado por um arranjo de dados **X** (40, 20) contendo as 40 amostras de fibras capilares e os 20 descritores calculados para cada imagem (veja Tabela 2). Os dados foram modelados utilizando o método de reconhecimento de padrões SIMCA[24] [Wold, 1976; Beebe, 1998].

2.2.2 Análise dos Dados

O conjunto de dados foi dividido em dois subconjuntos, um contendo 32 amostras, utilizado para a construção do modelo de classificação, e outro com 8 amostras, sendo duas amostras de cada classe, utilizadas como conjunto de predição.

Modelagem

Os descritores estimados para as fibras capilares são de diferentes escalas e unidades, como nm, graus e % (Tabela 2) e para que todas as variáveis tenham a mesma contribuição na construção do modelo SIMCA os dados foram autoescalados[25] antes da modelagem.

---

[24] Cf. Apêndice.
[25] Autoescalamento = cada variável é centrada na média, e dividida pelo respectivo desvio padrão, de forma que todas as variáveis apresentarão desvio padrão unitário [Beebe, 1998].

O método SIMCA utiliza análise de componentes principais (PCA[26]) para modelar o espaço ocupado pelas amostras de cada classe. Um modelo PCA é construído para cada classe de amostras, cujos escores são utilizados para a construção de caixas multidimensionais que delimitam o espaço ocupado individualmente pelas classes. A classificação de amostras novas é realizada pela projeção dos seus escores dentro dos limites destas caixas, o que permite determinar a qual das caixas a nova amostra pertence [Beebe, 1998].

O modelo SIMCA construído apresentou três amostras alocadas em classes erradas, as amostras 9, 21 e 22. A Figura 2.4 contém as imagens AFM destas fibras, onde é possível observar que a amostra 9 (ponta controle, classe 1) apresenta uma grande deformação nas cutículas no centro da figura. Este tipo de deformação é pouco comum nas amostras da sua classe, o que contribui para que esta amostra apresente um comportamento diferenciado.

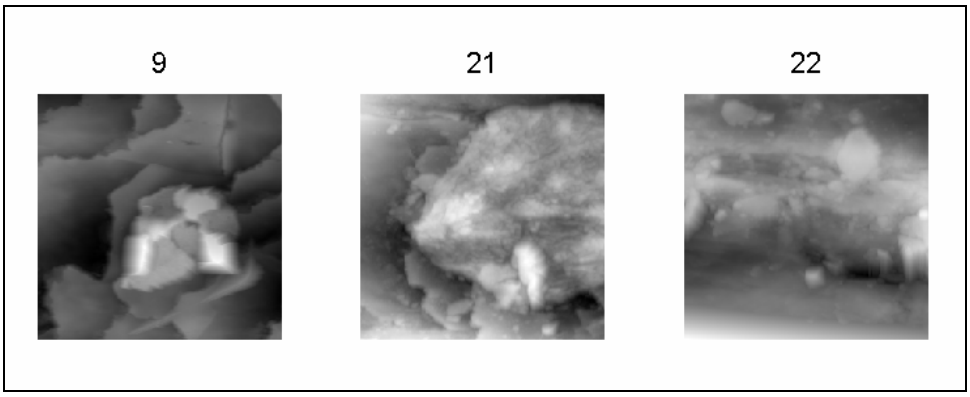


Figura 2. 4: Amostras atípicas detectadas pelo método SIMCA: 9 (ponta virgem), 21 (ponta descolorida), (c) 22 (ponta descolorida).

[26] Do inglês, *Principal Component Analysis* (Cf. Apêndice).

As amostras 21 e 22 pertencem à classe ponta descolorida. Esta classe possui um perfil extremamente heterogêneo e representa o maior nível de degradação da fibra devido principalmente à perda de cutículas causada pelos sucessivos processos de limpeza e escovação ao longo do crescimento, intensificada pelos efeitos do tratamento de descoloração sofrido [Robbins, 1994]. Estas duas amostras particularmente apresentam regiões extremamente degradadas e com ausência de cutículas em algumas regiões das imagens, o que impede o cálculo dos descritores *step height*, *tilt*, *backtilt*, *layer spacing*, *face distance e top distance*[27] nestas regiões, pois o número de cutículas é utilizado como base para a sua estimativa. Este fato leva a uma estimativa de descritores pouco representativa para estas imagens, pois as regiões com ausência de cutículas não serão consideradas, o que poderá comprometer sua correta classificação.

Pela observação das imagens destas amostras é possível concluir que a classificação errada se deve ao seu comportamento atípico e não por falha do modelo. Estas amostras foram excluídas e o modelo SIMCA foi refeito.

A Tabela 3 apresenta o número de fatores, a porcentagem de variância descrita nos modelos PCA para cada classe de amostras e a classificação das amostras do conjunto de treinamento. A classe 4 (raiz controle) é a mais homogênea e apenas 1 fator foi suficiente para descrever 83,4% das informações desta classe. A classe 2 (ponta descolorida) é a classe mais heterogênea e 2 fatores foram necessários para a modelagem dos

[27] Cf. seção 2.1.

dados, com 87,8 % da variância descrita. Estas duas classes representam os extremos do nível de degradação da fibra.

Tabela 3: Parâmetros do modelo SIMCA com dados autoescalados.

	Classe 1 Ponta controle	Classe 2 Ponta descolorida	Classe 3 Raiz descolorida	Classe 4 Raiz controle
Número de fatores do modelo	4	2	4	1
% de variância descrita	86,7 %	87,8 %	91,9 %	83,40 %
N$^{o.}$ de amostra da classe	11	5	9	4
N$^{o.}$ de amostras classificadas corretamente	11	5	9	4
No. de amostras mal classificadas	0	0	0	0

As classes 1 (ponta controle) e 3 (raiz descolorida), representam um nível de degradação da fibra intermediário, e apresentam características similares. Para estas duas classes, os modelos com até 3 fatores apresentaram amostras com classificação errada, sendo que estas foram alocadas sempre nas classes 1 ou 3. Apenas os modelos com 4 fatores proporcionaram a correta classificação das amostras.

A Figura 2.5 apresenta o gráfico do poder discriminante das variáveis, que estima a influência de cada variável na diferenciação das classes de amostras.

É possível observar que a variável rugosidade média (*Roughness_m*) se destaca, tendo apresentado alto poder discriminante. Fibras capilares não tratadas apresentam aumento gradual na rugosidade ao longo da sua extensão, da raiz em direção às pontas, como uma consequência do processo natural de deterioração [McMullen *et al.*, 2000; McMullen & Kelty, 2001], e este descritor tem sido utilizado como um bom indicador do estado de preservação da fibra [You & Yu, 1997; Monteiro, 2003]. O uso de descolorantes intensifica o processo de degradação da estrutura

protéica da fibra, com consequente aumento na rugosidade da superfície [Robbins 1994; Swift & Smith, 2000; Monteiro, 2003].

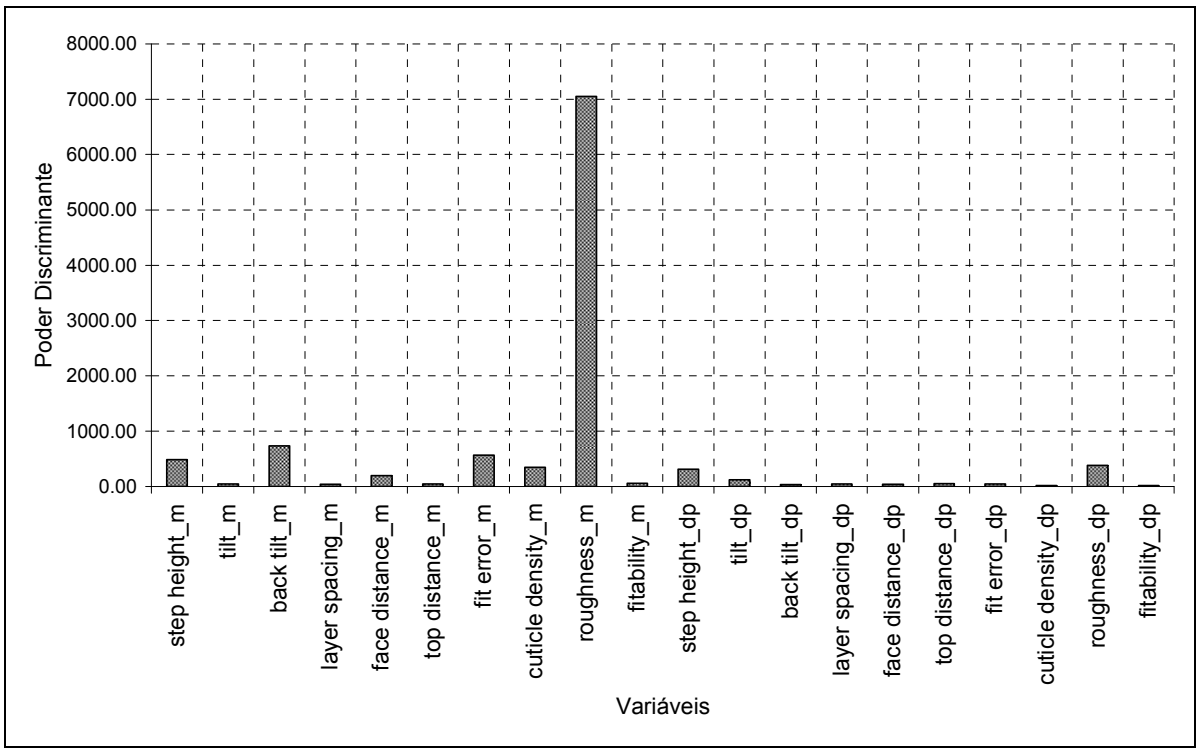


Figura 2. 5: Poder discriminante das variáveis utilizadas no modelo de classificação SIMCA.

Outras variáveis que também apresentaram maior poder discriminante (> 100) são: valor médio e desvio padrão de *step height*, desvio padrão de *tilt*, valor médio de *back-tilt*, valor médio de *face distance*, valor médio de *fit error*, valor médio de *cuticle density* e desvio padrão de *Roughness*. A variação nos valores destes descritores em função de cada classe de amostras será detalhada a seguir[28].

[28] Cf. seção 2.3.

Classificação

O modelo SIMCA com 29 amostras foi utilizado para classificar 8 amostras externas, e os resultados podem ser observados na Tabela 4, em que é possível observar que apenas uma amostra não foi classificada corretamente. A Figura 2.6 apresenta a imagem AFM da amostra 40, pertencente à classe 4, e que foi identificada como pertencente à classe ponta descolorida. Aparentemente ela não apresenta problemas nem diferenças em relação às amostras da sua classe. No entanto, todas as amostras utilizadas neste estudo foram obtidas numa área de 30 μm x 30 μm, e apenas a amostra 40 foi obtida com 20 μm x 20 μm, o que interferiu na estimativa dos descritores resultando na classificação incorreta.

A classificação incorreta da amostra 40 demonstra que o modelo não é robusto a variações na resolução das imagens.

Tabela 4: Classificação das amostras externas.

Classe 1 Ponta controle		Classe 2 Ponta descolorida		Classe 3 Raiz descolorida		Classe 4 Raiz controle	
Amostra	Classe Predita	Amostra	Classe Predita	Amostra	Classe Predita	Amostra	Classe Predita
5	1	17	2	27	3	37	4
8	1	19	2	30	3	40	**2**

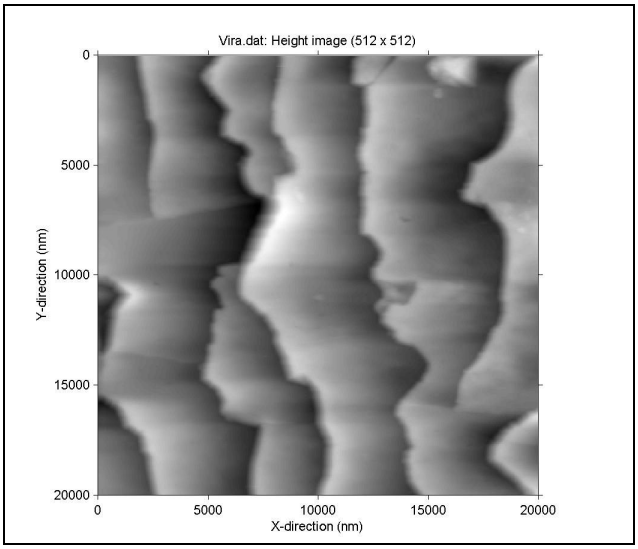


Figura 2. 6: Amostra 40 (raiz controle), que foi mal classificada pelo modelo SIMCA.

Modelos com baixa capacidade preditiva podem ser detectados quando ocorre algum dos seguintes erros de predição [Beebe, 1998]:

- número alto de amostras alocadas em classes erradas: inadequação do modelo causada por fatores como escolha inadequada do número de componentes principais e, consequentemente, altos resíduos dos modelos PCA;

- amostras classificadas em mais de uma classe: indica que o modelo contém sobreposição de classes, ou seja, os limites das caixas multidimensionais de cada classe apresentam sobreposição;

O modelo SIMCA desenvolvido não apresentou nenhum destes problemas, o que comprova sua capacidade preditiva. A distância limite para cada classe de amostras foi estabelecido com base nos parâmetros estatísticos Q e T^2 com 95% de confiança e podem ser observados na Figura 2.7 para o conjunto de treinamento.

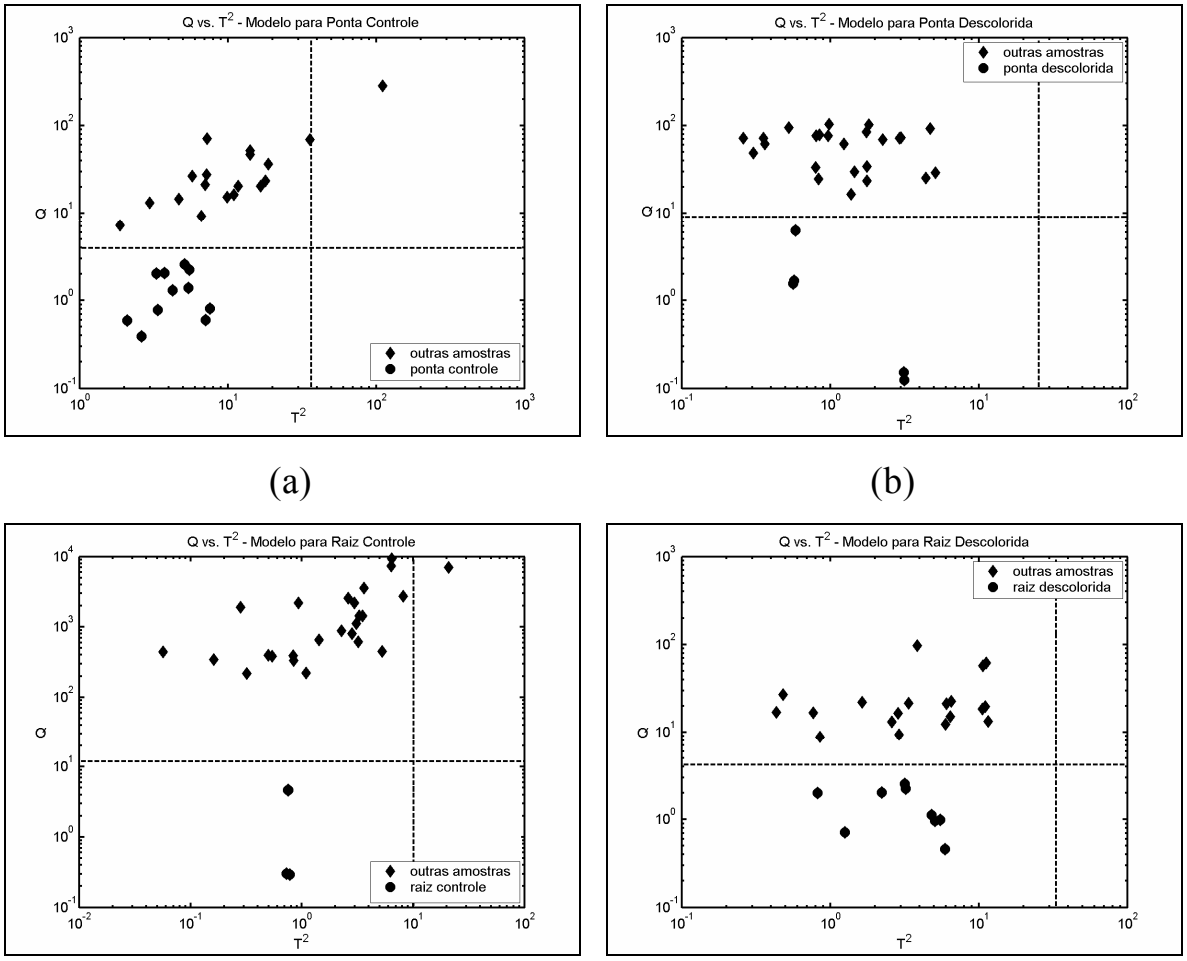


Figura 2. 7: Gráfico dos valores de Q vs. T^2 de Hotelling para cada classe do conjunto de treinamento, mostrando os seus respectivos limites de controle: (a) ponta controle, (b) ponta descolorida, (c) raiz descolorida e (d) raiz controle.

O parâmetro Q é estimado pela soma quadrática dos resíduos do modelo, e indica o quanto cada amostra está contemplada pelo modelo PCA. Q é, portanto, uma medida da diferença, ou seja, do resíduo, entre a amostra e sua projeção no espaço determinado pelas *k* componentes

principais do modelo [Wise *et al.*, 2005][29]. Um modelo adequado deverá fornecer valores baixos de Q para a classe de amostras descrita.

O parâmetro T^2 de Hotelling expressa a variação de cada amostra dentro do modelo PCA. Este parâmetro é estimado a partir da soma normalizada do quadrado dos escores e indica o quanto a amostra está próxima do centróide do conjunto de amostras da classe [Wise *et al.*, 2005]. Sendo assim, amostras pertencentes à classe modelada deverão apresentar valores baixos de T^2.

Na Figura 2.7, as amostras pertencentes a uma determinada classe deverão apresentar simultaneamente valores baixos de Q e T^2 para a modelagem PCA em questão.

2.3 Análise dos Dados com Métodos de Ordem Superior

2.3.1 Arranjo dos Dados

O cálculo dos descritores é baseado no número de cutículas identificadas na imagem. Fibras capilares muito degradadas apresentam poucas cutículas, e em alguns casos isto pode levar a uma estimativa pouco representativa de alguns descritores da fibra. Na tentativa de contornar este problema, foram geradas matrizes de descritores para cada imagem de cabelo, do tipo $\underline{\mathbf{X}}$ (I, J, K) em que J = 256, o número de linhas na matriz-imagem original, e K = 16, correspondendo aos 10 descritores, mais o desvio padrão dos descritores de 1 a 6 para cada

[29] Cf. Apêndice.

linha I, num total de 16 variáveis[30]. As matrizes de descritores de 256 x 16 foram empilhadas gerando um arranjo cúbico **X** (*I*, *J*, *K*), de *I* = 40 amostras, *J* = 256 linhas (correspondendo aos 256 pixels da imagem original) e *K* = 16 descritores (detalhes na Figura 2.8).

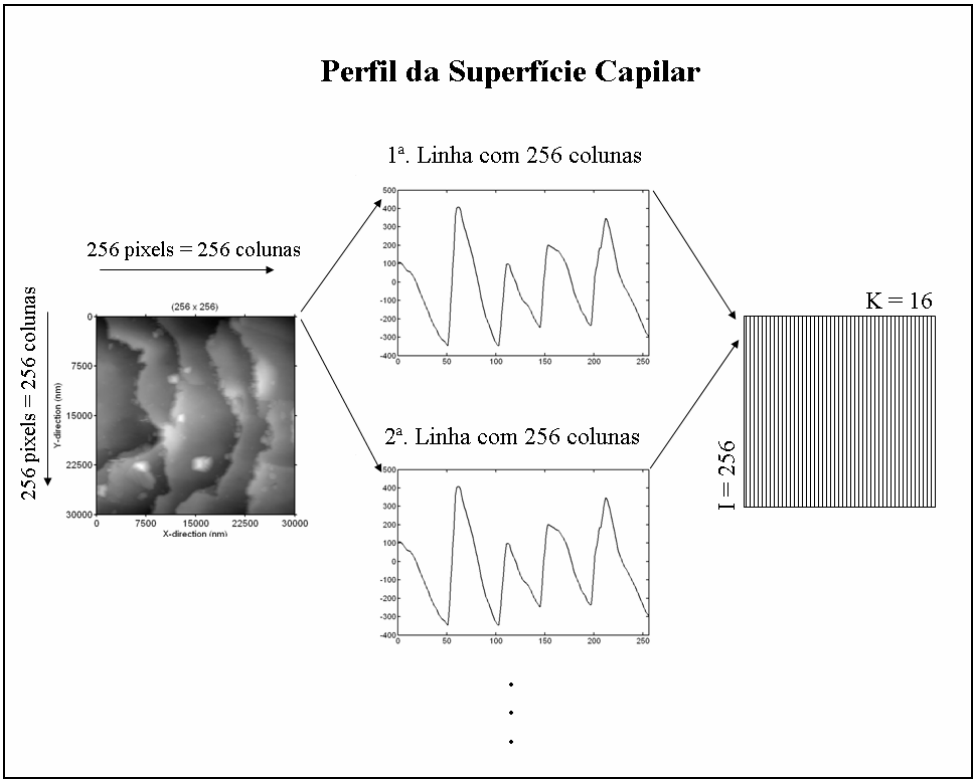


Figura 2. 8: Esquema do cálculo da matriz de descritores.

Espera-se com isto ter maior representatividade das características das fibras, pois a matriz de descritores fornecerá a distribuição dos mesmos ao longo de toda a imagem.

[30] Observando a Figura 2.8 é possível notar que dependendo do número de cutículas detectado, é possível obter mais de um valor para alguns descritores em cada linha da matriz, como: *step height*, *tilt*, *backtilt*, *layer spacing*, *face distance* e *top distance*. Para este caso utilizou-se a média e o desvio padrão dos valores obtidos para estes descritores em cada linha. A estimativa dos decritores *fit error*, *cuticle density*, *roughness* e *fitability* fornece apenas um valor por linha. Sendo assim, o número total de descritores utilizado foi 16, sendo a média e o desvio padrão dos descritores *step height*, *tilt*, *backtilt*, *layer spacing*, *face distance* e *top distance* para cada linha da imagem, e o valor dos descritores *fit error*, *cuticle density*, *roughness* e *fitability* restantes.

Com o objetivo de discriminar as quatro classes de fibras capilares a partir das informações contidas nos descritores, foi utilizado o método NPLS discriminante [Bro, 1998; Bro, 1996].

O conjunto de dados utilizado é formado por um arranjo de dados tridimensional $\underline{\mathbf{X}}$ (40, 256, 16) contendo as 40 matrizes das imagens empilhadas, e um arranjo bidimensional do tipo $\mathbf{Y}(I, J)$ com I = 40 amostras e J = 4 colunas, sendo cada uma referente a uma classe de amostra (y_1 = ponta virgem, y_2 = ponta descolorida, y_3 = raiz descolorida e y_4 = raiz virgem). A cada coluna da matriz **Y** foram atribuídos valores categóricos discretos de 0 ou 1, sendo 0 atribuído às amostras que não pertencem à categoria e 1 atribuído às amostras pertencentes à categoria. Detalhes da estrutura dos dados podem ser visualizados na Figura 2.9.

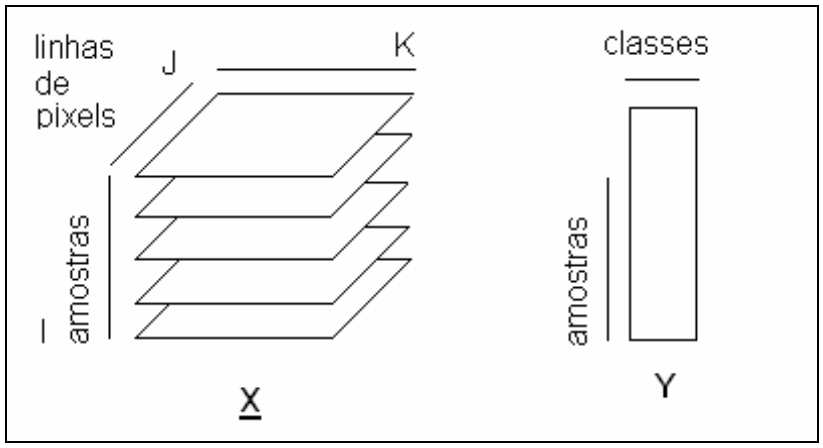


Figura 2. 9: Estrutura dos dados.

2.3.2 Análise dos Dados

Os descritores estimados para as fibras capilares são de diferentes ordens de grandeza (Tabela 2) e foi necessário escalar os dados no

Modo *K* (modo dos descritores) para desvio padrão unitário, para que todas as variáveis tenham a mesma contribuição na construção do modelo NPLS [Gurden *et al*., 2001; Bro & Smilde, 2003].

Inicialmente foram construídos modelos com até 10 fatores utilizando todas as 40 amostras. Os resultados da variância descrita em cada fator para os blocos **X** e **Y** podem ser observados na Tabela 5. A porcentagem de variância descrita no bloco **X** é alta já nos primeiros fatores, no entanto para o bloco **Y** é muito baixa, alcançando 50% somente no modelo com 6 fatores, o que pode ser um indicativo da presença de amostras anômalas [Beebe, 1998].

Tabela 5: Porcentagem de variância capturada pelo modelo NPLS.

Conjunto com 40 amostras		
Fatores	**% Variação Explicada no Bloco X**	**% Variação Explicada no Bloco Y**
1	99,84	27,10
2	99,83	28,33
3	99,83	30,99
4	99,83	46,17
5	99,84	48,96
6	**99,84**	**51,87**
7	99,86	57,58
8	99,87	63,35
9	99,87	72,41
10	99,88	79,10

Investigando as amostras atípicas

Para detectar quais amostras apresentaram comportamento atípico, foi realizado um estudo da influência de cada amostra no modelo, por meio da estimativa dos valores de *leverage* (h)[31]. Na Figura 2.10 é possível

[31] Para detalhes sobre a estimativa da *leverage*, confira o Apêndice.

observar que para a modelagem com 6 fatores, as amostras 1, 21 e 22 apresentam valores de *leverage* muito altos, e acima do valor limite estabelecido (*hcrítico* = 3*k*/*I*, em que *k* = número de fatores e *I* = número de amostras) estimado com 99% de confiança, que é de 0,450.

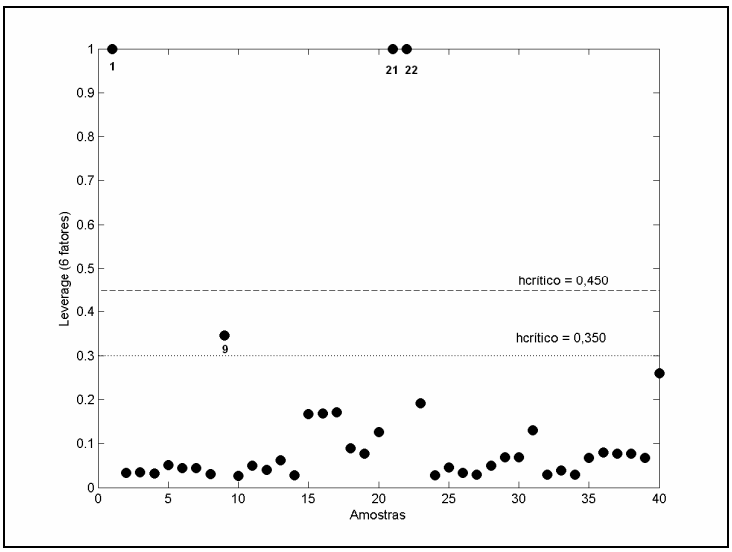


Figura 2. 10: Gráfico dos valores de *leverage* vs. amostra para os modelos NPLS com 6 fatores e dados escalados no Modo 3.

A amostra 9 ficou acima do limite estimado com 95% de confiança (hcrítico = 2*k*/*I*), e a amostra 40 apresentou valor alto de *leverage*, mas manteve-se dentro dos limites estimados. Com exceção da amostra 1, as amostras com comportamento atípico são as mesmas identificadas anteriormente nos testes com o método SIMCA (Cf. seção 2.2).

A Figura 2.11 apresenta as imagens das amostras atípicas, das quais apenas a amostra 1 (classe 1, ponta controle) não foi detectada com comportamento atípico na modelagem utilizando SIMCA. Esta amostra apresenta cutículas em bom estado, com algumas poucas deformações e pedaços soltos (representados pelas partes brancas). No entanto, a parte

inferior da micrografia está muito escura, aparentemente porque a imagem foi obtida na região próxima à curvatura do fio, quando deveria ter sido obtida na região junto ao eixo da fibra. Esta região escura não faz parte da fibra, e portanto não apresenta cutículas. Desta maneira o algoritmo gera matrizes de descritores incompletas para estas imagens, com número de linhas inferior a 256 e que precisam ser ajustadas para o formato 256 x 16 para serem incluídas na análise, pois todas as matrizes devem ter a mesma dimensão para serem empilhadas. O ajuste da matriz para a dimensão correta foi possível utilizando o recurso NAN para dados faltantes[32]. No entanto, se o número de células matriciais a ser preenchido com o recurso para dados faltantes for muito grande, como é o caso da amostra 1, a modelagem poderá ser comprometida. As amostras 21 e 22 também estão muito degradadas, como discutido anteriormente[33], e apresentam ausência de cutícula em áreas muito extensas, o que também gera grande quantidade de dados faltantes (Figura 2.11).

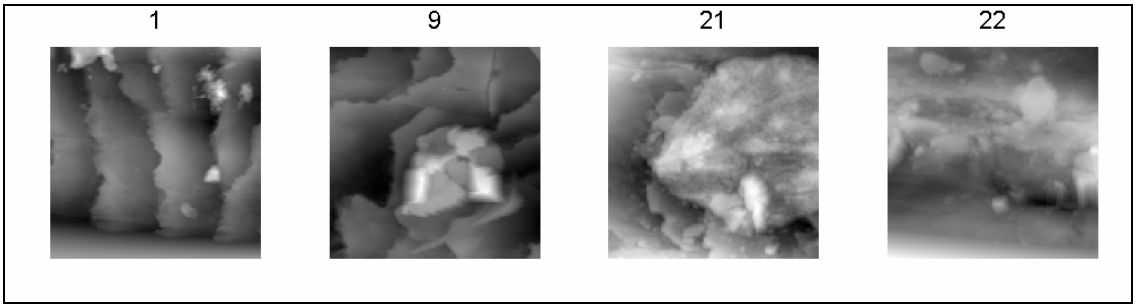


Figura 2. 11: Amostras atípicas detectadas pela análise NPLS com 6 fatores: amostras 1 (ponta virgem), 9 (ponta virgem), 21 (ponta descolorida), (c) 22 (ponta descolorida).

[32] NaN = *Not-a-Number*, recurso do software MatLab utilizado neste trabalho para realizar as análises quimiométricas. Este recurso permite a manipulação matemática de dados matriciais com células faltantes [Bro, 1997].
[33] Cf. seção 2.2.2.

Estas amostras também foram corrigidas para o formato 256 x 16 utilizando o recurso NaN. O comportamento atípico da amostra 9 foi discutido anteriormente na seção 2.2.2.

O uso de matrizes de descritores teve como objetivo contornar o problema da estimativa dos descritores nas regiões com ausência de cutícula. As matrizes de descritores contêm a distribuição dos descritores ao longo de toda a imagem, e esperava-se com isto melhorar de forma geral a correta classificação de todas as amostras, inclusive aquelas com problemas de dados faltantes. No entanto, as amostras atípicas detectadas pelo método SIMCA foram também detectadas pelo método das matrizes de descritores. Apesar de não ter sido possível contornar este problema, o método atual permitiu identificar também irregularidades em outra amostra (a amostra 1), o que não foi possível com o método SIMCA. Acredita-se, portanto, que por utilizar a distribuição dos descritores ao longo de toda a imagem, o método atual é mais representativo e identifica mais detalhes nas imagens, quando comparado ao método anterior.

Novas modelagens foram testadas após a exclusão das amostras mencionadas acima, na tentativa de obter valores maiores de % variância explicada em **Y** (Tabela 6).

O melhor modelo NPLS foi obtido após a exclusão das quatro amostras citadas, com 6 fatores descrevendo 99,92% da variância total do bloco **X** e 80,29% do bloco **Y**. A presença de amostras atípicas insere uma maior variabilidade no conjunto de dados, pois estas amostras apresentam um comportamento diferenciado em relação às outras amostras do modelo. Sendo assim, um número maior de fatores será

necessário para descrever as informações contidas na população de amostras em estudo.

Tabela 6: Variância dos blocos X e Y descritas pelos fatores dos modelos NPLS investigados. Dados escalados no Modo 3.

Fatores	Exclusão das Amostras 9, 21 e 22		Exclusão das Amostras 1, 9, 21 e 22	
	% Variância Bloco **X**	% Variância Bloco **Y**	% Variância Bloco **X**	% Variância Bloco **Y**
1	99,89	28,09	99,89	27,68
2	99,88	29,67	99,90	45,92
3	99,89	47,40	99,90	54,40
4	**99,90**	**55,65**	**99,91**	**62,87**
5	99,90	63,90	99,91	72,88
6	**99,91**	**73,60**	**99,92**	**80,29**

Modelagem: Análise Discriminante

As amostras foram divididas em dois conjuntos, um contendo 28 amostras e que foi utilizado para calibração com validação cruzada *leave-one-out*, e outro com 8 amostras (2 de cada classe), utilizado para validação externa. Inicialmente foi construído um modelo NPLS com as 28 amostras de calibração, mas a amostra 23 apresentou valor de *leverage* acima do valor limite para 99% de confiança (*hcrítico = 3k/I*, Figura 2.12(a)). O modelo foi refeito após a exclusão desta amostra e todas as amostras restantes apresentaram valores de *leverage* dentro deste limite (Figura 2.12(b)).

A Tabela 7 apresenta a variância descrita nos dois casos, onde é possível observar que a variância descrita em **X** aumentou de 99,92% para 99,94% no fator 6, após a exclusão da amostra 23, e em **Y** aumentou de 85,04% para 87,50%.

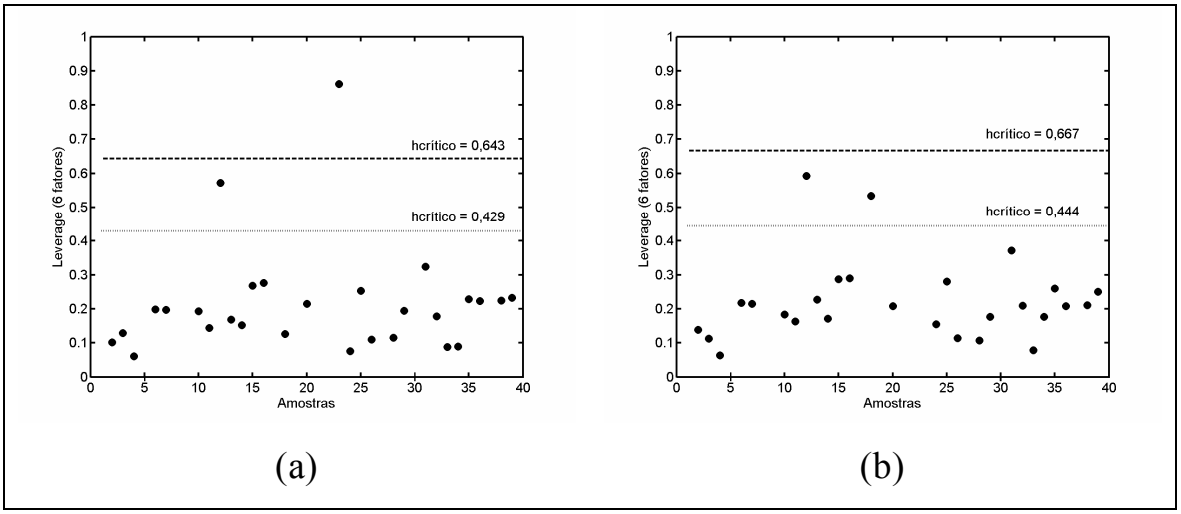


Figura 2. 12: Gráfico dos valores de *leverage* vs. amostra para o modelo com (a) 28 amostras e (b) 27 amostras (após a exclusão da amostra 23).

Tabela 7: % de variância descrita em cada fator para os blocos **X** e **Y** dos modelos NPLS investigados. Dados escalados no Modo 3 e modelos com validação cruzada *leave-one-out*.

	Conjunto com 28 amostras		Conjunto com 27 amostras	
	% Variância Bloco **X**	% Variância Bloco **Y**	% Variância Bloco **X**	% Variância Bloco **Y**
1	99,89	28,32	99,91	29,23
2	99,90	44,59	99,92	43,89
3	99,90	53,25	99,92	57,46
4	**99,91**	**62,20**	**99,93**	**66,48**
5	99,92	75,72	99,93	81,69
6	**99,92**	**85,04**	**99,94**	**87,50**
7	99,93	88,37	99,94	90,70
8	99,93	92,51	99,94	94,41
9	99,93	96,13	99,94	96,40
10	99,93	97,68	99,95	97,75

Para investigar o desempenho dos modelos com 4 e 6 fatores, os valores de **Y** preditos durante a calibração e a validação cruzada *leave-one-out* foram avaliados (Tabela 8).

Tabela 8: Classificação das amostras de calibração durante a etapa de validação cruzada.

	Classe Definida*	Classe Predita – Modelos com 27 amostras			
		Modelo com 4 Fatores		Modelo com 6 Fatores	
		Ycal**	Ycrossval**	Ycal**	Ycrossval**
2	1	**3**	**3**	1	**3**
3	1	**3**	**3**	1	**3**
4	1	1	**4**	1	**3**
6	1	1	**4**	1	1
7	1	1	**4**	1	1
10	1	1	1	1	1
11	1	1	**3**	1	**3**
12	1	1	1	1	1
13	1	1	**3**	1	**3**
14	1	1	1	1	**3**
15	2	2	2	2	2
16	2	2	2	2	2
18	2	2	2	2	2
20	2	2	2	2	2
24	3	**1**	**1**	3	**1**
25	3	3	**1**	3	**1**
26	3	3	3	3	3
28	3	3	3	3	**1**
29	3	3	3	3	3
31	3	3	3	3	3
32	3	3	3	3	3
33	3	3	**1**	**1**	**1**
34	3	3	3	3	3
35	4	4	4	4	4
36	4	4	**1**	4	4
38	4	4	**1**	4	4
39	4	4	4	4	4
Acertos		24	15	26	17
Erros na classe 1		2	7	0	6
Erros na classe 2		0	0	0	0
Erros na classe 3		1	3	1	4
Erros na classe 4		0	2	0	0
% Total de Erros		11,1%	44,4%	3,7%	37,0%

Legenda: * classes: (1) ponta controle, (2) ponta descolorida, (3) raiz descolorida, (4) raiz controle; ** Ycal = Y predito de calibração; Yval = Y predito de validação cruzada.

A classificação é fornecida pelo valor do vetor **y** característico da classe nos modelos com 4 e 6 fatores, sendo que os critérios adotados para identificar a qual classe a amostra pertence foram:

- os valores de **y** preditos pelo modelo foram aproximados para 0 (não pertence à classe) ou 1 (pertence à classe).
- nos casos em que os valores de **y** para as quatro colunas da matriz **Y** ficaram abaixo de 0,5, identificou-se a classe da amostra como sendo aquela que apresentou o maior valor de **y**.

O modelo construído com 6 fatores revelou melhor capacidade preditiva, com apenas 1 amostra classificada incorretamente durante a calibração, o que representa apenas 3,7% das amostras. Os erros durante a validação cruzada também foram menores para o modelo com 6 fatores. De forma geral, observa-se que:

- Em todos os casos, as amostras da classe 2 (ponta descolorida) foram classificadas corretamente. Este é o grupo de amostras mais heterogêneo.
- As amostras da classe 4 (raiz controle) também foram bem classificadas, mas o modelo com 4 fatores apresentou 2 amostras classificadas incorretamente durate a validação cruzada. Este é o grupo de amostras mais homogêneo.
- As amostras da classe 1 (ponta controle) e 3 (raiz descolorida) apresentaram o maior número de predições incorretas. Estas duas classes apresentam características similares em termos de degradação da fibra e são difíceis de serem distinguidas. Somente o modelo com 6 fatores foi capaz de discriminar estas duas classes, que apresentaram valores de **Y** de calibração corretos, com apenas uma exceção (a amostra 33). Os valores de **Y** preditos

durante a validação cruzada demonstram a dificuldade de distinguir estas duas classes de amostras, pois amostras da classe 1 classificadas incorretamente nesta etapa foram identificadas como pertencentes à classe 3 e vice-versa.

Com base nestas observações, o modelo construído com 6 fatores foi considerado mais adequado, por ter apresentado menor número de amostras classificadas incorretamente durante as etapas de calibração e validação cruzada *leave-one-out*.

Validação Externa - Predição

O modelo construído com 27 amostras e 6 fatores foi utilizado para classificar 8 amostras externas. Os resultados podem ser observados na Tabela 9, em que é possível observar que apenas uma amostra (de número 19), não foi classificada corretamente.

Tabela 9: Classificação das amostras de validação externa

Classe 1 Ponta controle		Classe 2 Ponta descolorida		Classe 3 Raiz descolorida		Classe 4 Raiz controle	
Amostra	Classe Predita	Amostra	Classe Predita	Amostra	Classe Predita	Amostra	Classe Predita
5	1	17	2	27	3	37	4
8	1	19	**1**	30	3	40	4

Esta amostra pertence à classe ponta descolorida e está bastante deteriorada, praticamente sem cutículas e com o córtex bastante exposto (Figura 2.13).

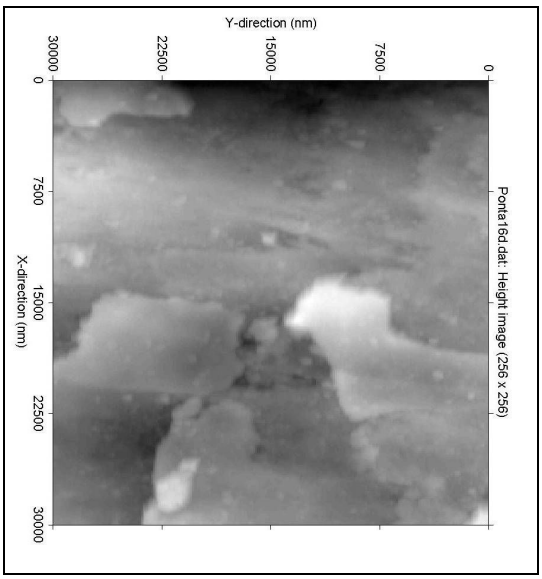


Figura 2. 13: Amostra número 19, que foi mal classificada pelo modelo NPLS.

Neste estudo, seis variáveis latentes foram necessárias para a modelagem. No entanto, a maior parte desta variância está concentrada nos três primeiros fatores. Isto significa que a maior parte da informação necessária para discriminar as quatro classes de amostras está nestes fatores, como pode ser observado no gráfico dos escores dos fatores 1, 2 e 3 para o Modo *I* (modo das amostras) do bloco **X** (Figuras 2.14 e 2.15).

O fator 1 mostra a separação de algumas poucas amostras não relacionadas especificamente a nenhuma das quatro classes. O fator 2 descreve a origem da fibra e o fator 3 o efeito do tratamento cosmético.

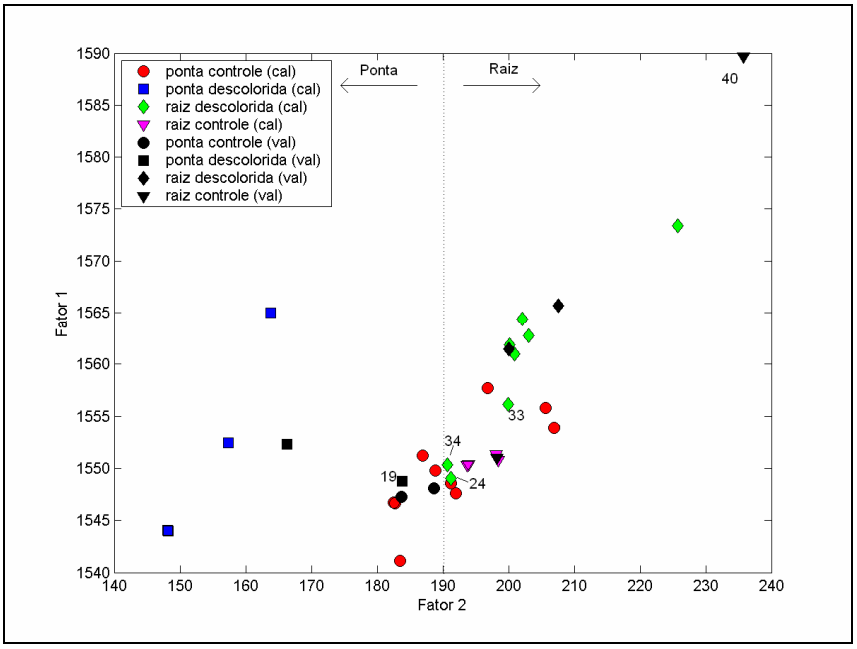


Figura 2. 14: Escores dos fatores 1 e 2 para a modelagem NPLS com dados escalados no Modo *K*.

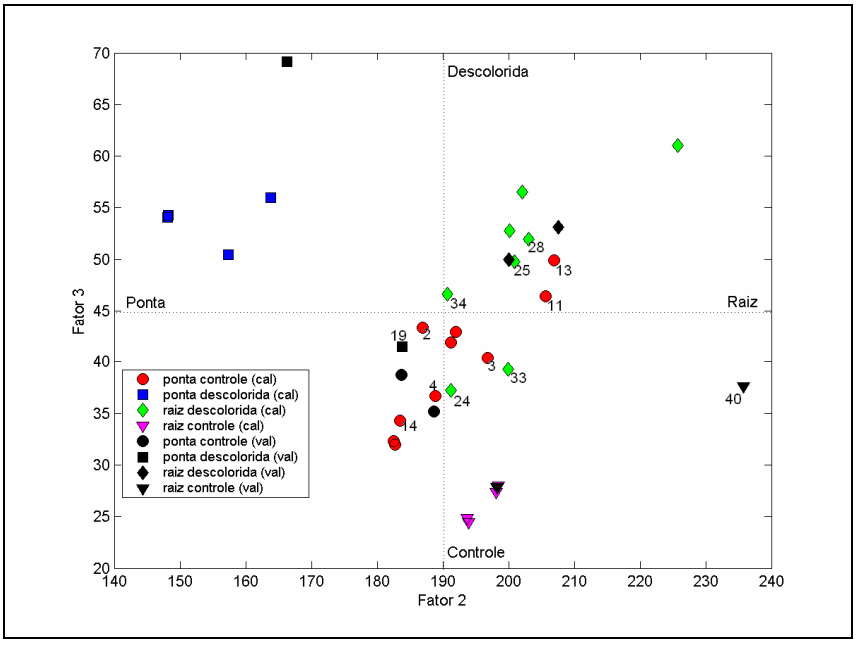


Figura 2. 15: Escores dos fatores 2 e 3 para a modelagem NPLS com dados escalados no Modo *K*.

Na Figura 2.15 nota-se claramente a presença de três grupos distintos de amostras: o grupo homogêneo das amostras da classe 4 (raiz controle); o

grupo disperso das amostras da classe 2 (ponta descolorida) e o grupo das amostras das classes 1 (ponta controle) e 3 (raiz descolorida), cujas amostras estão próximas, mas apresentam uma tendência de separação. Neste último grupo, em destaque na Figura 2.15, estão indicadas as amostras do conjunto de calibração que apresentaram erros de classificação no **Y** de calibração (amostra 33), ou no **Y** de validação cruzada (amostras 2, 3, 4, 11, 13, 14, 24, 25, 28).

Na Figura 2.15 é possível ainda observar que a amostra externa 19 (ponta descolorida), classificada incorretamente pelo modelo, está próxima à região central do gráfico. A amostra externa 40 (raiz controle), apesar de ter sido corretamente classificada e de estar localizada na região das amostras da sua classe, apresenta valores de escores um pouco mais altos nos fatores 2 e 3. Esta amostra apresenta resolução diferente das outras amostras e já havia sido classificada incorretamente pelo modelo SIMCA[34].

As Figuras 2.16(a) e 2.16(b) mostram os escores para os fatores 3, 4, 5 e 6. No gráfico de escores dos fatores 3 e 4 é possível notar uma tendência para a separação das amostras das classes controle e descolorida. Os escores dos fatores 5 e 6 não apresentam nenhum padrão significativo, porém, como demonstrado anteriormente (Tabela 8) estes fatores foram necessários para a adequada modelagem de todas as 4 classes.

[34] Cf. seção 2.2.

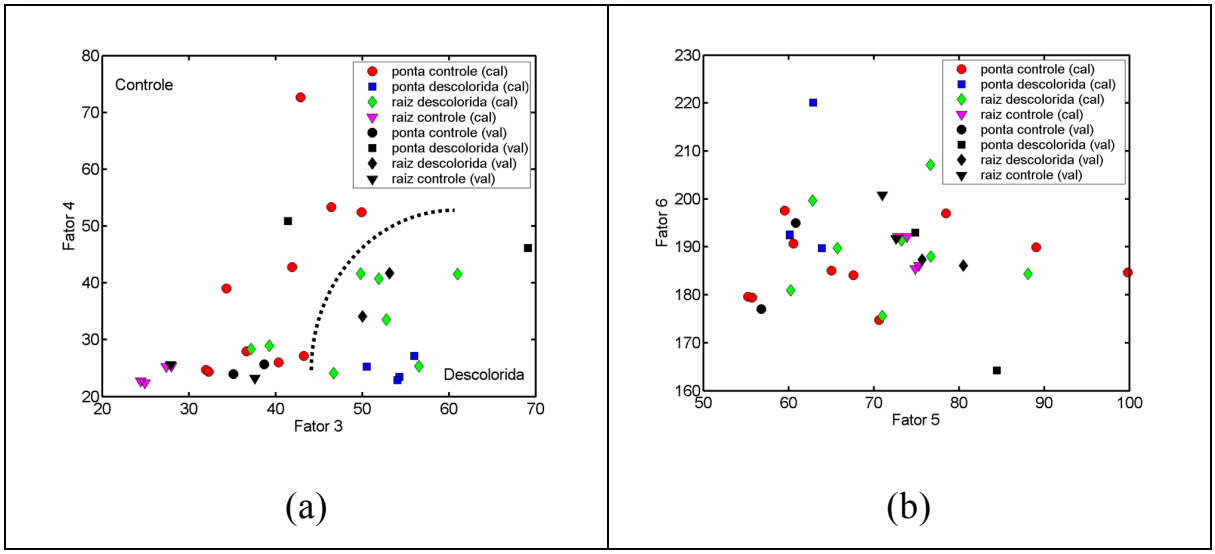


Figura 2. 16: Escores dos fatores (a) 3 e 4 e (b) 5 e 6 para a modelagem NPLS com dados escalados no Modo *K*.

As Figuras 2.17(a)-2.17(f) contêm os gráficos dos pesos do Modo 3 (modo dos descritores), que descrevem a contribuição de cada descritor em cada fator.

A rugosidade está associada ao estado de degradação da fibra e costuma ser descrita na literatura como um dos fatores mais importantes na sua caracterização [You & Yu, 1997; McMullen *et al.*, 2000; Swift & Smith, 2000; McMullen & Kelty, 2001; Monteiro, 2003]. Na modelagem NPLS, a rugosidade é descrita pelo fator 1, o que comprova a importância desta variável na descrição do conjunto de fibras. No entanto, este descritor separou apenas algumas amostras do conjunto total, o que indica que apesar da sua importância, outros descritores tiveram peso maior para a discriminação das quatro classes de amostras. Esta informação somente foi descrita nos fatores 2 e 3.

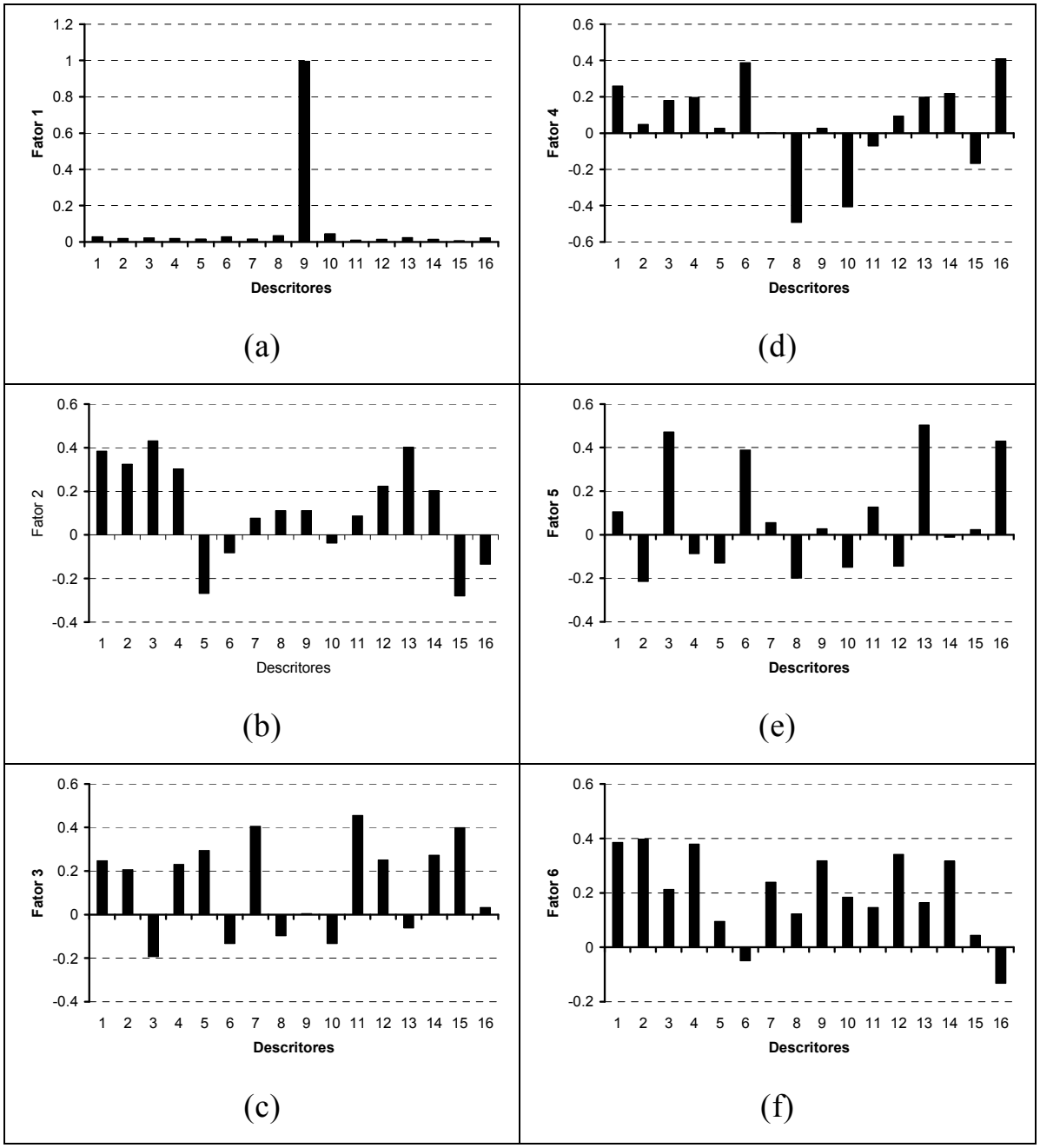


Figura 2. 17: Pesos (a) 1, (b) 2, (c) 3, (d) 4, (e) 5 e (f) 6 do Modo *K* (descritores).

Nos fatores 2 e 3 (Figuras 2.14 e 2.15), as amostras de cada classe ocupam regiões distintas: a origem da fibra é identificada no fator 2, enquanto o efeito do tratamento cosmético é descrito pelo fator 3. As amostras das classes ponta descolorida e raiz controle representam os

níveis mais extremos de degradação e alteração da fibra, e por este motivo ocupam regiões bem características no gráfico dos escores. As amostras das outras duas classes, ponta controle e raiz descolorida, são mais similares e ocupam uma região intermediária.

De uma geral o que se investiga em ambos os casos é o grau de degradação da fibra capilar ocasionado tanto pelos processos naturais ocorridos ao longo da sua extensão (raiz/ponta), quanto pela oxidação por peróxido (descolorida/controle). Pela observação simultânea dos escores dos fatores 2 e 3 na Figura 2.15, é possível perceber claramente uma tendência ao longo da diagonal do gráfico, que reflete o nível de degradação da fibra, com contribuições tanto da origem (fator 2) quanto do tratamento cosmético (fator 3). Nesta diagonal, as amostras apresentam três principais agrupamentos:

- *na região dos valores altos do fator 2 e valores baixos do fator 3 encontram-se as amostras de raiz controle*: As maiores contribuições no fator 2 foram dos descritores 1 (*step height*), 2 (*tilt*), 3 e 13 (*backtilt* e seu respectivo desvio padrão), 4 (*layer spacing*), que apresentaram valores de pesos acima de 0,3 (Figura 2.17(b)). Isto significa que fibras mais preservadas apresentarão valores maiores para os descritores *step height*, *tilt*, *backtilt* e *layer spacing*, além de valores mais heterogêneos de *backtilt*, indicado pelos altos valores de desvio padrão de *backtilt* (descritor 13).

- *na região dos valores baixos do fator 2 e valores altos do fator 3 localizam-se as amostras de ponta descolorida*: Os descritores 5 e 15 (*face distance* e seu respectivo desvio padrão) apresentaram valores de pesos

próximos de -0,3 no fator 2 (Figura 2.17(b)), o que significa que fibras mais deterioradas apresentarão valores maiores e mais homogêneos para *face distance* (descritores 5 e 15). No fator 3, as maiores contribuições são dos descritores 5 e 15 (*face distance* e seu respectivo desvio padrão), 7 (*fit error*) e 11 (desvio padrão de *step height*) (Figura 2.17(c)), que apresentaram valores de pesos > 0,295.

- *na região de valores intermediários dos fatores 2 e 3 localizam-se as amostras de raiz descolorida e ponta controle*. Apesar destas amostras ocuparem regiões próximas, as amostras de raiz descolorida tendem a ocupar a região dos valores mais altos dos fatores 2 e 3.

A descoloração por peróxido promove a degradação da CMC, que mantém as fibras unidas, e da estrutura protéica da superfície da fibra, favorecendo a quebra e consequente perda de partes das cutículas [Robbins 1994; Monteiro, 2003]. Os valores de *step height* refletem a espessura da cutícula em relação ao eixo central da fibra, e quanto mais degastada é a superfície cuticular, menor sua espessura e menor é o valor deste descritor [Monteiro, 2003].
Consequentemente, os valores de *tilt*, *backtilt* e *layer spacing* serão menores para fibras descoloridas, pois a superfície da cutícula será mais plana, e os valores de *face distance* serão maiores. Fibras mais antigas também sofrem perda de cutículas, que resulta em valores menores para estes descritores. A Figura 2.18 apresenta os descritores e como o seu cálculo é afetado pela quebra das bordas das cutículas.

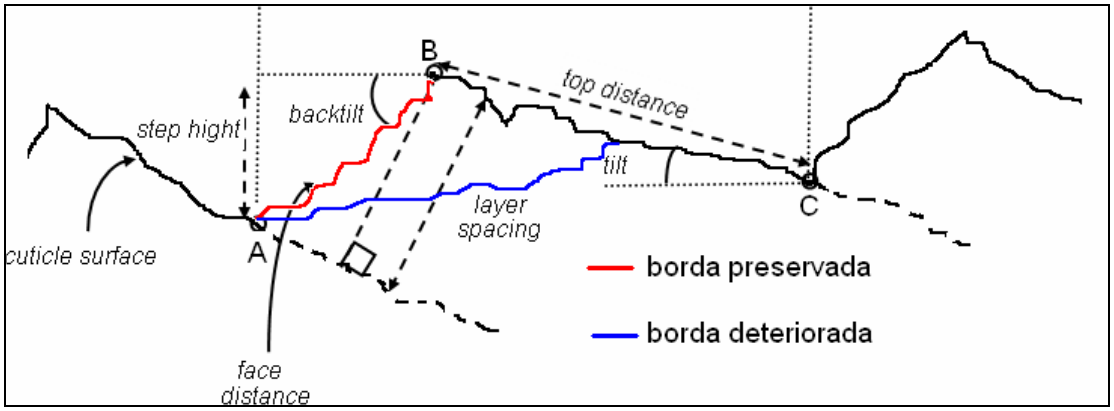


Figura 2. 18: Ilustração da perda de partes da cutícula e os efeitos na estimativa dos descritores.

O aumento dos valores de *face distance* (descritor 5), e seu desvio padrão (descritor 15) está relacionado ao aumento do estado de degradação da fibra. O descritor *fit error* (7) descreve a qualidade do ajuste do perfil à superfície da fibra (diferença entre o *perfil longitudinal* e o *perfil ajustado*), como descrito anteriormente[35], e é sensível à presença de irregularidades na superfície da fibra. Fibras mais degradadas, principalmente as fibras da ponta, apresentam maiores irregularidades e, consequentemente, irão apresentar valores maiores para este descritor [Robbins, 1994; Gurden *et al*., 2004; Swif & Smith, 2000].

2.4. Discussão Geral dos Resultados

Os dois métodos quimiométricos empregados neste trabalho tiveram como objetivo discriminar as quatro classes de amostras de fibras capilares e construir modelos para a classificação de novas amostras. As

[35] Cf. seção 2.1.

variáveis utilizadas nos modelos foram os descritores capilares estimados a partir de imagens AFM das fibras e que fornecem suas principais características morfológicas.

Os dois métodos foram capazes de discriminar as quatro classes de amostras com sucesso e de classificar novas amostras adequadamente, além de permitir identificar amostras com comportamento atípico. Neste último item o modelo de ordem superior NPLS mostrou-se mais eficaz, pois considera a distribuição dos descritores ao longo de toda a imagem da fibra, e como conseqüência é capaz de identificar mais detalhes. O modelo SIMCA por outro lado é construído a partir do valor médio e do desvio padrão dos descritores para toda a imagem, e a informação referente à distribuição destas informações ao longo da imagem fica comprometida.

De forma geral os dois métodos identificam o estado de degradação da fibra, ocasionado tanto pelos processos de deterioração natural, e que pode ser notado na distinção entre amostras da ponta e da região da raiz, quanto pela ação do peróxido. É possível observar três estágios de degradação:

- a classe ponta descolorida (□): constituída por fibras que sofreram maior degradação, com cutículas deterioradas e regiões com ausência de cutículas. Devido ao aspecto extremamente heterogêneo, as amostras desta classe apresentaram-se dispersas no gráfico dos pesos (Figuras 2.14 e 2.15).

- a classe raiz virgem (∇): formada por fibras novas e preservadas, com cutículas intactas. Seu aspecto é extremamente uniforme e as amostras formam um agrupamento bastante compacto e homogêneo nos gráficos dos pesos.

Estas duas classes de amostras representam os extremos do nível de deterioração da fibra capilar, e tendem a apresentar um aspecto visual totalmente distinto, que foi facilmente identificado pelos tratamentos quimiométricos aplicados nas imagens.

- as classes ponta virgem (O) e raiz descolorida (◊): estas duas classes são extremamente parecidas e, portanto, de difícil distinção entre si. Pode-se considerar que elas apresentam aspecto intermediário entre as amostras da classe raiz virgem (∇) e ponta descolorida (□), e por causa desta similaridade sua discriminação no gráfico dos pesos é mais difícil (Figuras 2.14 e 2.15). Apesar desta similaridade, os dois métodos quimiométricos foram capazes de classificar corretamente amostras destas duas classes.

CAPÍTULO 3

Trabalhando Diretamente com as Imagens – Métodos de Ordem Superior: NPLS

Efeito de Tratamento Cosmético: Discriminação Quantitativa de Imagens de Fibras Capilares Submetidas ao Tratamento de Descoloração

3.1 Arranjo dos Dados

O conjunto de dados utilizado é formado por um arranjo tridimensional $\underline{\mathbf{X}}$ (40, 256, 256) contendo 40 matrizes das imagens de 256 pixels x 256 pixels empilhadas, e um arranjo bidimensional do tipo $\mathbf{Y}$ (I, J) com I = 40 amostras e J = 4 colunas, sendo cada uma referente a uma classe de amostra (y_1 = ponta controle, y_2 = ponta descolorida, y_3 = raiz descolorida e y_4 = raiz controle). A cada coluna da matriz $\mathbf{Y}$ foram atribuídos valores categóricos discretos de 0 ou 1, sendo 0 atribuído às amostras que não pertencem à categoria e 1 atribuído às amostras pertencentes à categoria. O arranjo $\underline{\mathbf{X}}$ de imagens foi submetido a transformada de Fourier bidimensional [Bharati *et al.*, 2004; Geladi, 1992; Huang *et al.*, 2003] e os dados foram modelados utilizando o método NPLS [Bro, 1998; Bro, 1996].

3.2 Transformações

O uso do método NPLS pressupõe trilinearidade dos dados. As imagens utilizadas neste estudo são do tipo nível de cinza e neste caso a posição de cada cela da matriz fornece a coordenada geográfica dos pixels e seu valor fornece a intensidade da cor, que no caso, vai do branco até o preto, passando pelos diferentes graus de cinza. Sendo assim, toda imagem *nível de cinza* é por natureza univariada e quando empilhadas caracterizam um arranjo do tipo objeto x objeto x objeto (OOO) que não atinge o critério de trilinearidade necessário à aplicação direta dos métodos de ordem superior. Neste caso, faz-se necessária a aplicação de uma transformação, de forma a obter-se um arranjo trilinear do tipo objeto x variável x variável (OVV). Diferentes técnicas como transformada de Fourier e transformada wavelet têm sido utilizadas para resolver este problema [Bharati *et al.*, 2004; Simoncelli & Olshausen, 2001; Huang *et al.*, 2003 Geladi, 1992]. Neste trabalho utilizamos o espectro de potência da transformada de Fourier bidimensional, 2D-FFT, seguida de uma transformação logarítmica. A transformada de Fourier converte uma função no domínio espacial para o domínio da frequência, e assim transforma o arranjo OOO em OVV. Ao aplicar a transformada de Fourier na matriz da imagem, o resultado é uma matriz com números complexos. É necessário então utilizar o espectro de potência, que nada mais é do que o quadrado do módulo da transformada de Fourier. A transformação logarítmica tem a função de reduzir a dispersão dos valores e aproximá-la da distribuição normal. A Figura 3.1 apresenta uma imagem de cada classe deste estudo antes e após a aplicação de 2D-FFT e LOG. A Figura 3.2(a) apresenta uma

imagem submetida somente à transformação 2D-FFT e a Figura 3.2(b) apresenta a mesma imagem submetida à 2D-FFT (espectro de potência) seguida de LOG, em que é possível demonstrar como é possível visualizar mais detalhes na imagem 2D-FFT após a aplicação do logarítmo. As Figuras 3.2(c) e 3.2(d) apresentam respectivamente as superfícies destas mesmas imagens.

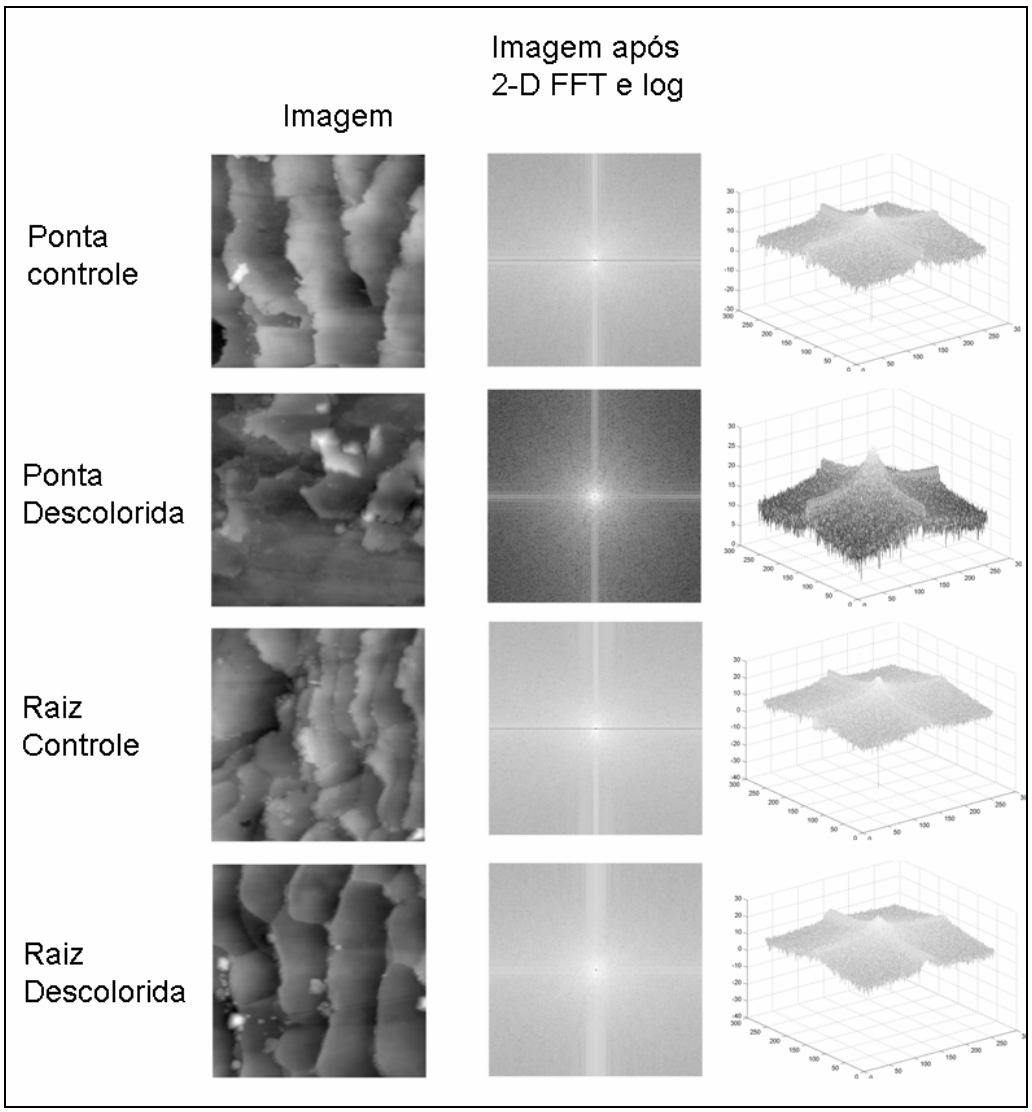


Figura 3. 1: Representação da imagem univariada (OO) após a transformação 2D-FFT (espectros de potência) e LOG, em que o domínio da frequência 2-D forma os dois novos modos de variáveis (VV).

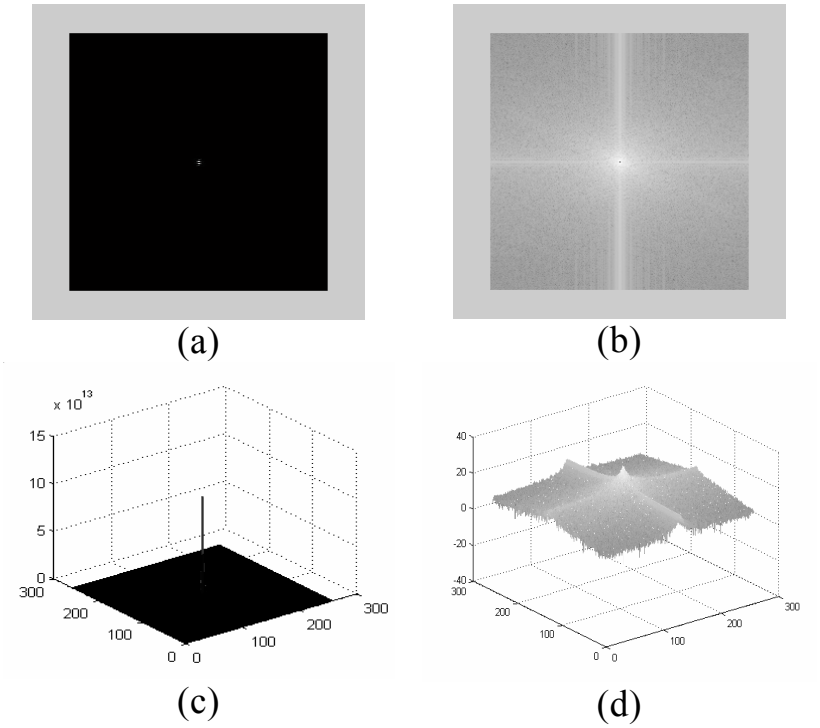


Figura 3. 2: Representação de uma imagem após a transformação 2D-FFT (a) e (c), e após a transformação 2D-FFT seguida de LOG (b) e (d).

3.3 Estudo do Tipo de Preprocessamento

Neste caso, as variáveis dos modos *J* e *K* são da mesma ordem de grandeza, e em analogia com o que acontece com dados bidimensionais, um preprocessamento centrado na média seria útil na remoção de *offset*. No entanto, preprocessamento em arranjos *multi-way* não é algo trivial, pois muitas vezes a estrutura trilinear dos dados poderá ser comprometida [Bro & Smilde, 2003; Gurden *et al.*, 2001].

O efeito do tipo de preprocessamento dos dados foi testado, comparando os resultados da variação descrita para os blocos **X** e **Y** obtidos para a modelagem com os dados não processados e com os

dados centrados na média nos Modos *J* e *K* separadamente. Os melhores resultados foram obtidos para os dados não processados (Tabela 10).

Tabela 10: Variância dos blocos **X** e **Y** descritas pelas variáveis latentes do modelo para os dados não processados e dados centrados na média nos modos *J* e *K*.

	Dados não processados		Dados Centrados na Média no Modo *J*		Dados Centrados na Média no Modo *K*	
	% Variância no Bloco **X**	% Variância no Bloco **Y**	% Variância no Bloco **X**	% Variância no Bloco **Y**	% Variância no Bloco **X**	% Variância no Bloco **Y**
1	99,06	27,19	70,45	27,27	70,45	27,27
2	99,10	40,32	71,88	37,18	71,88	37,18
3	99,16	56,24	72,09	51,82	72,09	51,82
4	99,18	65,72	72,96	55,47	72,96	55,47
5	99,21	72,59	73,20	63,58	73,20	63,58
6	99,22	79,34	73,46	75,51	73,46	75,51
7	99,23	84,26	73,58	82,22	73,58	82,22
8	99,24	88,49	73,70	87,04	73,70	87,04
9	99,25	91,11	73,82	89,88	73,82	89,88
10	99,25	93,06	73,94	93,16	73,94	93,16

3.4 Análise dos Dados

Os dados não processados foram modelados utilizando NPLS com validação cruzada *leave-one-out*, e até 10 fatores. A Tabela 11 apresenta a porcentagem de variação descrita em cada fator, na qual é possível observar que a partir dos modelos com 5 e 6 fatores a porcentagem de variância descrita para o bloco **Y** praticamente não apresenta ganhos significativos.

As Figuras 3.3(a) e 3.3(b) apresentam os valores de *leverage* obtidos nestas condições para as amostras do conjunto, em que é possível observar que para o modelo com 6 fatores, todas as amostras encontram-se dentro do limite de *hcrítico = 2k/I*. O modelo com 5

fatores apresentou uma amotra (36) acima do limite de hcrítico = *3k/I*, mas dentro de *2k/I*.

Tabela 11: % de variação descrita em cada fator para os blocos **X** e **Y** dos modelos NPLS investigados. Dados não escalados e modelos com validação cruzada *leave-one-out*.

Conjunto com 40 amostras		
Fatores	**% Variação Explicada no Bloco X**	**% Variação Explicada no Bloco Y**
1	99,06	27,19
2	99,10	40,32
3	99,16	56,24
4	99,18	65,72
5	**99,21**	**72,59**
6	**99,22**	**79,34**
7	99,23	84,26
8	99,24	88,49
9	99,25	91,11
10	99,25	93,06

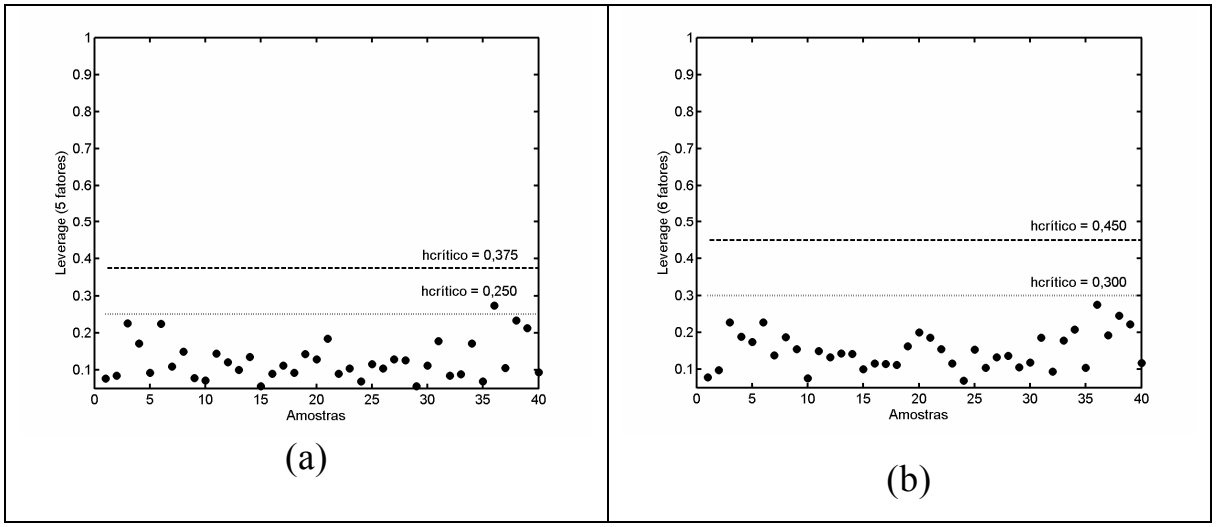


Figura 3. 3: Gráfico dos valores de *leverage* vs. amostra para modelo NPLS com (a) 5 fatores e (b) 6 fatores. Dados não processados.

Modelagem: Análise Discriminante

As amostras foram divididas em dois conjuntos, um contendo 32 amostras, utilizado para calibração com validação cruzada *leave-one-out*, e outro com 8 amostras (2 de cada classe), utilizado para validação externa. A Tabela 12 contém a % de variação descrita em cada fator para os blocos **X** e **Y** do modelo NPLS, e a Figura 3.4 contém o gráfico de *leverage* para os modelos com 5 e 6 fatores, em que é possível observar que em ambos os modelos todas as amostras encontram-se dentro do valor limite para 95% de confiança (*hcrítico = 2k/I*).

Tabela 12: % de variação descrita em cada fator para os blocos **X** e **Y** do modelo NPLS investigado. Dados não escalados e modelos com validação cruzada *leave-one-out*.

	Conjunto com 32 amostras	
Fatores	**% Variação Explicada no Bloco X**	**% Variação Explicada no Bloco Y**
1	99,07	28,32
2	99,11	42,53
3	99,16	55,14
4	99,19	60,94
5	**99,20**	**75,92**
6	**99,21**	**81,61**
7	99,22	87,38
8	99,23	90,84
9	99,24	93,58
10	99,25	95,25

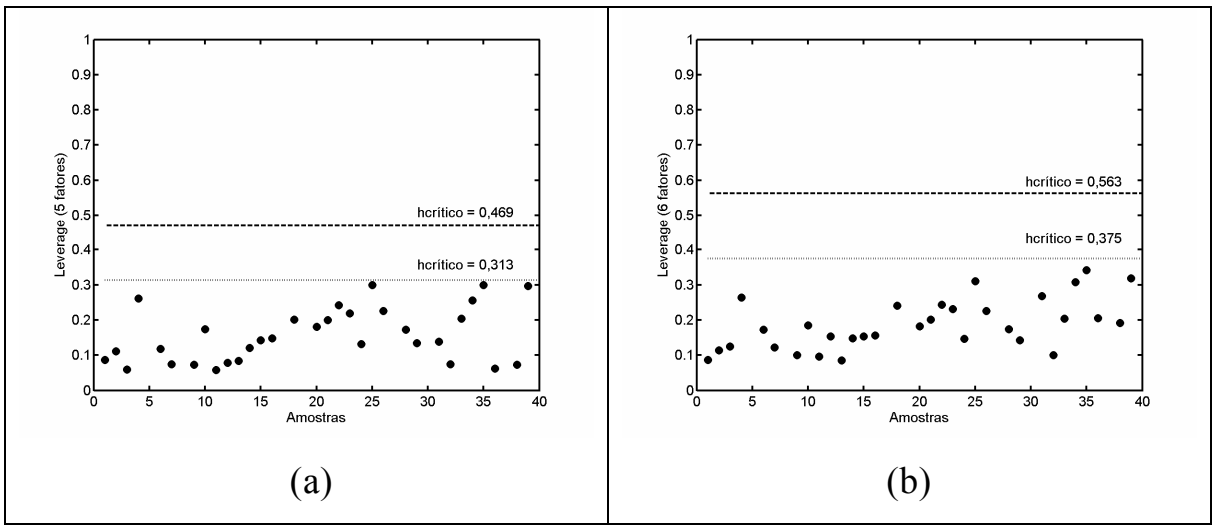


Figura 3. 4: Gráfico dos valores de *leverage* vs. amostra para o conjunto de calibração: (a) modelo com 5 fatores; (b) modelo com 6 fatores.

A Tabela 13 contém a classificação das amostras do conjunto de calibração e os resultados da validação cruzada para os modelos com 5 e 6 fatores.

O modelo com 6 fatores apresentou os menores erros para a predição do **Y** de calibração, apesar de ter apresentado erros de validação maiores do que o modelo com 5 fatores. No entanto, o modelo com 5 fatores não foi capaz de classificar corretamente as amostras da classe 4 (raiz controle).

Como descrito anteriormente, a classificação é fornecida pelo valor do vetor **y** característico da classe no fator escolhido, em que os valores de **y** preditos pelo modelo foram aproximados para 0 (não pertence à classe) ou 1 (pertence à classe). Nos casos em que o valor de **y** em todas as colunas foi menor que 0,5, considerou-se o maior valor.

Tabela 13: Classificação das amostras do conjunto de calibração durante a etapa de validação cruzada *leave-one-out*.

Amostras	Classe Definida*	Classe Predita (5 fatores)		Classe Predita (6 fatores)	
		Ycal	**Yval**	**Ycal**	**Yval**
1	1	1	1	1	1
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	**3**	1	**3**
6	1	1	1	1	1
7	1	1	1	1	1
9	1	1	1	1	1
10	1	1	1	1	1
11	1	1	1	1	1
12	1	1	**2**	1	**2**
13	1	1	1	1	1
14	1	1	1	1	1
15	2	**3**	**3**	**3**	**3**
16	2	**3**	**3**	**3**	**3**
18	2	2	2	2	2
20	2	2	2	2	2
21	2	2	2	2	2
22	2	2	2	2	2
23	2	2	2	2	2
24	3	3	3	3	3
25	3	3	3	3	3
26	3	3	3	3	3
28	3	3	3	3	3
29	3	3	3	3	3
31	3	**4**	**1**	**4**	**4**
32	3	3	**1**	3	**1**
33	3	3	3	3	3
34	3	3	3	3	**1**
35	4	4	4	4	4
36	4	**1**	**1**	4	**1**
38	4	**1**	**1**	4	**1**
39	4	4	4	4	4
	Acertos	27	24	29	23
	Erros na classe 1	0	2	0	2
	Erros na classe 2	2	2	2	2
	Erros na classe 3	1	2	1	3
	Erros na classe 4	2	2	0	2
	% Total de Erros	15,6%	25%	9,4%	28,1%

* classes: (1) ponta controle, (2) ponta descolorida, (3) raiz descolorida, (4) raiz controle.

Os altos erros de validação cruzada indicam que o conjunto de dados é heterogêneo e algumas classes de amostras estão pouco representadas no conjunto. É o caso, por exemplo, da classe raiz controle, que contém apenas 4 amostras no conjunto de calibração. Apesar deste erro durante a validação cruzada *leave-one-out*, considerou-se mais adequado o modelo com 6 fatores.

A Figura 3.5 apresenta o gráfico dos escores para os fatores 1, 3 e 5 do Modo *I* para as amostras de calibração, no qual é possível observar que as quatro classes de amostras ocupam regiões distintas. Há ocorrência de amostras em regiões intermediárias ou regiões de outra classe, algumas das quais já tiveram o comportamento atípico detectado na análise anterior[36].

[36] Cf. Capítulo 2.

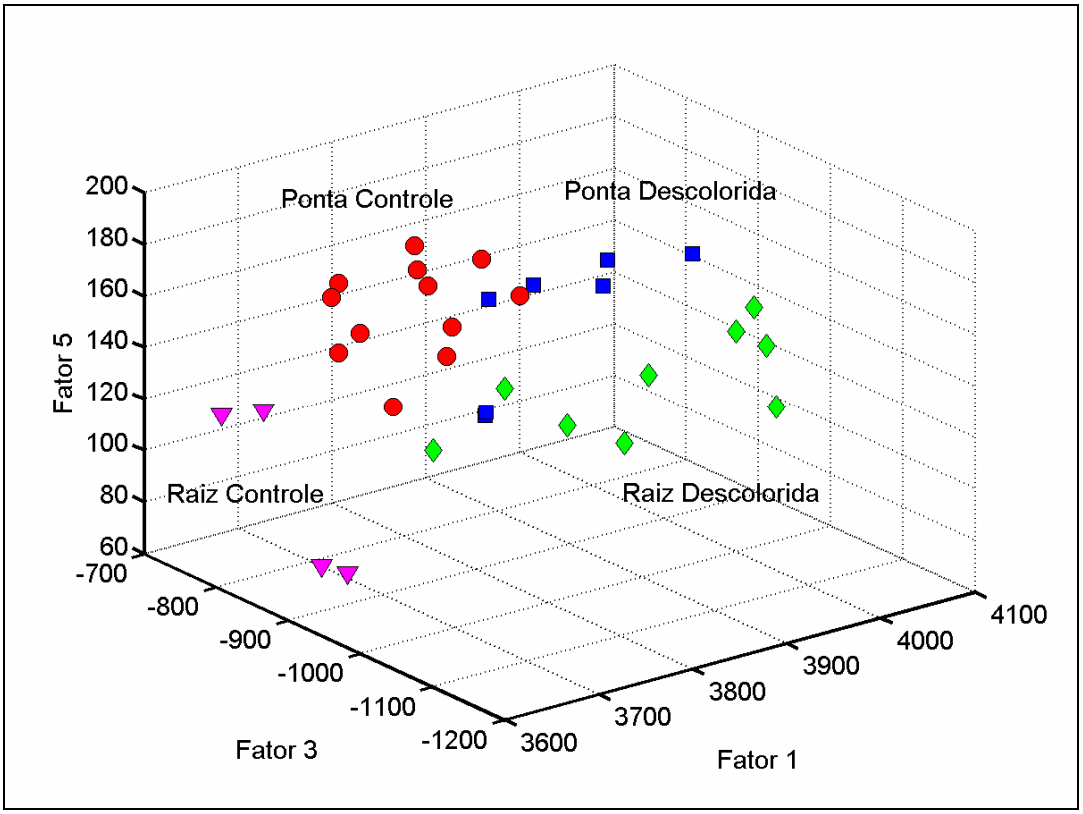


Figura 3. 5: Gráfico dos escores dos fatores 1, 3 e 5 para o conjunto de calibração. Legenda: (O) ponta virgem, (□) ponta descolorida, (∇) raiz virgem, (◊) raiz descolorida.

Validação Externa - Predição

O modelo NPLS com 6 fatores foi utilizado para classificar as 8 amostras externas, das quais apenas 1 foi classificada incorretamente (Tabela 14). Esta amostra pertence à classe ponta descolorida, que apresenta maior heterogeneidade quando comparada às outras amostras do conjunto. Esta classe também foi a que apresentou o maior número de amostras classificadas incorretamente (2) para o **Y** de calibração (Tabela 13).

Tabela 14: Classificação das amostras de validação externa

Classe 1 Ponta controle		Classe 2 Ponta descolorida		Classe 3 Raiz descolorida		Classe 4 Raiz controle	
Amostra	Classe Predita	Amostra	Classe Predita	Amostra	Classe Predita	Amostra	Classe Predita
5	1	17	**1**	27	3	37	4
8	1	19	2	30	3	40	4

A Figura 3.6 contém o gráfico dos escores para os fatores 1, 3 e 5 contendo as amostras de calibração e de validação externa.

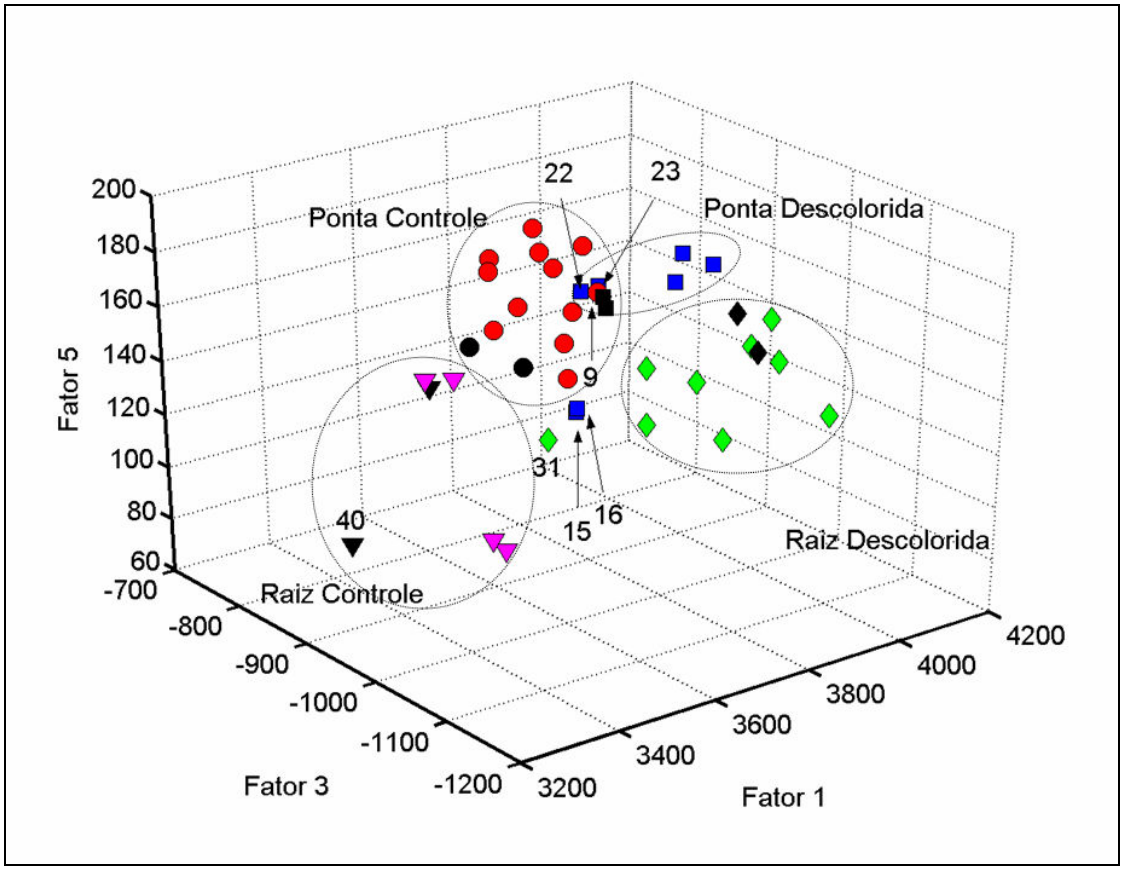


Figura 3. 6: Gráfico dos escores dos fatores 1, 3 e 5 para as amostras de calibração e de validação externa. Legenda: (O) ponta virgem, (□) ponta descolorida, (∇) raiz virgem, (◊) raiz descolorida.

A Figura 3.7 apresenta cada fator separadamente e, desta maneira, permite visualizar a influência de cada um deles na discriminação das classes de amostras.

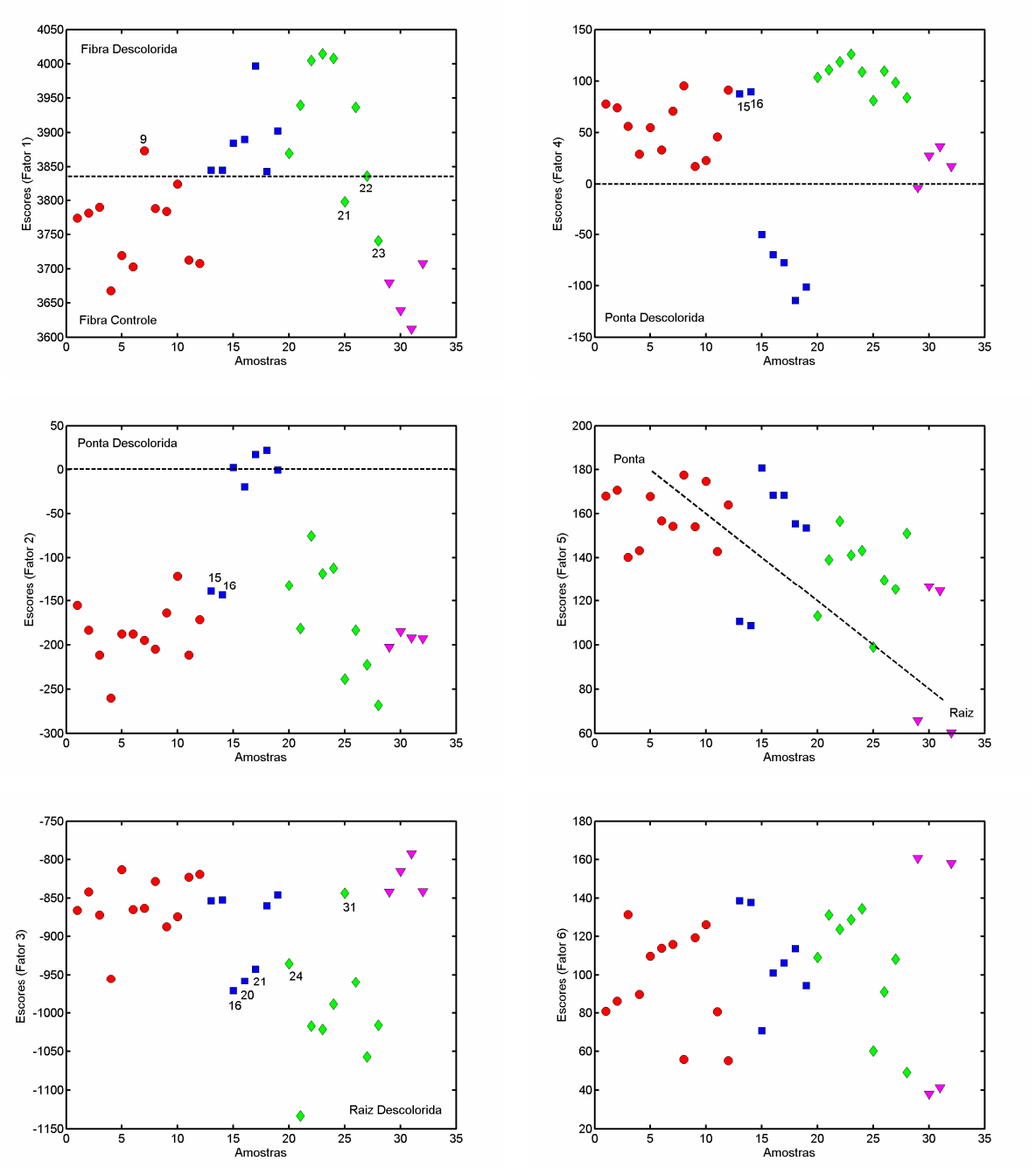


Figura 3. 7: Gráficos dos escores 1, 2, 3, 4 e 5 do Modo *I* para o modelo NPLS. Legenda: (O) ponta virgem, (□) ponta descolorida, (∇) raiz virgem, (◊) raiz descolorida.

O fator 1, que descreve 99,07 % da variação em **X** e 28,32% em **Y**, é responsável pelo efeito do tratamento cosmético e separa as amostras das classes controle e descolorida.

O fator 2 (0,04% em **X** e 14,2% em **Y**) caracteriza a classe ponta descolorida, que apresentou valores de escores próximos de zero neste fator e bem diferentes dos valores para as outras 3 classes.

No fator 3 (0,05% em **X** e 12,6% em **Y**), as amostras da classe raiz descolorida tendem a ocupar a região dos valores mais negativos, quando comparadas às outras amostras do conjunto. As amostras de ponta descolorida estão claramente associadas aos escores negativos do fator 4 (0,03% em **X** e 5,8% em **Y**), e o fator 5 (0,01% em **X** e 14,3% em **Y**) caracteriza a tendência na separação das amostras das classes raiz e ponta.

O fator 6 (0,01% em **X** e 5,7% em **Y**) não apresentou nenhum padrão significativo, mas sua inclusão foi necessária para a correta modelagem da classe 4. As Figuras 3.8 e 3.9 apresentam o gráficos dos pesos para os Modos *J* e *K*, respectivamente.

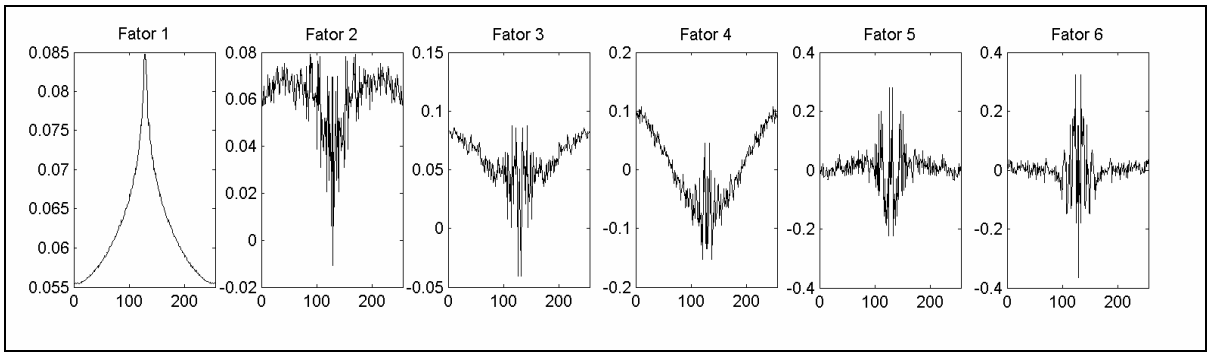


Figura 3. 8: Gráficos dos pesos do Modo *J*.

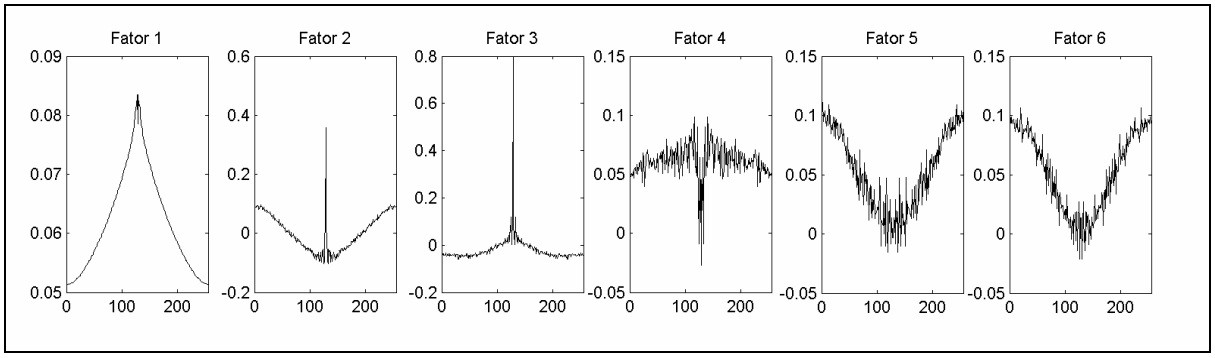


Figura 3. 9: Gráficos dos pesos do Modo *K*.

3.5 Discussão dos Resultados

O método empregado teve como objetivo discriminar as quatro classes de amostras de fibra capilares diretamente a partir das imagens nível de cinza e construir modelos de classificação para identificar posteriormente a classe de novas amostras. O modelo construído foi capaz de classificar corretamente as amostras externas, com exceção de uma amostra pertencente à classe 2 (ponta descolorida), a mais heterogênea do conjunto. Considerando a heterogeneidade natural de fibras capilares, o erro de predição obtido pode ser considerado relativamente baixo. Uma alternativa para construir um modelo mais robusto seria incluir novas amostras no conjunto de calibração.

O método foi capaz ainda de incluir na modelagem as amostras com comportamento atípico detectadas anteriormente pelo método dos descritores, sem prejuízo para a capacidade preditiva do modelo. O comportamento anômalo destas amostras foi claramente identificado no presente estudo pela análise do gráfico dos escores do Modo *I*.

A principal vantagem deste método em relação ao dos descritores é a análise direta a partir das imagens, o que permite sua aplicação no estudo quantitativo de qualquer tipo de imagem nível de cinza.

CAPÍTULO 4

Trabalhando Diretamente com as Imagens – Métodos de Ordem Superior: PARAFAC e MPCA

Características Étnicas: Discriminação Quantitativa de Imagens de Fibras Capilares Provenientes de Diferentes Etnias

4.1 Arranjo dos Dados

O conjunto de dados utilizado é formado por um arranjo tridimensional **X** (36, 256, 256) contendo 36 matrizes das imagens empilhadas de fibras capilares de três etnias diferentes: 11 amostras de fibras caucasianas, 13 de fibras afro e 12 de fibras orientais. Com o intuito de obter um arranjo do tipo OVV[37], o arranjo **X** de imagens foi submetido à transformada de Fourier bidimensional (2D-FFT) seguido de uma transformação logarítmica [Huang *et al.*, 2003]. Os dados não escalados foram modelados com os métodos PARAFAC [Bro, 1998; Bro, 1997] e MPCA [Wold *et al.*, 1987; Henrion *et al.* 1992; Wise *et al.*, 1999].

[37] Cf. Seção 3.2.

4.2 Análise Exploratória: PARAFAC

4.2.1A escolha do número de fatores

A escolha do número de fatores foi realizada pela análise da soma quadrática dos resíduos em **X**, da consistência trilinear e do número de iterações necessárias durante a modelagem. Para tanto, foram testados modelos PARAFAC com todas as amostras, variando-se o número de fatores de 1 a 10, sendo que cada modelagem foi realizada em triplicata para verificar a reprodutibilidade dos parâmetros obtidos. Os resultados podem ser observados na Figura 4.1.

Um modelo com número apropriado de fatores deverá apresentar valores baixos para os resíduos obtidos e com poucas iterações. Em geral, espera-se que as replicatas apresentem valores de resíduos próximos, e um indicativo de que ruído foi modelado ocorre quando os valores dos resíduos para modelos em replicatas utilizando o mesmo número de fatores não são idênticos. Aumentos no número de iterações necessárias podem indicar também que um número muito grande de fatores foi utilizado para a modelagem [Bro, 1998; Bro, 1997].

O teste de consistência trilinear (CORCONDIA[38]) examina a adequação do modelo PARAFAC [Bro & Kiers, 2003]. Modelos apropriados devem fornecer valores de CORCONDIA entre 80 e 100%. Modelos com valores em torno de 50% são instáveis e o uso de restrições no modelo poderá ajudar a estabilizá-lo. Valores próximos de zero ou

[38] *Core Consistency Diagnostic*.

negativos indicam que os dados não podem ser descritos por um modelo trilinear, ou que o número de fatores utilizado é maior que o necessário. Os valores de CORCONDIA devem diminuir com o aumento do número de fatores. Sendo assim, um modelo adequado deverá apresentar o maior número de fatores em que seja possível obter um valor válido de consistência trilinear [Bro, 1998; Bro, 1997].

O modelo com 2 fatores atendeu aos critérios descritos acima (Figura 4.1), tendo fornecido um valor médio de CORCONDIA igual a 98,86% e 99,06% da variância total dos dados [Bro, 1998; Bro, 1997].

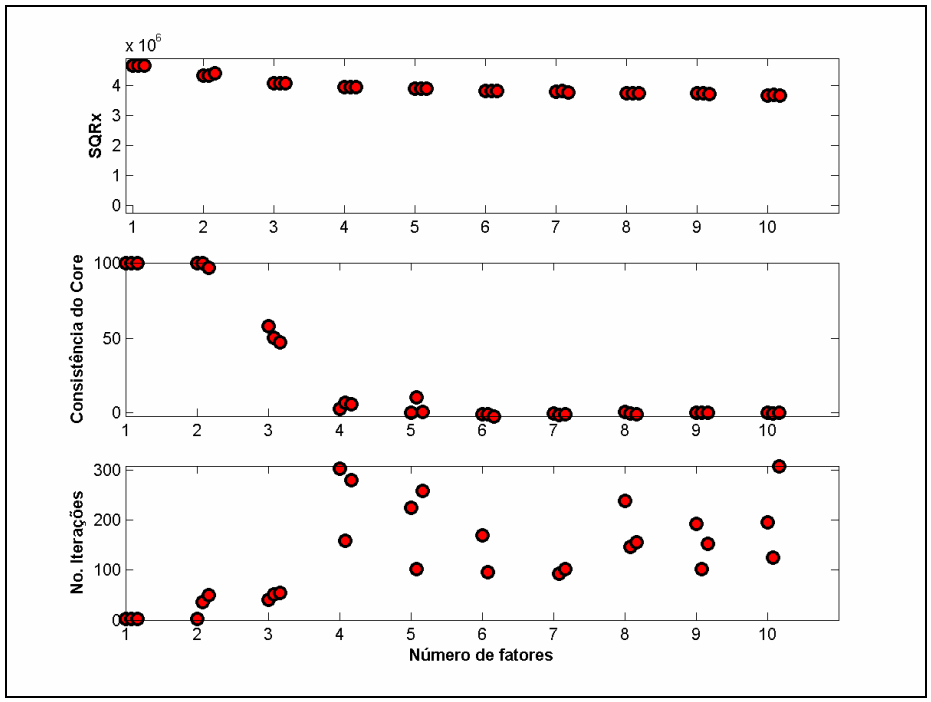


Figura 4. 1: Teste de consistência trilinear para escolha do número de fatores para o modelo PARAFAC (SQRx = soma quadrática dos resíduos de **X**).

4.2.2 Análise dos Dados

Todas as amostras utilizadas no modelo apresentaram valores de *leverage* abaixo do limite de *3k/I* (Figura 4.2). No gráfico dos pesos do Modo *I* (modo das amostras) para os 2 fatores (Figura 4.3), as amostras de cada classe ocupam regiões características e 5 amostras encontram-se em regiões intermediárias ou foram mal classificadas. Estas amostras representam 13,9% do conjunto, e pertencem às classes caucasiana e afro. As fibras capilares possuem uma natureza heterogênea, [Swift & Brown, 1972; Swift, 1999; Smith, 1998; Swift & Smith, 2000; Hadjur *et al.*, 2002; McMullen *et al*., 2000; McMullen & Kelty, 2001], e dentre as fibras de diferentes etnias, as orientais são as mais uniformes [Robbins, 1994; McMullen *et al.*, 2000]. Esta característica pode ser facilmente observada na análise dos *pesos* do modelo PARAFAC, pois as amostras pertencentes a esta classe formaram o grupo mais homogêneo (Figura 4.3). No outro extremo estão as fibras afro, cujas amostras apresentaram maior dispersão. A Figura 4.4 apresenta o gráfico dos pesos dos Modos *J* e *K*.

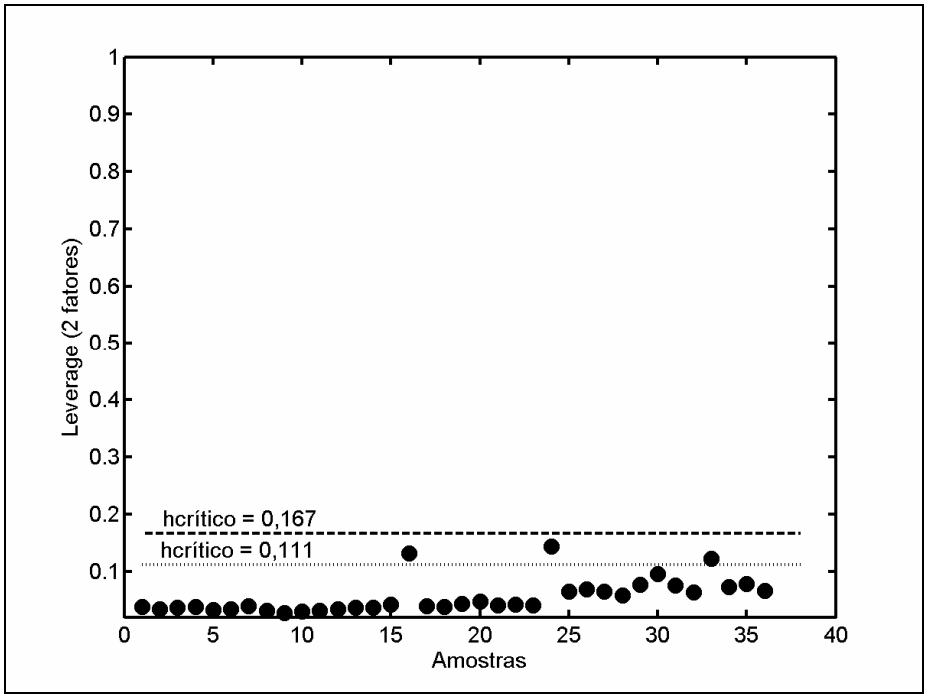


Figura 4. 2: *Leverage* do modelo PARAFAC com 2 fatores.

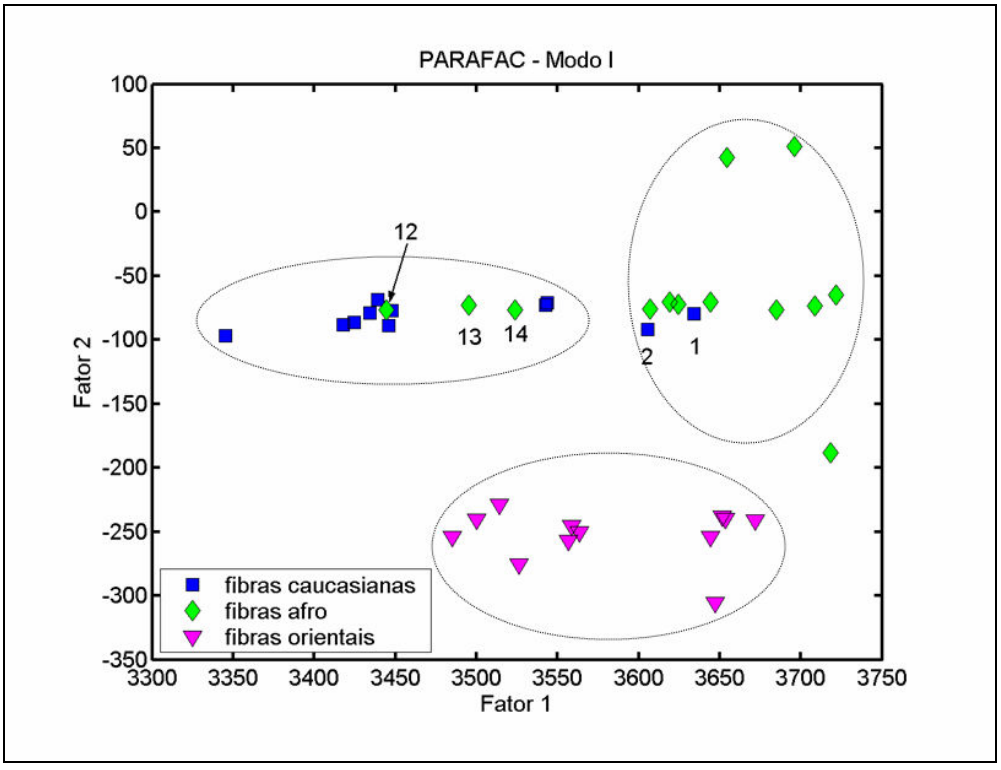
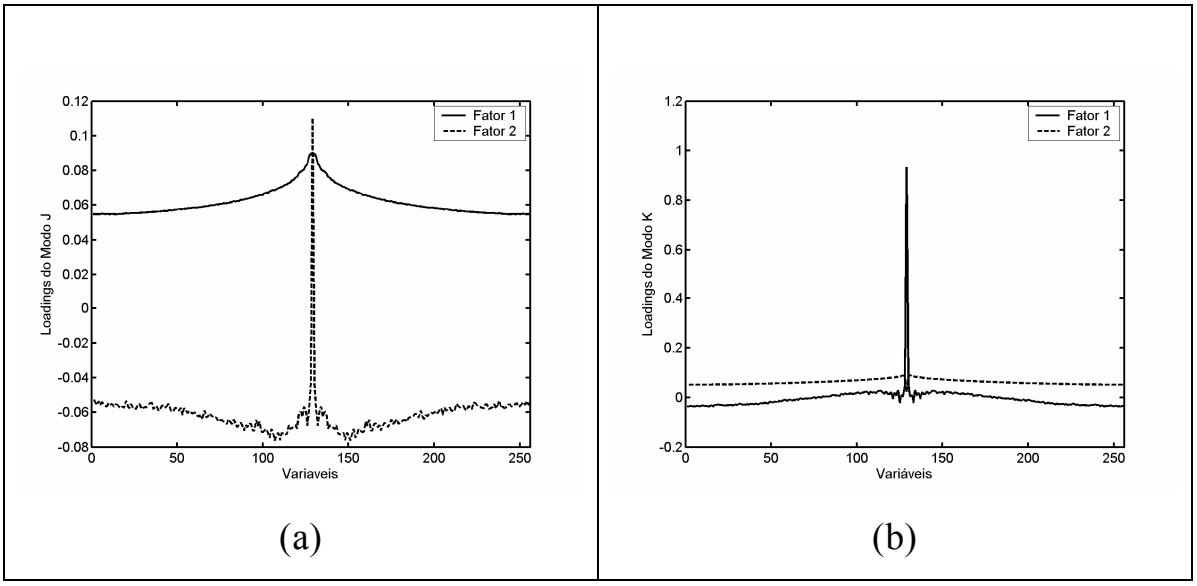


Figura 4. 3: Gráfico dos *pesos* do Modo *I* para o modelo PARAFAC com 2 fatores.



(a) (b)

Figura 4. 4: Gráfico dos *pesos* dos Modos (a) *I* e (b) *K* para o modelo PARAFAC com 2 fatores.

4.3 Classificação Não Supervisionada: MPCA

O método PARAFAC tem sido utilizado com sucesso tanto na análise exploratória, em que o intuito é observar o comportamento dos dados como, por exemplo, agrupamentos, tendências e presença de amostras atípicas, quanto na calibração de segunda ordem de dados espectroscópicos [Bro, 1998; Bro, 1997]. Neste último caso é possível modelar propriedades das amostras, como concentração e propriedades físico-químicas, e posteriormente estimar estas propriedades em novas amostras. Para tanto, o comportamento da propriedade de interesse, caracterizada por uma variável contínua, é modelado em função dos pesos por meio de uma equação linear, o que permite posteriormente realizar a predição destas propriedades para novas amostras. No caso das amostras de imagens este ajuste é inviável, pois a propriedade que

se deseja modelar é a classe, uma variável categórica cuja dependência dos pesos não pode ser descrita adequadamente por uma equação linear. Com o intuito de realizar a classificação de amostras externas, o método MPCA [Wold *et al.*, 1987; Henrion *et al.*, 1992; Henrion 1994; Wise *et al.*, 1999] foi utilizado como alternativa ao método PARAFAC, para construir um modelo de classficação que permitisse posteriormente a classificação de novas amostras. Neste caso, o conjunto de dados **X** foi dividido em dois subconjuntos: o *conjunto de treinamento*, utilizado na construção do modelo MPCA, e o *conjunto de validação externa*, utilizado na etapa de predição.

4.3.1 Seleção das amostras para validação externa

O conjunto de dados foi dividido em dois subconjuntos, um com 30 amostras que foi utilizado como conjunto de treinamento e outro com 6 amostras, utilizado para validação externa. Para selecionar as amostras de validação, cada classe foi modelada separadamente com o método PARAFAC, e duas amostras foram selecionadas com base nos valores obtidos dos *pesos* do Modo *I* e *leverage* (Figura 4.5). Procurou-se selecionar para o conjunto de validação, amostras com teores intermediários de *leverage* e *pesos*.

Os três modelos apresentaram 99,07% (fibras caucasianas), 99,15% (fibras afro) e 99,12% (fibras orientais) de variação explicada com apenas 2 fatores, e foram selecionadas as amostras 10, 11 (caucasianas), 23, 24 (afro), 35 e 36 (orientais). Duas amostras das classes afro e oriental apresentaram valores de *leverage* acima do limite de *2k/I*, no

entanto, estas amostras estão dentro do limite de 3k/I e foram mantidas no conjunto de dados (Figura 4.5).

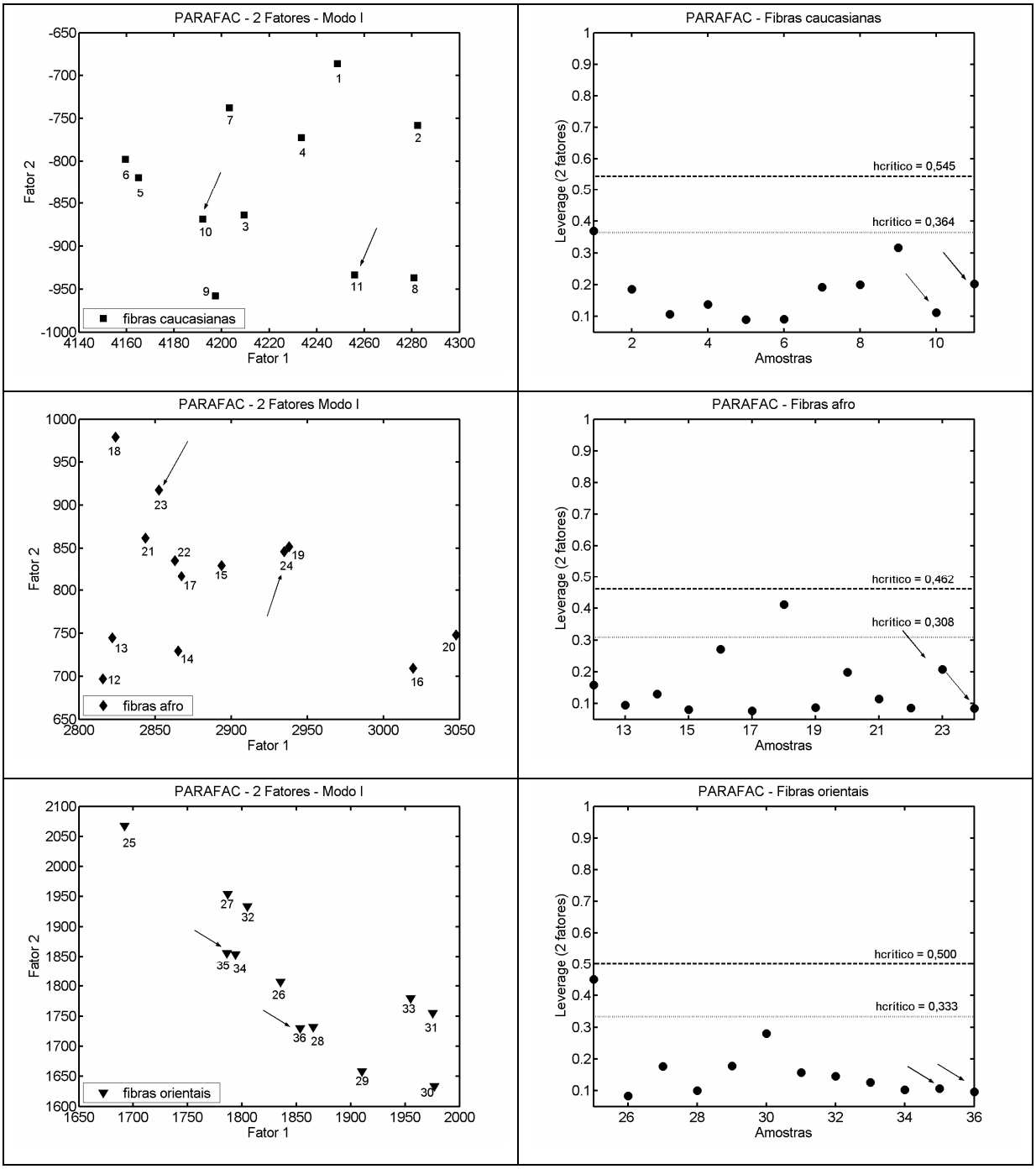


Figura 4. 5: Pesos (Modo das amostras) e leverage dos modelos PARAFAC para cada classe de fibras capilares, utilizados na seleção de amostras para validação externa.

4.3.2 Modelagem

O método MPCA desdobra o arranjo de dados $\underline{\mathbf{X}}$ (I, J, K) em uma matriz $\mathbf{X}$ (I, JK), que então é modelada utilizando PCA [Wold *et al.*, 1987; Henrion *et al.*, 1992; Henrion 1994; Wise *et al.*, 1999]. O desdobramento pode ser realizado em qualquer dimensão, mas no presente estudo ele foi feito com os modos J e K, pois o objetivo foi investigar o modo das amostras (I).

Inicialmente cada coluna da matrix $\mathbf{X}$ contendo os dados desdobrados é centrada na média, e somente depois a decomposição é realizada. A MPCA explica a variação das variáveis medidas pelas suas trajetórias médias. Ao subtrair a trajetória média de cada variável, o que é realizado ao centrar na média as colunas da matriz desdobrada, ocorre a remoção de não linearidades dos dados ao longo do eixo que foi desdobrado. Desta maneira, os *escores* da MPCA descreverão o comportamento das amostras (tendências, agrupamentos e anomalias), e os *pesos* descreverão a variação das variáveis medidas relacionadas às suas trajetórias médias [Wise *et al.*, 2003].

A Figura 4.6 apresenta a porcentagem de variância descrita em cada fator, cujos valores são relativamente baixos. Isto se deve ao fato de que durante a MPCA, a trajetória média modela a maior parte da variação normal dos dados, e esta média é extraída antes do cálculo do modelo PCA nos dados desdobrados, resultando desta maneira numa menor porcentagem de variação descrita [Wise *et al.*, 2003]. A partir destas considerações, e pela análise da Figura 4.6, escolheu-se o modelo com 4 fatores.

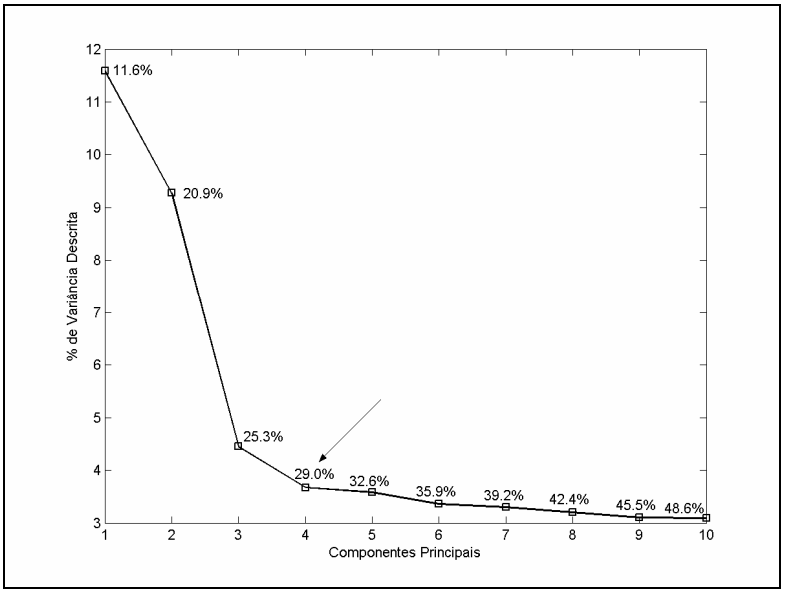


Figura 4. 6: Porcentagem de variância descrita nas componentes principais.

Com este modelo, a amostra 8 (fibra caucasiana) apresentou valor de *leverage* acima de *3k/I* (Figura 4.7). Esta amostra foi excluída e o modelo refeito, após o qual, a amostra 9 ficou com valor de *leverage* acima deste limite. Este comportamento é conhecido como "bola de neve" [ASTM1655]. A norma ASTM, que estabelece os procedimentos para o uso de métodos de calibração multivariada em dados espectroscópicos, orienta que nestes casos deve-se utilizar o primeiro modelo construído após a retirada da amostra com *leverage* alta, desde que neste modelo nenhuma das amostras restantes apresente valor de *leverage* acima de 0,5 [ASTM1655].

Seguiu-se a orientação da norma ASTM para os dados de imagem deste trabalho, e novas modelagens foram testadas até que nenhuma amostra apresentasse *leverage* acima do valor indicado. O melhor modelo foi obtido após a exclusão das amostras 8 e 9, e apresentou 28,7% da

variância descrita com 4 componentes, cujos pesos e escores podem ser observados respectivalmente nas Figuras 4.8 e 4.9.

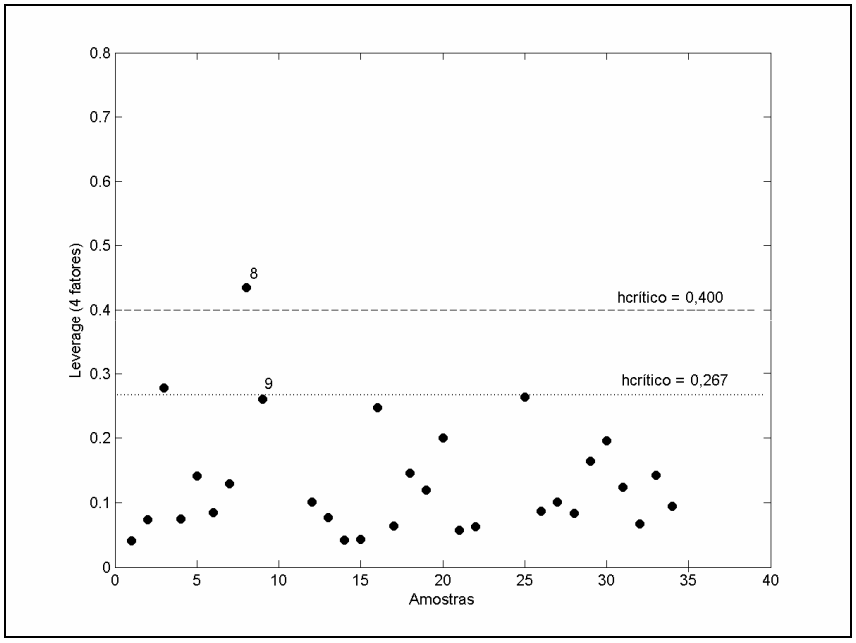


Figura 4. 7: Gráfico de *leverage* para o modelo com 4 fatores (30 amostras do conjunto de treinamento).

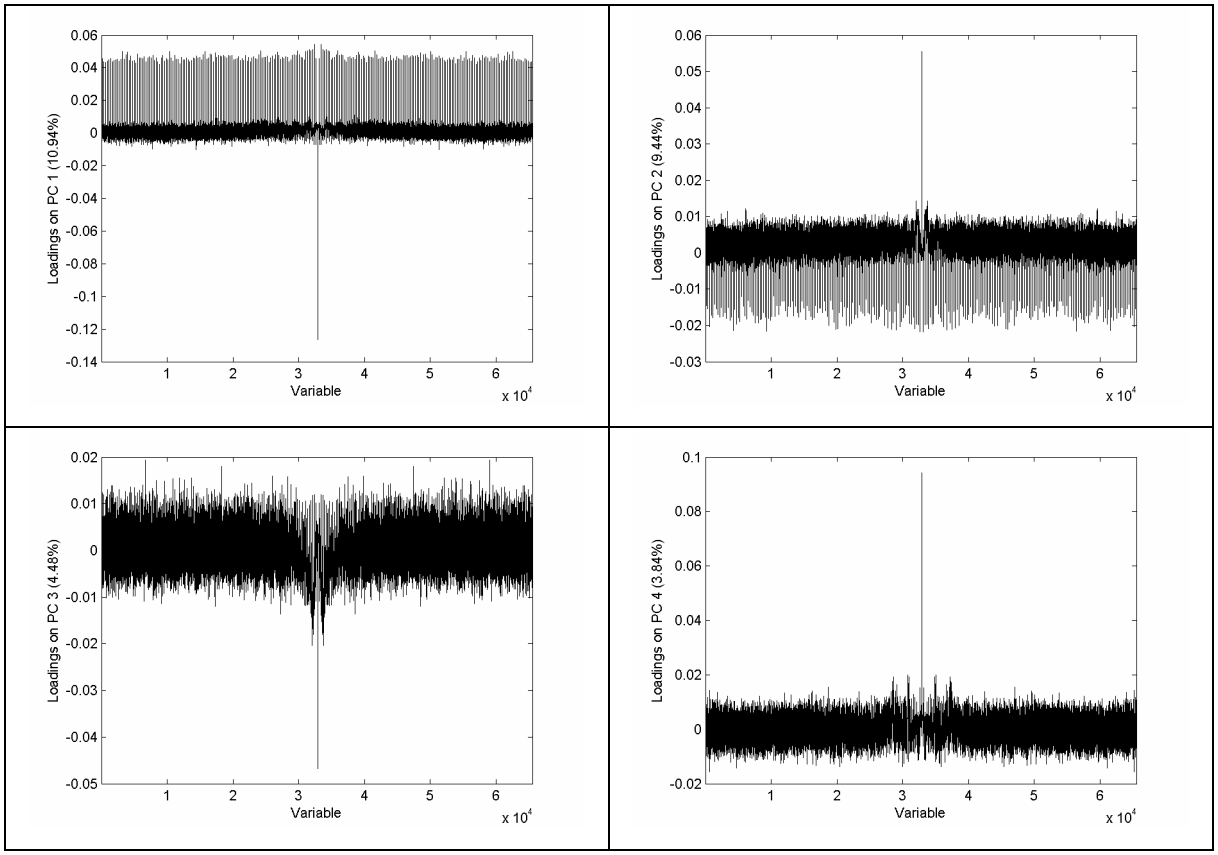


Figura 4. 8: *Pesos* das componentes principais 1, 2, 3 e 4.

4.3.3 Validação Externa

O poder preditivo do modelo foi testado realizando a classificação de seis amostras externas, sendo duas de cada classe, por meio da estimativa dos escores destas amostras a partir dos pesos gerados durante a etapa de modelagem. A projeção dos escores das novas amostras no espaço descrito pelos escores das amostras do conjunto de treinamento permitiu classificar com sucesso as seis amostras de validação (Figura 4.9), apesar do conjunto de treinamento ter apresentado algumas amostras em regiões intermediárias ou mal classificadas.

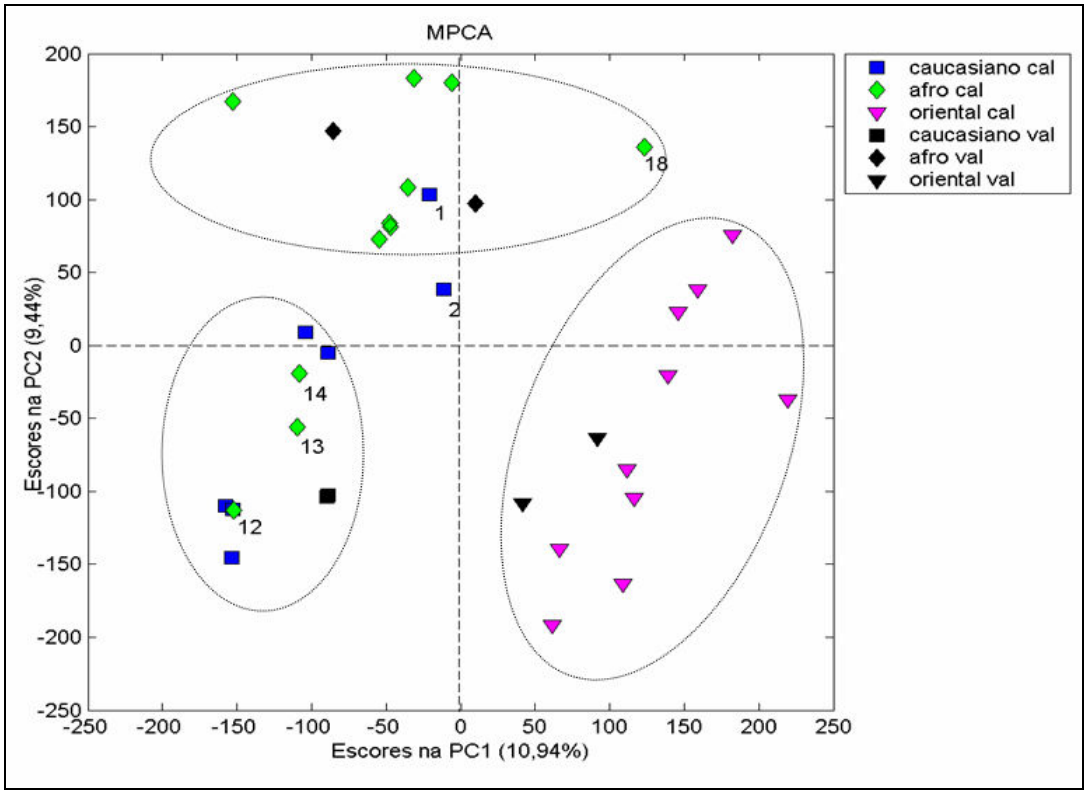


Figura 4. 9: Escores nas componentes 1 e 2, para as amostras dos conjuntos de treinamento e de validação externa, para o modelo MPCA com 4 fatores.

4.4 Discussão dos Resultados

A classificação geral de fibras capilares é baseada em três etnias majoritárias: negróide, caucasiana e mongol. Apesar da natureza extremamente heterogênea das fibras capilares, cada etnia apresenta aspectos característicos, cujos padrões podem ser facilmente reconhecidos. Das três classes, cabelos caucasianos em geral apresentam o menor diâmetro, e podem apresentar diversas graduações em termos de ondulação, que vão do liso ao muito ondulado. Fibras negróides apresentam desvios na circularidade, que levam ao formato encarapinhado, e fibras do tipo mongol, por outro lado, possuem um

formato mais circular e maior raio de curvatura, que determinam o aspecto extremamente liso. Estas últimas apresentam ainda bordas serrilhadas e maior número de cutículas por área.

As características de cada classe resultam em agrupamentos de amostras bem definidos no gráfico dos escores tanto do modelo PARAFAC (Figura 4.3) quanto do modelo MPCA (Figura 4.9).

Nestes gráficos é possível observar que as fibras orientais, que são por natureza mais homogêneas, apresentaram um agrupamento bem distinto das outras duas classes de amostras. As fibras afro, que apresentam uma variabilidade maior, estão mais dispersas e com algumas amostras (12, 13 e 14) na região ocupada pelas amostras caucasianas. Duas amostras de fibras caucasianas (1 e 2) encontram-se ainda na região ocupada pelas fibras afro.

A presença de amostras em regiões de outras classes é resultante da heterogeneidade natural das fibras capilares, mas esta característica não comprometeu a capacidade preditiva do modelo, pois todas as seis amostras do conjunto de validação foram corretamente classificadas.

CAPÍTULO 5

Conclusão Geral

Este trabalho teve como objetivo a construção de modelos quantitativos para a análise de imagens *nível de cinza*, utilizando métodos quimiométricos, em especial os métodos de ordem superior. Estes métodos constituem poderosas ferramentas na extração de informações quantitativas de imagens, principalmente devido à possibilidade de analisar diversas imagens simultaneamente, o que permite a construção de modelos de regressão e reconhecimento de padrões, que poderão ser utilizados para a predição de propriedades ou classificação de novas amostras.

Para este estudo foram utilizadas imagens de Microscopia de Força Atômica (AFM) de fibras capilares de dois conjuntos: um com fibras submetidas ao tratamento de descoloração com peróxido de hidrogênio e persulfato de amônio, e um outro conjunto, com fibras de diferentes etnias.

A descoloração das fibras com peróxido e persulfato tem sido amplamente descrita na literatura cosmética, e seu efeito tanto na superfície da fibra quando na sua estrutura interna foi documentado com diversas técnicas tais como MEV, AFM, TEM, DSC, TGA e Espectroscopia Raman. Os aspectos morfológicos das fibras capilares de diferentes etnias também são bem conhecidos, e para efeitos de estudo e classificação, as fibras são divididas em três etnias majoritárias (negróide, caucasiana e mongol), cada uma com características distintas. No entanto, estas características representam apenas

tendências nas etnias puras, pois fibras capilares apresentam um aspecto naturalmente heterogêneo.

Diversas estratégias foram testadas para a análise de imagens *nível de cinza* de fibras capilares, tendo como objetivo a sua discriminação quantitativa.

No capítulo 2 foi realizado um estudo da superfície das fibras capilares por meio da estimativa de descritores baseados em parâmetros morfológicos, tais como quantidade de cutículas por área superficial, ângulos internos e externos da abertura das cutículas e rugosidade. Fibras de diferentes classes tendem a apresentar descritores capilares característicos, e com base nestas informações, modelos de classificação foram construídos utilizando os métodos SIMCA e NPLS.

Primeiramente testou-se o uso do método SIMCA em uma matriz de dados contendo a média e o desvio padrão dos descritores capilares. No entanto, esta estratégia poderá gerar lacunas na representação das características morfológicas da fibra capilar, devido à distribuição heterogênea destes descritores ao longo da superfície da fibra. Na tentativa de inserir esta distribuição espacial no modelo, utilizou-se o método NPLS, no qual uma das dimensões do arranjo de dados apresenta a estimativa da média e do desvio padrão dos descritores ao longo de toda a superfície da fibra. Ambos os métodos foram capazes de identificar amostras com comportamento anômalo e classificar corretamente novas imagens. No entanto, o método NPLS apresentou altos erros durante a etapa de validação cruzada *leave-one-out*, devido à heterogeneidade das amostras e ao pequeno número de amostras de algumas classes.

No capítulo 3, o método NPLS foi utilizado no mesmo conjunto de dados, mas diretamente nas imagens após aplicar a transformada de Fourier, e os resultados mostraram-se superiores ao modelo construído com os descritores. Na modelagem anterior, algumas amostras haviam sido identificadas como atípicas devido à limitação do algoritmo de estimar determinados descritores em regiões muito deterioradas e com falta de cutículas. Na modelagem utilizando as imagens puras, o comportamento diferenciado destas amostras foi identificado, porém a sua inclusão no modelo foi possível, sem prejuízo para a sua capacidade preditiva.

No capítulo 4, fibras caucasianas, afro e orientais foram utilizadas para representar as três etnias majoritárias, na construção de modelos de reconhecimento de padrões por meio do uso dos métodos PARAFAC e MPCA. A análise exploratória realizada com o método PARAFAC foi capaz de discriminar as três classes de amostras com sucesso, bem como de auxiliar na seleção das amostras de treinamento e de validação. O método MPCA permitiu a correta classificação de um conjunto de amostras externas, apesar da presença de algumas amostras anômalas no conjunto de treinamento. Os dois métodos foram capazes de identificar também o grau de heterogeneidade característico de cada classe, por meio da maior ou menor dispersão dos escores das amostras, o que concorda com a informação prévia da heterogeneidade natural de fibras capilares.

De uma forma geral, os métodos apresentados neste trabalho permitiram discriminar com sucesso as diversas classes de amostras, bem com identificar tendências e amostras com comportamento atípico. Com exceção dos modelos construídos com os descritores, que são

específicos para fibras capilares, os modelos construídos com as imagens puras poderão ser aplicados a qualquer tipo de imagem *nível de cinza*. Estes métodos apresentam amplo potencial para a construção de modelos de classificação e calibração, que uma vez validados, poderão ser utilizados para a predição de propriedades e classificação de futuras imagens.

APÊNDICE:

Introdução aos Métodos Quimiométricos:

O texto a seguir apresenta uma breve descrição dos principais métodos quimiométricos e os casos nos quais cada um deles se aplica. A intenção não é expor minuciosamente os fundamentos algébricos envolvidos, e apenas os métodos utilizados neste trabalho são detalhados. Para o estudo aprofundado do tema, há inúmeras e ótimas publicações na literatura científica, algumas das quais serão citadas como referência bibliográfica ao longo do texto.

A.1 Estrutura dos Dados e Métodos de Análise Multivariada

Os métodos multivariados podem ser classificados de acordo com a dimensão dos dados analisados [Olivieri *et al.*, 2006; Sanchez & Kowalski, 1986, 1988, 1990]. Dados univariados fornecem como resposta um escalar, que na linguagem algébrica são classificados como tensor de ordem zero (Figura I.1). Neste caso, a análise dos dados é feita pela observação do comportamento de uma única variável de cada vez, como por exemplo, a concentração da espécie de interesse ou uma propriedade físico-química (densidade, viscosidade, ponto de fusão, pH, ponto de ebulição, coeficiente de partição octanol-água). Nos dados multivariados, é possível analisar diversas variáveis simultaneamente, gerando estruturas de dados mais complexas e que podem apresentar-se como um vetor (tensor de ordem 1), uma matriz (tensor de ordem 2) ou

um cubo (tensor de ordem 3) e assim por diante. Exemplos deste tipo de resposta são descritos respectivamente por espectros (vetor), imagens *nível de cinza* (matriz de pixels x pixels) ou superfícies de respostas geradas por instrumentos hifenados tais como GC/MS e LC/MS (matriz formada pelo espectro de massa em cada tempo de retenção), e imagens multivariadas (imagens formadas pelos pixels x pixels x canais de cores).

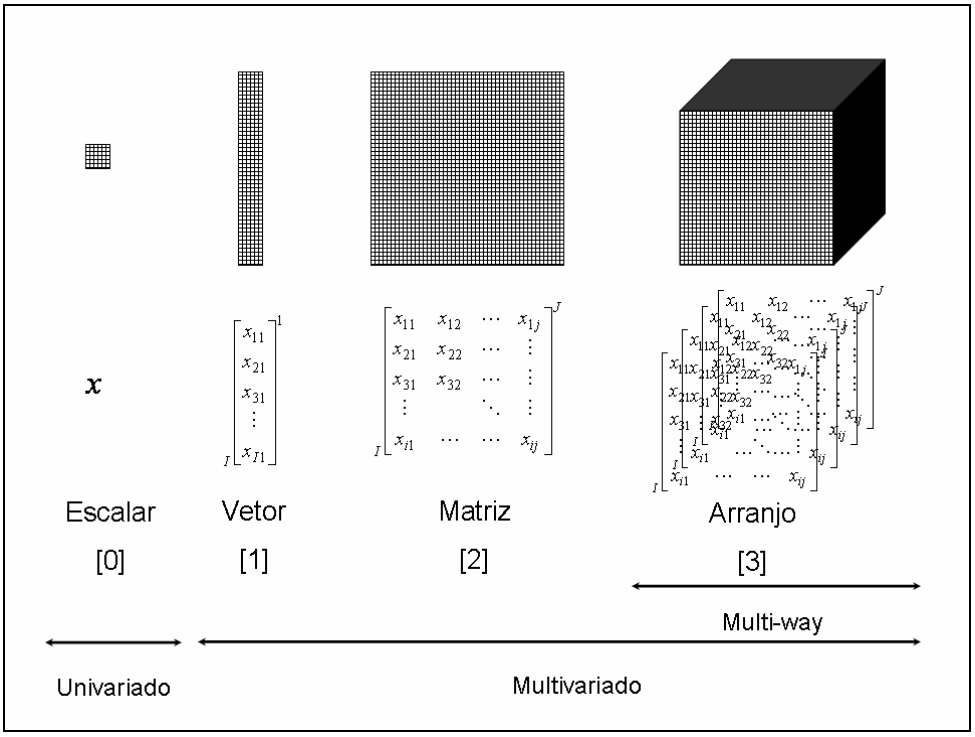


Figura I. 1: Representação da dimensão dos dados com as respectivas ordens entre colchetes. Adaptado de Olivieri *et al.* (2006).

A Figura I.1 ilustra estes arranjos de dados, e os métodos de análise adequados a cada estrutura: métodos estatísticos univariados e métodos multivariados, sendo que estes últimos podem ainda ser divididos em

métodos de primeira ordem e métodos de ordem superior, ou métodos *multi-way*.

Por convenção, escalares são representados por letra minúscula em itálico (x), vetores por letra minúscula em negrito ($\mathbf{x}$), matrizes por letra maiúscula em negrito ($\mathbf{X}$) e arranjos de ordem superior são representados por letra maiúscula em negrito e sublinhada ($\underline{\mathbf{X}}$) [Bro, 1998].

De uma forma geral, os métodos de análise multivariada podem ser utilizados com três propósitos definidos [Huang *et al.*, 2003]: (i) descrição dos dados, ou análise exploratória; (ii) classificação e discriminação; e (iii) correlação e regressão. Esta divisão não contempla todas as possibilidades da aplicação destes métodos, mas sistematiza e orienta a escolha da estratégia de análise.

A.2 Métodos de Primeira Ordem ou Métodos Multivariados[39]

Nos métodos de primeira ordem, os dados são organizados em uma estrutura matricial $\mathbf{X}$ (I, J), constituída por I amostras e J variáveis. Esta estrutura implica que cada amostra i é caracterizada por um conjunto de J variáveis. De uma forma geral, estes métodos podem ser classificados em métodos de reconhecimento de padrões e métodos de calibração multivariada, e as suas diferentes subdivisões podem ser observadas na Figura I.2 [Beebe, 1998; Infometrix Inc., 1990-2007].

[39] Por definição, tanto os métodos de primeira quanto segunda ordem são métodos multivariados. No entanto, alguns autores utilizam o termo *Métodos Multivariados* para descrever apenas os métodos de primeira ordem, no qual arranjos matriciais de dados são utilizados. Neste contexto, os métodos de segunda ordem são mais conhecidos por Métodos *Multi-way*, *N-way* ou Multi-modos.

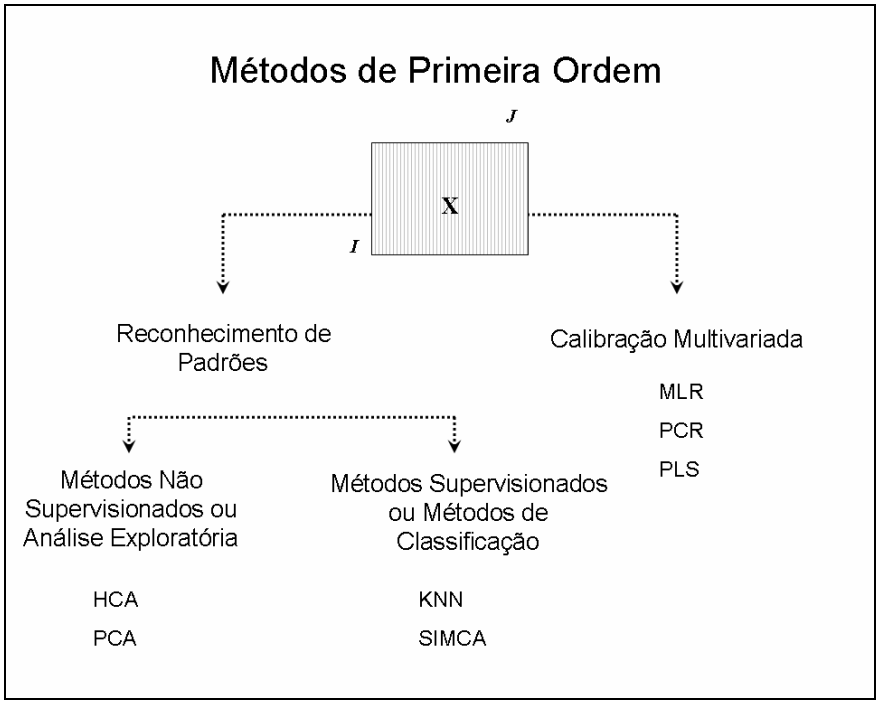


Figura I. 2: Fluxograma dos métodos de primeira ordem.

A.2.1 Reconhecimento de Padrõe Não Supervisionados ou Análise Exploratória

Os métodos de reconhecimento de padrões são utilizados para detectar similaridades e diferenças entre grupos de amostras, e também permitem a classificação de novas amostras em grupos pré-estabelebidos por meio da construção de modelos. São divididos em métodos não supervisionados e métodos supervisionados [Figura I.2]. Os primeiros são utilizados para uma análise prévia dos dados, no intuito de detectar a formação de agrupamentos e correlações entre amostras, e a presença de amostras com comportamento atípico. Estes métodos são denominados não supervisionados porque os algoritmos utilizados não são previamente direcionados pela informação da

existência de classes. Os métodos atualmente mais conhecidos e difundidos deste tipo são: a análise de agrupamentos hierárquicos (HCA[40]) e a análise de componentes principais (PCA[41]).

A HCA utiliza o perfil das amostras para agrupá-las segundo seus diferentes e hierárquicos graus de similaridade. A estimativa desta similaridade é realizada pela medida da distância entre as amostras num espaço *n*-dimensional [Beebe, 1998; Infometrix Inc., 1990-2007; Adamson & Bawden, 1981]. O resultado é fornecido na forma de um dendrograma[42], que apresenta os ramos formados pelos agrupamentos das amostras e a hierarquia presente entre eles (Figura I.3).

---

[40] *Hierarchical Cluster Analysis*.

[41] *Principal Component Analysis*.

[42] O termo *Dendograma* é muito popular na língua portuguesa para definir esta técnica de agrupamento, e foi traduzido do termo *dendogram* da Língua Inglesa. Da mesma forma, *dendrograma* provém de *dendrogram*. Uma busca pelos quatro verbetes no site de buscas *Google* (www.google.com, acesso em 12/07/2007), considerado um dos mais completos, revelou os seguintes números de entradas: dendograma (19800), dendrograma (43700), dendogram (923000) e dendrogram (732000). A mesma busca realizada no banco de dados *Web of Science*, um dos mais respeitáveis da área científica, revelou os seguintes números de artigos rescuperados: dendogram (85) e dendrogram (1789). Para o termo *dendogram*, o artigo mais antigo foi de 1985 [Bucher *et al.*, 1985], enquanto para o termo *dendrogram*, o artigo mais antigo recuperado foi de 1971 [Phipps, 1971]. Então foi realizada uma pesquisa dos dois termos em português no *Novo Dicionário Eletrônico Aurélio* (versão 5.11©, 2004 by Regis Ltda) e nenhum dos dois verbetes foi encontrado, tendo sido entretanto encontrados os termos: **dendrite** ([Do gr. dendrítes, 'de uma árvore ou relativo a ela'.] Substantivo feminino. 1.Dendrolite. 2.Histol. Prolongamento de neurônio, e que pode ser numeroso, especializado na função de receber estímulos ambientais, de células epiteliais sensoriais ou de outros neurônios.); **dendrito** ([De dendr(o)- + -ito2.] Substantivo masculino. 1.Geol. Deposição arborescente, em rochas, em virtude da infiltração de águas carregadas de óxido de ferro, manganês, etc.); **dendr(o)-** ([Do gr. dendro- < gr. déndron, ou.] Elemento de composição. 1.= 'árvore'; 'caule', 'haste': dendrite, dendroclasta. [Equiv.: -dendr(o)-, -dendro: acrodendrofilia, clerodendro, siringodendro.]. Os dois verbetes em português foram ainda procurados na Wikipedia, uma enciclopédia livre e disponível na *Internet* e os resultados foram: **Dendrograma** (dendro = árvore) é um tipo específico de diagrama ou representação icónica que organiza determinados factores e variáveis. Resulta de uma análise estatística de determinados dados, em que se emprega um método quantitativo que leva a agrupamentos e à sua ordenação hierárquica ascendente - o que em termos gráficos se assemelha aos ramos de uma árvore que se vão dividindo noutros sucessivamente. Isto é, ilustra o arranjo de agrupamentos derivado da aplicação de um "algoritmo de clustering" (ver agrupamento de dados); **dendograma**: não foram encontrados resultados para o termo pesquisado.

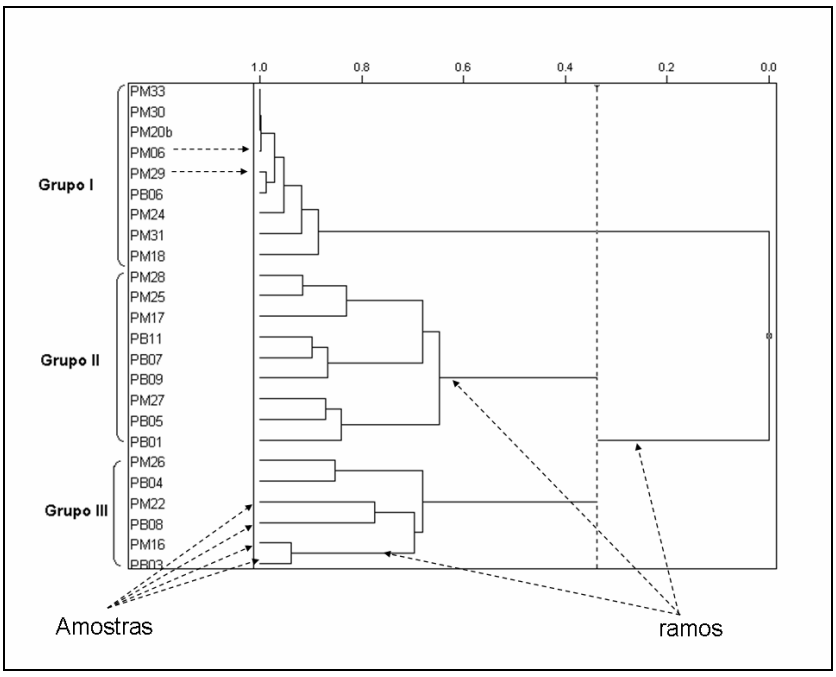


Figura I. 3: Exemplo de um dendrograma, mostrando a formação de três agrupamentos de amostras e a hierarquia entre eles.

A PCA se baseia na projeção das amostras em um novo sistema de eixos, mais favorável à visualização dos dados, que é construído a partir das variáveis originais. Este método é de particular importância, pois é a base para o desenvolvimento de diversos outros métodos multivariados, portanto será visto em mais detalhes [Wold *et al.*, 1987; Beebe, 1998; Infometrix Inc., 1990-2007].

PCA

Para entender o cálculo da PCA, considere uma matriz de dados **X** (*I*, *J*), em que as amostras são descritas pelas *I* linhas e as variáveis são descritas pelas *J* colunas (Figura I.4). Exemplos deste tipo de arranjo podem ser: os *J* comprimentos de onda dos espectros das *I* amostras; as

concentrações de *J* espécies de interesse para as *I* amostras; os *J* descritores moleculares para as *I* amostras; etc. Sendo assim, qualquer conjunto de dados com esta estrutura pode ser descrito por *J* informações diferentes, e em termos algébricos, dizemos que este conjunto de dados poderá ser descrito num espaço de *J* dimensões.

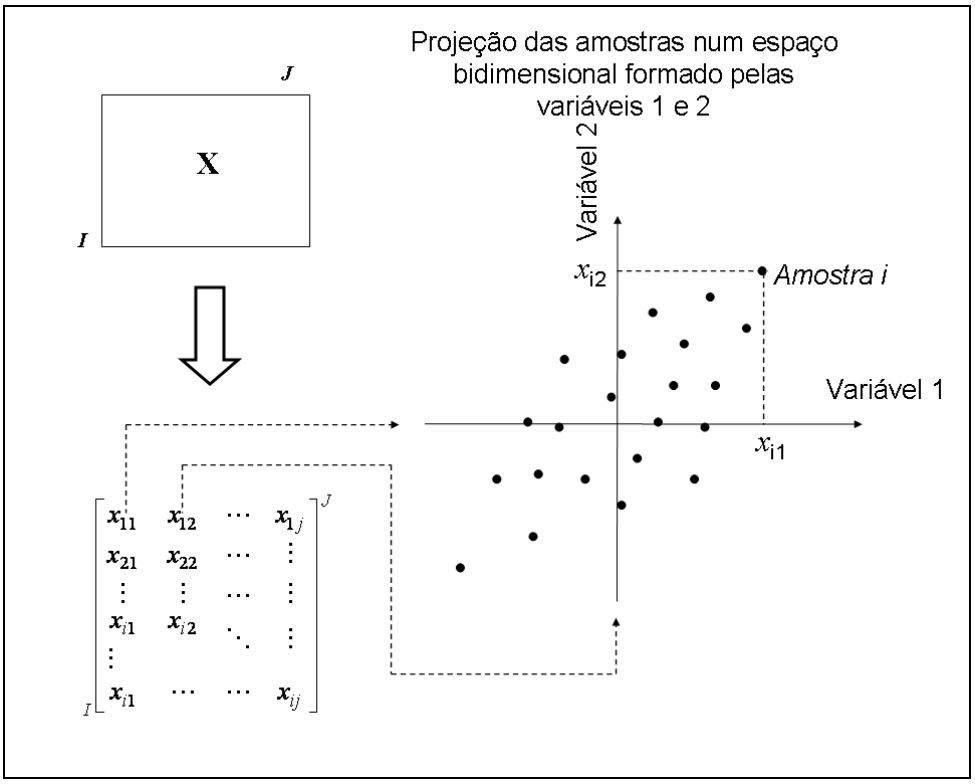


Figura I. 4: Matriz de dados **X** (*I*, *J*) e a projeção das amostras num espaço bidimensional formado pelas variáveis 1 e 2.

Pelo uso da estatística univariada é impossível analisar simultaneamente a influência das *J* variáveis no comportamento das *I* amostras, e corre-se o risco de perder informações importantes durante a análise dos dados. Uma forma de reduzir a dimensão dos dados e enfatizar as informações relevantes é reconstruir este conjunto de dados em um novo sistema de eixos, mais adequado à sua visualização. Durante a PCA, as amostras

são projetadas num novo sistema de eixos construídos a partir da combinação linear das *j*-variáveis originais. Estes novos eixos são denominados fatores ou componentes principais (algumas vezes denominadas PC, do inglês *Principal Components*), e são definidos de tal forma que a primeira componente principal descreve a maior parte da variabilidade dos dados, e as sucessivas componentes descreverão quantidades decrescentes da variabilidade restante, até que toda a informação presente nos dados tenha sido descrita. Estas componentes são ortogonais (perpendiculares) entre si, isto é, são totalmente não correlacionadas. Desta forma, cada eixo contém a informação de todas as variáveis, mas as informações mais importantes se concentram nas primeiras componentes principais.

Pela escolha adequada do número de componentes principais a ser utilizado durante a modelagem, é possível utilizar e visualizar toda a informação importante para a análise, e eliminar as informações redundantes ou o ruído presente. Desta forma, a visualização dos dados neste sistema de novos eixos permite a análise de todas as variáveis simultaneamente em um número menor de fatores.

Um exemplo clássico para introduzir os fundamentos da PCA é apresentado nas Figura I.5(a) e I.5(b), no qual um conjunto de amostras é projetado num espaço formado por duas variáveis. A primeira componente principal (PC1) indica o sentido da maior variabilidade dos dados (Figura I.5(a)), enquanto a variabilidade restante é descrita pela segunda componente principal (PC2) (Figura I.5(b)). Apesar das variáveis 1 e 2 fornecerem a exata localização de cada amostra, apenas pela observação da PC1 é possível ter uma boa idéia da sua dispersão.

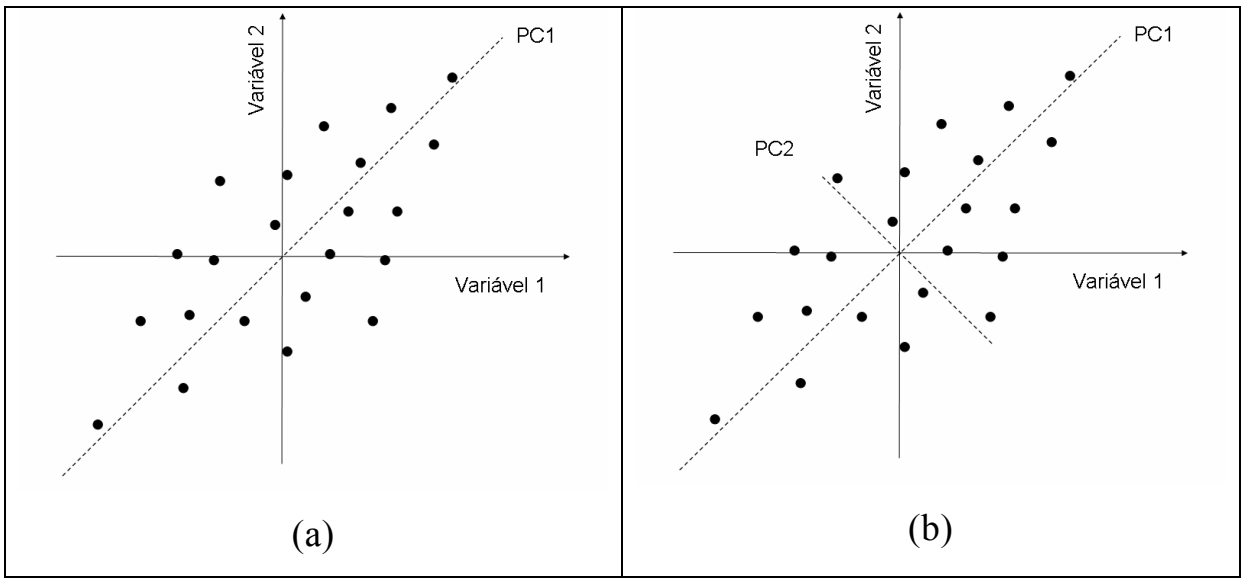


Figura I. 5: Projeção da primeira (a) e da segunda (b) componente principal para um conjunto de dados em um espaço de duas dimensões.

A PCA decompõe a matriz de dados **X** (*I*, *J*) em duas matrizes: uma matriz de escores **T** (*I*, *J*) e outra matriz de pesos **P** (*J*, *J*)[43]. Esta decomposição está representada na Figura I.6, em que o índice da matriz $\mathbf{P}^{\mathbf{T}}$ significa que esta matriz é transposta[44]. Os escores são as coordenadas das amostras projetadas neste novo sistema de eixos (Figuras I.7(a) e I.7(b)), e os pesos são descritos pelo coseno do ângulo entre a variável original e o novo eixo (Figuras I.7(c) e I.7(d)). Sendo assim, os escores contêm as informações das amostras, como por exemplo, a formação de agrupamentos e a presença de amostras atípicas, e os pesos contêm a informação das variáveis. Pela análise simultânea dos escores e dos pesos, é possível identificar quais são as variáveis que influenciam o comportamento das amostras. Pela escolha

[43] Há uma tendência na literatura científica brasileira de utilizar os termos em português. No entanto, alguns termos ainda são mais conhecidos pelos nomes em inglês. É o caso dos escores e pesos, que podem ser encontrados em diversos textos científicos pelos nomes em inglês *scores* e *loadings*.

adequada do número *k* de componentes principais, a dimensão dos escores e pesos diminui respectivamente para **T** (*I*, *K*) e **P** (*J*, *K*), gerando a matriz de resíduos **E** (*I*, *J*), que contém a informação irrelavante. A escolha do número de componentes a ser utilizado na modelagem é realizada durante a etapa de otimização do modelo PCA.

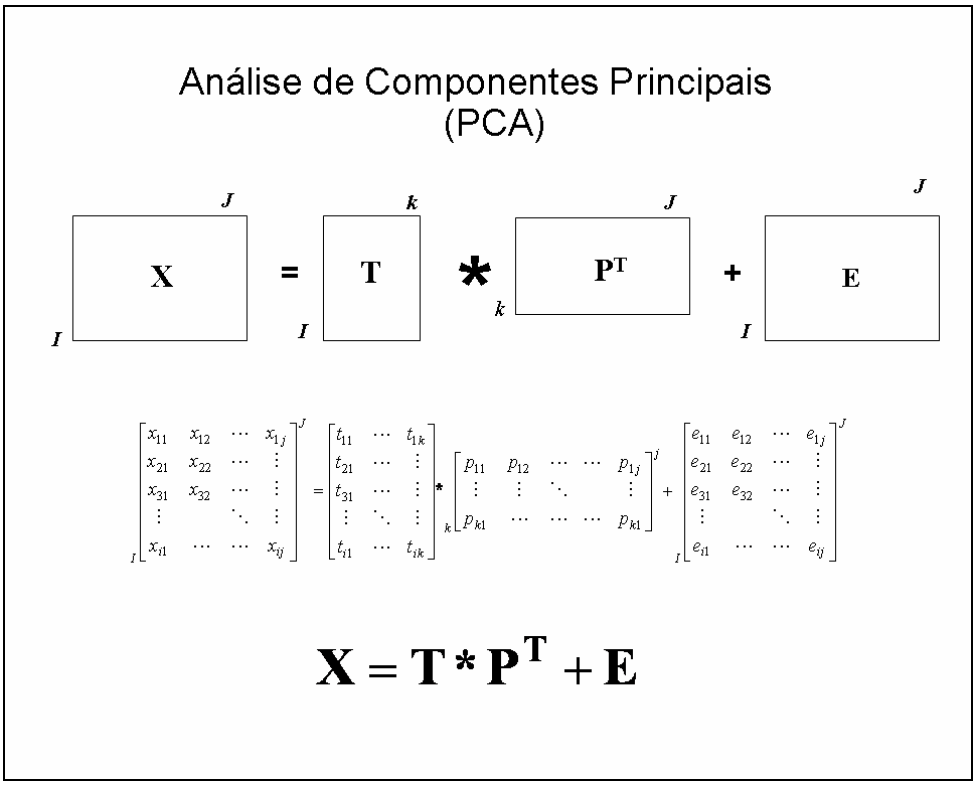


Figura I. 6: Decomposição da matriz **X** nas matrizes dos escores e pesos durante a análise de componentes principais.

A análise de componentes principais pode ser utilizada tanto para propósitos exploratórios quanto classificatórios e o seu cálculo pode ser realizado por diversos algoritmos, sendo os mais comuns a

[44] A transposta de uma é obtida quando permutados linhas e colunas. Por exemplo, a transposta de uma matriz **X** de dimensão *I* x *J* é a matriz $\mathbf{X}^{\mathbf{T}}$ de dimensão *J* x *I*.

decomposição de valor singular, a diagonalização da matriz de correlação e o NIPALS[45] [Wold *et al.*, 1987].

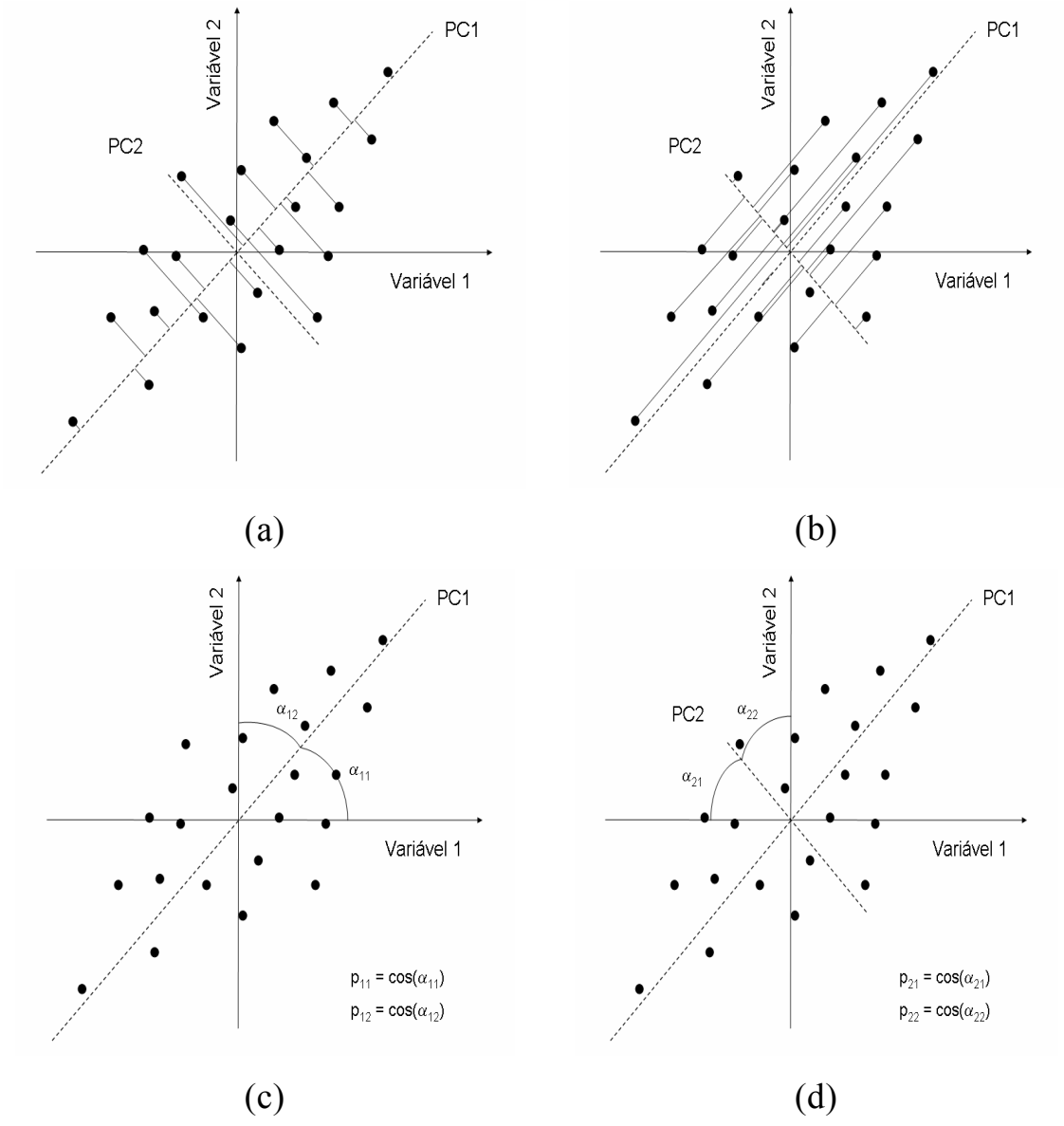


Figura I. 7: PC1 e PC2 para um conjunto de dados em um espaço de duas dimensões, demonstrando a estimativa dos escores (a) e (b) e dos pesos (c) e (d).

[45] *Non linear iterative partial least squares*.

Leverage: detectando outliers

A *leverage* é a medida da distância da amostra ao centro do conjunto dos dados, e é estimada a partir dos escores das amostras como descrito na Equação (6), no qual $\mathbf{T_k}$ são os escores das amostras para as *k* componentes principais. Os valores da *leverage* (*h*) são fornecidos pelos elementos da diagonal da matriz **H** [ASTM 1655, 2000].

$$\mathbf{H} = \mathbf{T_k}\left(\mathbf{T}^{\mathbf{T}}\mathbf{T}\right)^{-1}\mathbf{T_k}^{\mathbf{T}} \qquad (6)$$

A *leverage* indica a influência de cada amostra no modelo, e é uma poderosa ferramenta para a detecção de amostras atípicas para todos os métodos baseados em PCA (PCR, PLS, MPCA, MPLS, PARAFAC, NPLS). Um valor alto de *leverage* indica que a amostra tem grande influência na construção do modelo, e ao ser retirada, causará mudanças significativas. Estas amostras são facilmente visualizadas nos gráficos dos escores. O limite crítico para uma amostra ser considerada anômala é estimado por $3k/I$, no qual *k* é o número de componentes do modelo e *I* é o número de amostras. O limite de controle é estimado como sendo $2k/I$, isto é, amostras acima deste limite apresentam comportamento atípico, mas não devem necessariamente ser retiradas do conjunto.
Existem outras técnicas para a detecção de amostras anômalas, apenas a *leverage* foi descrita por ter sido a única utilizada neste trabalho.

A.2.2 Reconhecimento de Padrões Supervisionado ou Classificação

Nos métodos de reconhecimento de padrões supervisionados, também conhecidos por métodos de classificação, o objetivo é desenvolver um modelo preditivo por meio do uso de amostras alocadas em classes previamente conhecidas, denominadas conjunto de treinamento. Uma vez validado, este modelo poderá ser utilizado para predizer a classe de amostras desconhecidas. Por exemplo, amostras de hidrocarbonetos policíclicos aromáticos (HPA´s) podem ou não apresentar fototoxicidade, que é a propriedade de uma molécula de se tornar tóxica quando exposta à luz, e esta característica depende do intervalo entre a energia do HOMO[46] e a energia do LUMO[47]. Sendo assim, dependendo da estrutura molecular, estas substâncias poderão pertencer à classe "tóxica" ou "não tóxica" [Ribeiro & Ferreira, 2005].

As duas técnicas de reconhecimento de padrões supervisionado mais utilizadas na Química são: a análise do *K*-ésimo vizinho próximo (KNN[48]) e a Modelagem Independente Suave de Analogia de Classes (SIMCA[49]). Estes dois métodos permitem a construção de modelos de classificação baseados no conceito de proximidade entre amostras [Beebe, 1998; Derde & Massart, 1986].

No método KNN, durante a etapa de modelagem a distância euclidiana entre todos os pares de amostras do conjunto de treinamento é estimada para o espaço multidimensional das *J* variáveis, e em seguida cada

[46] *Highest occupied molecular orbital*.
[47] *Lowest unoccupied molecular orbital*.
[48] *K^{th} Nearest Neighbor*.
[49] *Soft Independent Modeling of Class Analogy*.

amostra é alocada em uma das classes previamente definidas por meio de um sistema de votos. Neste sistema, os *K* vizinhos mais próximos de cada amostra são escolhidos para votar e indicar a classe à qual a amostra deve pertencer. A classe mais votada será então atribuída à amostra. Este processo é repetido para diferentes valores de *K*, e o número ótimo será aquele que produzir o menor erro de classificação para as amostras do conjunto de treinamento.

A classificação de novas amostras é realizada também pelo sistema de votos. A distância euclidiana entre a amostra desconhecida e todas as amostras do conjunto de treinamento é estimada, e a atribuição da classe é realizada pela sua proximidade com os *K*-"vizinhos mais próximos". A amostra é então alocada na classe que receber mais votos [Beebe, 1998; Kowalski & Schartzki, 1972; Derde & Massart, 1986].

KNN é uma técnica simples de implementar e funciona bem quando há poucas amostras em cada classe. O conjunto de treinamento pode ser facilmente atualizado pela inclusão de novas amostras. No entanto, não é uma técnica muito eficiente na detecção de amostras anômalas, pois o modelo sermpre irá alocar a amostra em alguma das classes do conjunto de treinamento, mesmo que a amostra não pertença a nenhuma das classes pré-existentes. Outra desvantagem é que não é possível estimar o nível de confiança no resultado da classificação [Beebe, 1998].

Método SIMCA

No método SIMCA, cada classe do conjunto de treinamento é modelada separadamente utilizando análise de componentes principais (PCA), e a partir dos escores e dos resíduos do modelo é possível construir caixas

multidimensionais que determinam com certo grau de confiança os limites de cada classe [Beebe, 1998; Derde & Massart, 1986; Wold & Sjörström, 1977; Wold, 1976].

A modelagem SIMCA neste trabalho foi realizada com o software PLS Toolbox 3.0® [Eigenvector Inc., 2003] para uso no ambiente MatLab [Mathworks Inc., 2002], que utiliza decomposição de valores singulares (SVD[50]) para o cálculo da PCA. Para entender como os limites das caixas dimensionais são estimados esta decomposição será vista com um pouco mais de detalhes. Outros algoritmos podem ser utilizados para a modelagem PCA, como é o caso do software Pirouette, que utiliza o algoritmo NIPALS [Infometrix Inc., 2007].

Decomposição de Valor Singular (SVD) e Análise de Componentes Principais (PCA)

Qualquer matriz **X** (*I*,*J*) pode ser decomposta de acordo com a Equação (7), no qual **U** é ortogonal[51] e possui a dimensão *I* x *I*, **V** é ortogonal e *J* x *J*, e **S** é *I* x *J* e diagonal[52], no qual os elementos da diagonal são os valores singulares (λ), e diminuem gradativamente no sentido da primeira linha até a última linha.

$$\mathbf{X} = \mathbf{USV}^{\mathbf{T}} \qquad (7)$$

[50] *Singular value decomposition*.
[51] Uma matriz **X** é dita ortogonal se $\mathbf{X}^{-1}=\mathbf{X}^{T}$, isto é, se $\mathbf{X}.\mathbf{X}^{T}=\mathbf{I}$. **I**(*I*,*I*) é a matriz identidade, uma matriz quadrada contendo 1 nos elementos $i=j$ e 0 nos elementos $i \neq j$.
[52] A matriz diagonal contém elementos nulos fora da diagonal principal.

Em termos de componentes principais, a matriz de dados **X**(*I*,*J*) é decomposta na matriz **T**(*I*,*K*) dos escores e na matriz **P**(*J*,*K*) dos pesos, gerando a matriz de resíduos **E**(*I*,*J*), em que *k* é o número de componentes principais do modelo (Equação (8)) e deve ser menor ou igual à dimensão de **X**, ou seja, $k \leq \min\{I, J\}$. **T** é a matriz de escores, que contêm a informação das amostras, e **P** é a matriz de pesos, que contêm as informações das variáveis.

$$\mathbf{X} = \mathbf{T_k}\mathbf{P_k^t} + \mathbf{E} \tag{8}$$

Na decomposição PCA, os vetores $\mathbf{p}_i$ da matriz **P** são os autovetores da matriz de covariância (Equação (9)). Para cada vetor $\mathbf{p_i}$, temos um autovalor λ_i associado a ele (Equação (10)), sendo que **T** não é quadrada e é ortogonal ($\mathbf{T}^T\mathbf{T} = 0$ para $i \neq j$) e **P** é ortonormal ($\mathbf{P}^T\mathbf{P} = 0$ para $i \neq j$ e $\mathbf{P}^T\mathbf{P} = 1$ para $i = j$).

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X^T X}}{\mathrm{j}\text{-}1} \tag{9}$$

$$\mathrm{cov}(\mathbf{X})\mathbf{p_i} = \lambda_i \mathbf{p_i} \tag{10}$$

Em termos de SVD, temos as seguintes relações de equivalência:

$$\mathbf{V} = \mathbf{P} \tag{11}$$

$$\mathbf{US} = \mathbf{T} \qquad (12)$$

Os elementos da diagonal da matriz **S** fornecem os valores singulares, cujo quadrado resulta nos autovalores associados a cada componte principal e que representam a variância descrita em cada uma delas. Os autovalores indicam o quanto de informação do conjunto de dados cada componente contém, e seu valor diminui gradativamente da primeira até a última componente, sendo muito próximos de zero nas últimas componentes. A escolha do número adequado de componentes principais (*k*) a ser utilizado na modelagem é realizada com base nos autovalores mais significativos.

Limites das caixas multidimensionais na modelagem SIMCA: T^2 e Q

No PLS Toolbox® [Eigenvector Inc., 2003], o poder discriminante do SIMCA é fornecido por dois parâmetros: o parâmetro T^2 de Hotelling, que estima a distância entre as amostras; e o parâmetro Q, que estima a falta de ajuste do modelo.

O parâmetro Q é estimado pela soma do quadrado dos resíduos de cada linha da matriz dos resíduos (a matriz **E**), isto é, para cada amostra *i* da matriz **X**, Q_i é calculado como descrito na Equação (13), em que $\mathbf{e_i}$ é um vetor linha (Figura I.6). Este parâmetro indica o quanto cada amostra está contemplada pelo modelo PCA. Q é uma medida da diferença, ou seja, do resíduo, entre a amostra e sua projeção no espaço determinado pelas *k* componentes principais do modelo [Wise *et al.*, 2005].

$$Qi = \mathbf{e_i e_i}^{\mathbf{T}} \quad (13)$$

O parâmetro T^2 de Hotelling expressa a variação de cada amostra dentro do modelo PCA. Este parâmetro é estimado a partir da soma normalizada do quadrado dos escores, definido pela Equação (14), no qual $\mathbf{t_i}$ é um vetor linha. Nesta equação, $\mathbf{t_i}$ são os escores da *i*-ésima amostra, $\boldsymbol{\lambda}^{-1}$ é a matriz diagonal contendo o inverso dos autovalores associados com as *k* componentes principais do modelo [Wise et al., 2005].

$$T_i^2 = \mathbf{t_i \lambda^{-1} t_i}^{\mathbf{T}} \quad (14)$$

Amostras pertencentes a uma determinada classe deverão apresentar simultaneamente valores baixos de Q e T^2 para a modelagem PCA característica da classe em questão, e os limites de cada classe podem ser obtidos pela estimativa dos intervalos de confiança para Q e T^2.

Os limites para Q são estimados considerando-se os autovalores λ, como descrito na Equação (15), em que Θ fornece o somatório dos autovalores não utilizados no modelo (Equação (16)), $c\alpha$ é o desvio da distribuição normal de Q correspondente ao percentil superior (ou seja, 1 - α, sendo α o grau de confiança desejado), e h_0 é descrito na Equação (17). Na Equação (16), *p* é a potência utilizada para a aproximação da distribuição Gaussiana dos resíduos, *j* é a variável, *k* é o número de componentes principais utilizadas no modelo e $n \leq \min \{I, J\}$ [Jensen & Solomon, 1972; Jackson & Mudholkar, 1979; *Wise et al.*, 2005].

$$Q\alpha = \Theta_1 \left[\frac{c_\alpha \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{\frac{1}{h_0}} \quad (15)$$

$$\Theta_p = \sum_{j=k+1}^{n} \lambda_j^p \text{ , para } p = 1, 2, 3 \quad (16)$$

$$h_0 = 1 - \frac{2\Theta_1 \Theta_3}{3\Theta_2^2} \quad (17)$$

Os limites de confiança para os valores de T^2 são estimados utilizando os limites da distribuição F (Equação (18)), em que I_q é o número de amostras da classe q, k é o número de componentes principais utilizadas para desenvolver o modelo PCA. $F(k,I_q\text{-}k,\alpha)$ é o fator crítico da distribuição F com k graus de liberdade no denominador, I_q-k graus de liberdade no denominador, e α é o grau de confiança desejado [Wise *et al.*, 2005; NIST, 2006].

$$T^2_{k,I,\alpha} = \frac{k(I_q - 1)}{I_q - k} F_{k,I_q\text{-}k,\alpha} \quad (18)$$

De uma forma geral, Q é uma estimativa da informação dos dados que foi deixada de fora do modelo, e T^2 é uma medida da distância que a amostra está do centro do conjunto de dados. Sendo assim, o método

SIMCA utiliza análise de componentes principais (PCA) para modelar a distribuição do agrupamento das amostras de cada classe do conjunto de treinamento, formando uma caixa multidimensional com limites bem definidos. A classificação das novas amostras é então realizada pela sua projeção nas caixas adequadas [Beebe, 1998; Wold, 1976; Derde & Massart, 1986].

A Figura I.8 contém a representação de um exemplo tridimensional da modelagem SIMCA, no qual é possível notarmos a existência de quatro classes de amostras. Cinco novas amostras (em preto) foram classificadas com base neste modelo, das quais quatro foram identificadas como pertencente a alguma das classses pré-existentes e uma como não pertencente a nenhuma das classes, apesar ter ficado bem proxima a um dos agrupamentos.

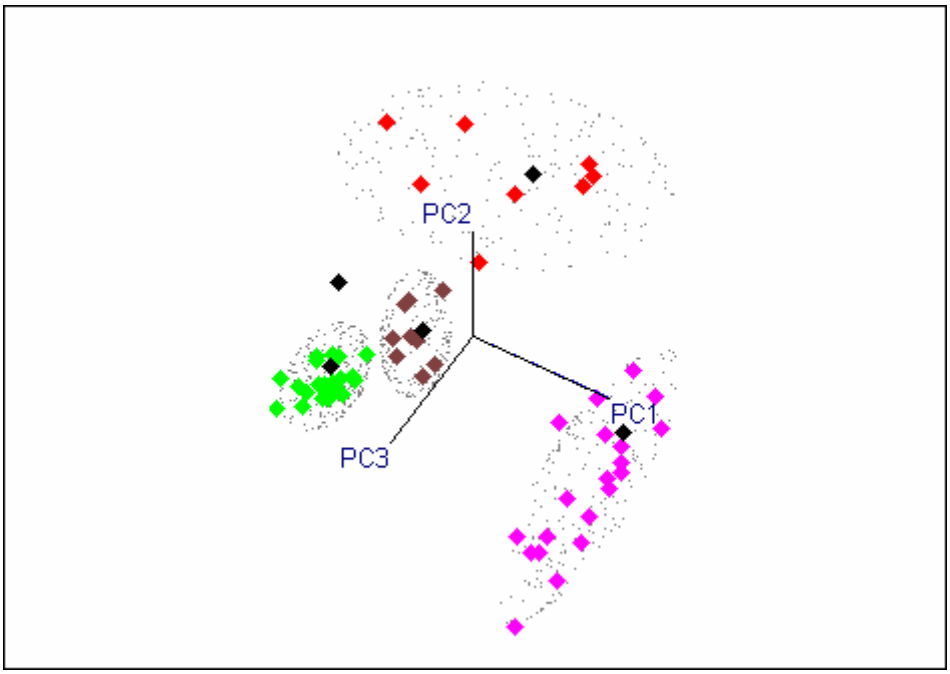


Figura I. 8: Exemplo tridimensional de SIMCA mostrando os limites de cada classe de amostras, e a projeção das amostras desconhecidas (em preto).

O método SIMCA fornece também uma medida para a relevância de cada variável para a construção do modelo, por meio da estimativa do poder de modelagem, o que permite a inclusão das variáveis importantes para a análise, e a exclusão das variáveis irrelevantes [Wold & Sjöström, 1977]. O poder de modelagem fornece a habilidade de cada variável de modelar a informação contida nos dados e sua estimativa é realizada pela comparação do desvio padrão residual de cada variável na classe em estudo ($s_{j,res}$, Equação (19)), com o desvio padrão residual da variável *j* ($s_{j,X}$, Equação (20)) para todo o conjunto de amostras, como descrito na Equação (21). Nas Equações (19) e (20), os índices *q* e p representam respectivamente classe e a amostra, I_q = número de amostras da classe *q*, *k* = número de componentes principais da classe *q*, res_{pj}^{q} são os resíduos de cada variável *j* na classe *q*, x_{pj}^{q} são os valores da variável *j* da classe *q* e $\bar{x}$ é a média dos valores da variável *j* para todo o conjunto de dados. É possível observar que o poder de modelagem (MPj) de cada variável descreve a razão sinal/ruído da variável *j* e varia de zero a 1. Quanto mais próximo de 1, maior será a habilidade da variável de modelar as informações contidas no conjunto dos dados [Wold & Sjöström, 1977].

$$s_{j,res} = \sqrt{\frac{\sum_{p=1}^{Iq} \left(res_{pj}^{q}\right)^2}{Iq - Aq - 1}} \qquad (19)$$

$$s_{j,X} = \sqrt{\frac{\sum_{p=1}^{Iq}\left(x_{pj}^{q} - \bar{x}_j\right)^2}{Iq-1}} \tag{20}$$

$$MP_j = 1 - \frac{s_{j,res}}{s_{j,X}} \tag{21}$$

Outra ferramenta importante da modelagem SIMCA é o poder discriminante, que testa a habilidade das diversas variáveis do modelo de distinguir as classes presentes do conjunto de treinamento.

Para estimar o poder discriminante da variável *j*, cada amostra *p* pertencente à classe *r*, é projetada na classe *q*, sendo que res_{pj}^{q} é o resíduo desta projeção. A Equação (22) fornece o desvio padrão destes resíduos, no qual *k* = número de componentes principais do modelo, *J* é o número de variáveis e I_r o número de amostras na classe *r*. O poder discriminante é dado pela Equação (23), no qual quanto mais próximo de zero, menor o poder discriminante da variável.

$$s_{j,r}^{q} = \sqrt{\frac{\sum_{p=1}^{Ir}\left(res_{pj}^{q}\right)^2}{I_r\left(J-k\right)}} \tag{22}$$

$$\phi_j^{r,q} = \sqrt{\frac{\left(s_{j,r}^{r}\right)^2 - \left(s_{j,q}^{r}\right)^2}{\left(s_{j,r}^{r}\right)^2 - \left(s_{j,q}^{q}\right)^2}} \tag{23}$$

A.2.3 Calibração Multivariada

Calibração é o processo de construir um modelo matemático para um conjunto de amostras, que descreva a relação entre as variáveis independentes e uma ou mais variáveis dependentes [Beebe, 1998; Geladi *et al.*, 1992b]. Em química analítica a calibração é utilizada para relacionar a resposta instrumental (variável indenpendente) com a concentração de uma ou mais espécies de interesse (variável dependente) [Pimentel & Barros Neto, 1996]. Em estudos de QSAR, a calibração é utilizada para relacionar um conjunto de descritores moleculares e propriedades físico-químicas às atividades farmacológica ou tóxica das moleculas [Ferreira, 2002; Ribeiro & Ferreira, 2005]. Este modelo matemático é construído para um conjunto de amostras cujos valores das variáveis dependentes são conhecidos. Uma vez validado, o modelo poderá ser utilizado para predizer os valores das variáveis dependentes para novas amostras.

Na calibração univariada, a variável dependente é modelada em função de apenas uma variável independente. O exemplo clássico disto na Química Analítica, é a modelagem do sinal analítico obtido por técnicas instrumentais tais como cromatografia (área cromatográfica) ou espectroscopia (absorção em determinado comprimento de onda) em função de uma espécie de interesse (Figura I.9). No entanto, para que o método seja confiável, é necessário garantir que o sinal analítico seja característico somente da espécie de interesse, ou seja, que ele seja seletivo, o que requer uma etapa prévia de preparo da amostra, o que aumenta os custos e o tempo da análise.

Para contornar o problema da seletividade e permitir a análise na presença de interferentes sem comprometer a eficiência do método, uma alternativa viável e econômica é o uso de métodos multivariados, que correlacionam a concentração do analito com toda a faixa espectral, possibilitando predições estatisticamente confiáveis mesmo na presença de interferentes, desde que estes sejam conhecidos (Figura I.10).

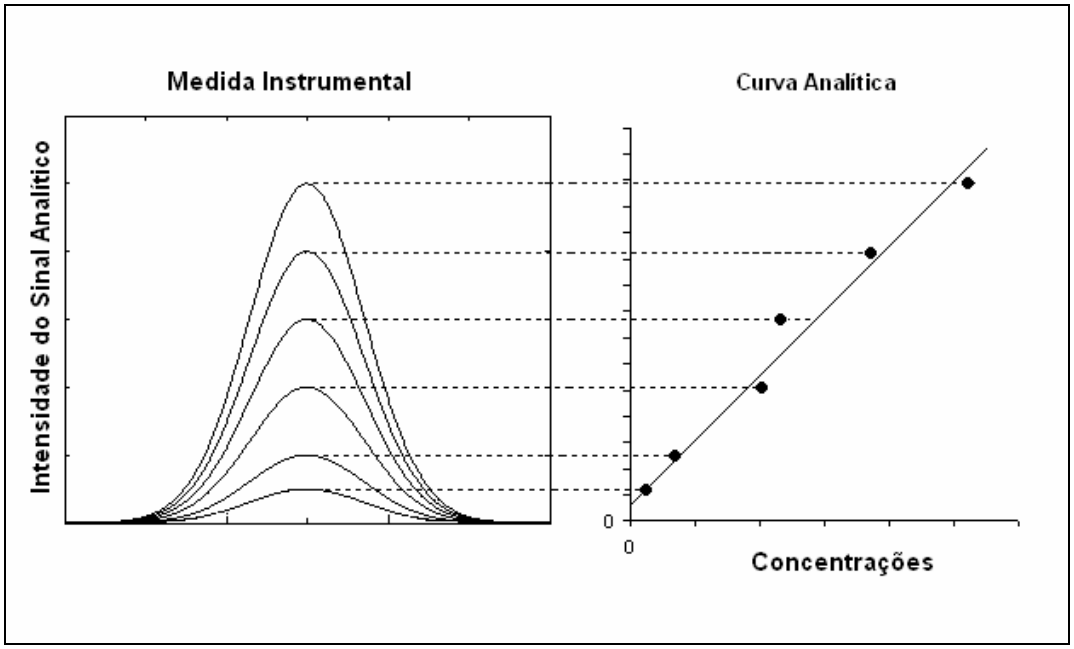


Figura I. 9: Modelagem univariada do sinal analítico em função da concentração.

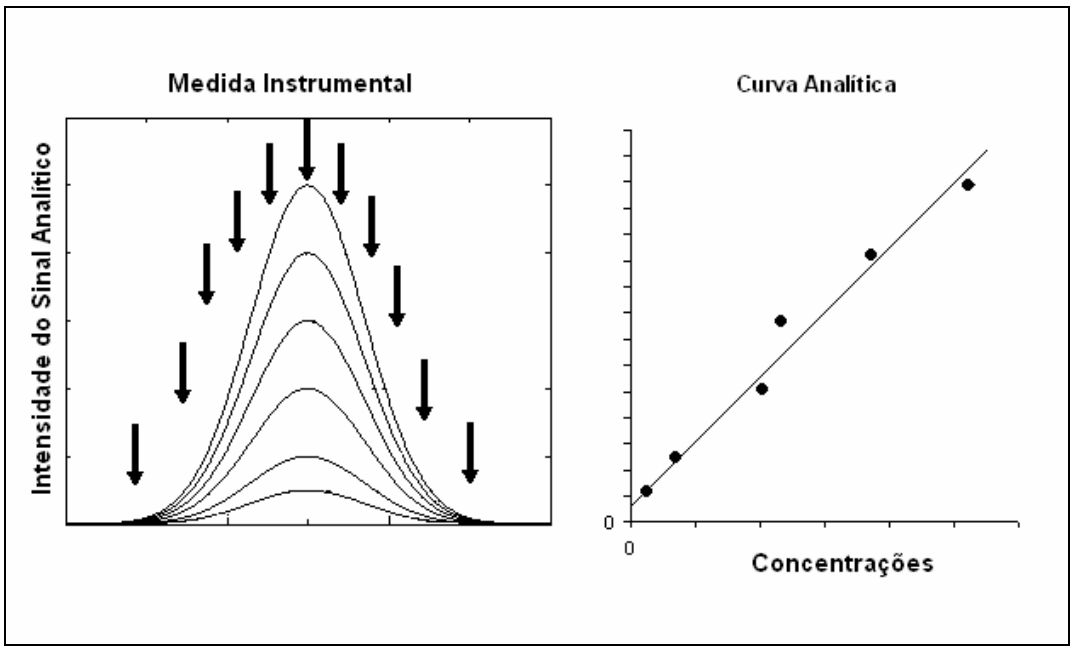


Figura I. 10: Modelagem multivariada do sinal analítico em função da concentração.

Os métodos de calibração multivariada mais difundidos atualmente são: a regressão linear múltipla (MLR[53]), a regressão por componentes principais (PCR[54]) e a regressão por mínimos quadrados parciais (PLS[55]) [Geladi & Kowalski, 1986; Geladi, 1988; Martens & Naes, 1993; Ferreira et al, 1999]. De uma forma geral, a regressão linear múltipla é adequada quando o número de variáveis é pequeno e menor do que o número de amostras. Este tipo de calibração requer uma etapa prévia de seleção de variáveis, e é útil na determinação de diversas espécies simultaneamente quando é possível obter-se um sinal analítico seletivo para cada uma delas. No entanto, na maioria das vezes isto não é possível, e uma etapa prévia de redução da dimensionalidade dos dados faz-se necessária. Nestes casos, métodos de calibração multivariada tais

[53] *Multiple linear regression*.
[54] *Principal component regression*.
[55] *Partial least square regression*.

como PCR e PLS são mais adequados. De uma forma geral, estes métodos utilizam a análise de componentes principais como estratégia para reduzir a dimensão dos dados, e os escores resultantes são modelados em função da concentração [Geladi & Kowalski, 1986; Geladi, 1988; Geladi *et al.*, 1992b; Martens & Naes, 1993; Beebe, 1998; Ferreira *et al.*, 1999].

A.3 Métodos de Ordem Superior, Métodos Multi-way ou Métodos Multi-Modos

Os métodos *Multi-way* são uma extensão natural dos métodos de primeira ordem para arranjos tri- ou *N*-dimensionais do tipo $\underline{\mathbf{X}}$ (*I*, *J*, *K*, ...). A maioria das aplicações em química envolve o uso destes métodos em arranjos tridimensionais, então a sua descrição será generalizada para este tipo de arranjo.

A.3.1 Unfold-PCA & Unfold-PLS

Uma das maneiras de analisar dados *N-way* é realizar um desdobramento para transformá-los em uma matriz, e na sequência submetê-los ao tratamento com os métodos multivariados de primeira ordem, em geral, PCA e PLS [Smilde *et al.*, 2004; Bro, 1998]. Este desdobramento é a concatenação das matrizes que formam o arranjo *N-way*, e pode ser realizado ao longo de qualquer um dos modos do arranjo (Figura I.11).

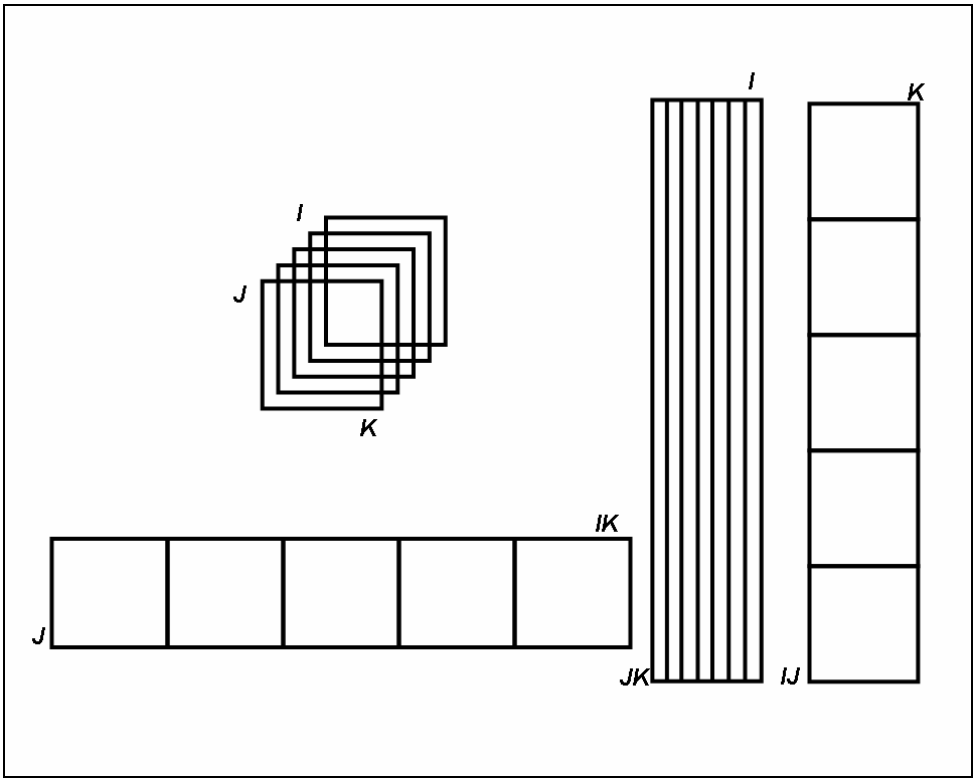


Figura I. 11: Ilustração do desdobramento de um arranjo 3-way no sentido dos três modos.

A aplicação de PCA após o desdobramento dos dados é usualmente conhecida como MPCA (de *Multilinear* PCA) ou *unfold*-PCA (do inglês, *unfolding*, que significa desdobramento). Da mesma forma, a utilização de PLS após o desdobramento recebe o nome de *unfold*-PLS.

No entanto, ao realizar o desdobramento a estrutura trilinear dos dados é ignorada, e a análise é realizada pelos métodos multivariados usuais, o que eventualmente poderá levar a modelos menos robustos e mais difíceis de interpretar [Smilde *et al.*, 2004; Bro, 1998; Huang *et al.*, 2003].

No caso dos dados de imagens descritos nesta tese, o arranjo $\underline{\mathbf{X}}$ (I, J, K) formado pelas imagens empilhadas e submetidas à transformação 2D-

FFT[56] (power spectrum) é descrito como um arranjo do tipo OVV (objeto x variável x variável). Ao desdobrar este arranjo em uma matriz **X** (*I*, *JK*), a direção das amostras é mantida intacta. Da mesma forma, após a modelagem com PCA, em cada componente os dados são decompostos em um vetor de escores (*I*, 1) no modo das amostras, e um vetor longo de pesos (*JK*), que pode ser subseqüentemente redobrado em uma matriz de pesos **P** (*J*, *K*), que descreve a variação no espaço 2-D do domínio das variáveis para cada componente principal. A estrutura do modelo com os resíduos é fornecida pela Equação (24), no qual o índice *f* é o número de componentes, e **E** é a imagem residual do *unfold*-PCA.

$$\underline{\mathrm{X}} = \sum_{f=1}^{F} t_f \underline{\mathrm{P}}_f + \underline{\mathrm{E}} \qquad (24)$$

A.3.2 PARAFAC e os modelos de Tucker

De uma forma geral, o modelo PARAFAC decompõe o arranjo $\underline{\mathbf{X}}$ (*I*, *J*, *K*) em três matrizes de pesos denominadas **A**, **B** e **C** (Figura I.12). Estas matrizes contêm respectivamente os elementos a_{if}, b_{if}, c_{kf}, referentes a cada dimensão de $\underline{\mathbf{X}}$: a dimensão das amostras, das variáveis do modo *J* e das variáveis do modo *K*. O modelo trilinear é ajustado para minimizar a soma quadrática dos resíduos, e_{ijk}, como descrito na Equação (25), em que *f* é o número de fatores do modelo. O modelo PARAFAC pode também ser representado pela Equação matricial (26),

[56] Cf. seção 3.2.

na qual as matrizes **A**, **B** e **C** possuem respectivamente as dimensões *I* x *F*, *J* x *F* e *K* x *F*, e o símbolo $|\otimes|$ representa o produto de Khatri-Rao [Smilde *et al.*, 2004; Bro, 1998; Bro, 1997].

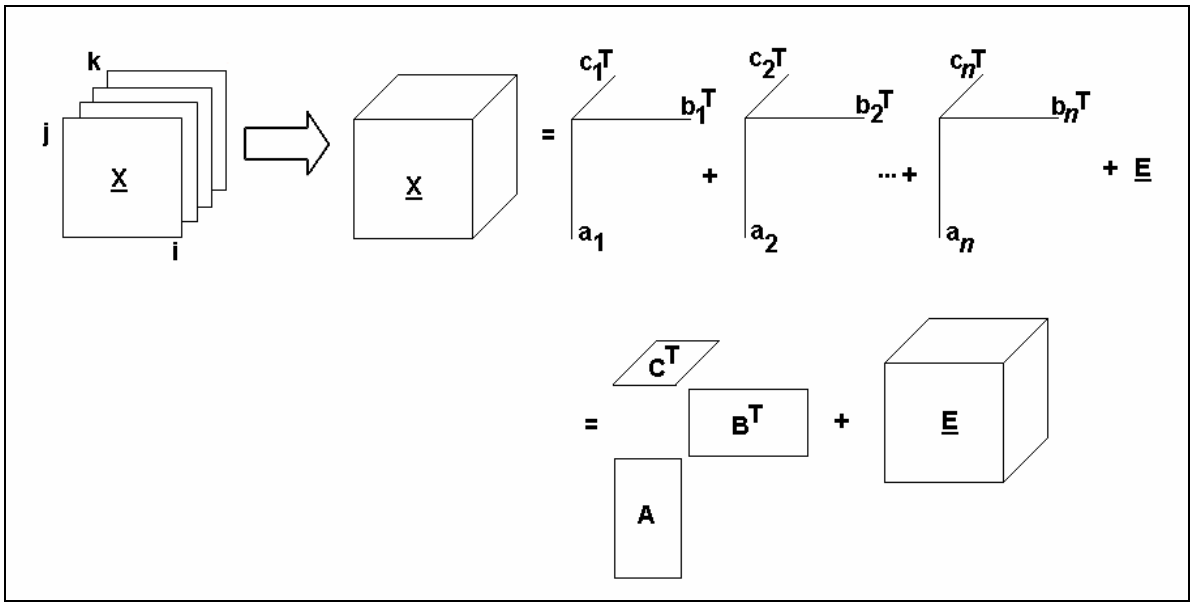


Figura I. 12: Representação gráfica do modelo PARAFAC.

$$x_{ijk} = \sum_{f=1}^{\mathbf{F}} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{25}$$

$$\underline{\mathbf{X}} = \mathbf{A}(\mathbf{C}|\otimes|\mathbf{B})^{\mathbf{T}} + \underline{\mathbf{E}} \tag{26}$$

A estimativa do modelo PARAFAC é realizada utilizando o algoritmo dos quadrados alternados (ALS[57]), que estima sucessivamente os valores de cada matriz de pesos a partir dos valores conhecidos das outras duas matrizes. A inicialização deste algoritmo é realizada por meio de algum

[57] *Alternating Least Squares*.

método de decomposição, usualmente DTLD (*direct trilinear decomposition*) ou SVD (*single value decomposition*). O algoritmo converge iterativamente até que um determinado critério de convergência, ou um número previamente definido de iterações, seja atingido [Smilde *et al.*, 2004; Bro, 1997; Bro, 1998].

Os modelos de Tucker constituem uma família de modelos, denominados Tucker1, Tucker2 e Tucker3. Em linhas gerais, o modelo Tucker1 corresponde ao desdobramento do arranjo $\underline{\mathbf{X}}$ em uma matriz bidimensional, seguido de uma decomposição bilinear utilizando PCA. O modelo Tucker3 é expresso pela Equação (27), em que as matrizes **A** (*I*, *D*), **B** (*J*, *E*) e **C** (*K*, *F*) contêm os pesos do modelo referente às três dimensões dos dados, o arranjo $\underline{\mathbf{G}}$ (*D*, *J*, *K*) é o *core*, e o tensor $\underline{\mathbf{E}}$ (*I*, *J*, *K*) contém os resíduos do modelo. O símbolo $\otimes$ é o produto de Kronecker, ou produto tensorial. Os índices *D*, *E* e *F* representam o número de componentes de cada modo, e uma característica importante deste modelo, quando comparado ao PARAFAC, é que o modelo Tucker3 permite que o número de fatores decompostos seja diferente em cada dimensão. O modelo Tucker2 é um caso particular do Tucker3, em que uma das dimensões é mantida fixa durante a decomposição.

$$\underline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{G}}(\mathbf{C} \otimes \mathbf{B})^{\mathbf{T}} + \underline{\mathbf{E}} \qquad (27)$$

Ambos, PARAFAC e Tucker3 são mais simples do que a aproximação alternativa de desdobrar o arranjo e aplicar os métodos de primeira ordem (*unfold*-PCA e *unfold*-PLS), e a principal vantagem da sua utilização é que neste caso, a estrutura trilinear dos dados é preservada.

No entanto, estes modelos são mais rígidos, pois qualquer desvio da trilinearidade será suficiente para invalidar a análise.

A escolha do número de fatores: Teste de Consistência Trilinear (Corcondia[58])

Existem diversas ferramentas para estimar o número adequado de fatores para um modelo e as mais simples são o conhecimento químico do sistema em estudo e a porcentagem de variância descrita. Outras estratégias mais refinadas podem ser utilizadas, tais como a validação cruzada, e no caso do PARAFAC, o uso do teste de consistência trilinear, denominado CORCONDIA [Bro & Kiers, 2003; Bro, 1998]. Este teste é baseado na interpretação do modelo PARAFAC como um modelo Tucker3 restrito. Neste teste, o arranjo $\underline{\mathbf{X}}$ é modelado com Tucker3, e se houver consistência trilinear os elementos da superdiagonal do tensor central ($\underline{\mathbf{G}}$, confira Equação (27)) terão valores próximos a 1, e os demais elementos valores próximos a zero. O valor de CORCONDIA será então fornecido em termos de porcentagem pela Equação (28), no qual gedf é o elemento da matriz central estimado com Tucker3 a partir dos pesos do PARAFAC, hdef é o elemento de um tensor binário contendo valores iguais a 1 na superdiagonal e zero nas demais posições, e **F** é o número de fatores do modelo [Bro & Kiers, 2003; Bro, 1997; Reis, 2002; Sena, 2004].

[58] *Core Consistency Diagnostic*.

$$\text{CORCONDIA} = 100 \times \left(1 - \frac{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}\left(g_{def} - h_{def}\right)^2}{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F} h_{def}^2} \right) \tag{28}$$

Modelos válidos apresentam valores de CORCONDIA entre 80 e 100%. Modelos com valores em torno de 50% são instáveis, e o uso de restrições no modelo poderá ajudar a estabilizá-lo. Valores próximos de zero ou negativos indicam que os dados não podem ser descritos por um modelo trilinear, ou que o número de fatores utilizado é maior do que o necessário. Os valores de consistência trilinear devem diminuir com o aumento do número de componentes, portanto, um número apropriado de fatores para o modelo deverá apresentar o maior número de componentes em que seja possível obter um valor válido de CORCONDIA [Bro, 1998; Bro, 1997].

Além dos valores de consistância trilinear, um modelo com número apropriado de fatores deverá apresentar valores baixos para os resíduos obtidos (**E**) e o critério de convergência deve ser obtido com poucas iterações.

A.3.3 NPLS: Calibração Multi-way

O método NPLS é uma extensão do método PLS para dados de ordem superior. O NPLS é um método de regressão que relaciona o arranjo trilinear de variáveis independentes **X** (*I*, *J*, *K*) com um vetor, ou matriz de variáveis dependentes **y** (ou **Y**). O NPLS decompõe o arranjo trilinear **X** de uma maneira semelhante ao modelo PARAFAC, mas

neste caso, o modelo deve descrever a covariância máxima entre as variáveis dependentes e independentes. Primeiramente, o NPLS decompõe o arranjo $\underline{\mathbf{X}}$ em uma matriz $\mathbf{T}(I,F)$, relacionada ao modo das amostras, e duas matrizes $\mathbf{W^J}(J,F)$ e $\mathbf{W^K}(K,F)$, relacionadas às outras duas dimensões. O modelo estrutural é descrito pela Equação (29), em que o símbolo $|\otimes|$ representa o produto de Khatri-Rao, e $\underline{\mathbf{E}}_\mathbf{x}$ são os resíduos de $\underline{\mathbf{X}}$ [Bro, 1996; Smilde, 1997; Bro, 1998].

$$\underline{\mathbf{X}} = \mathbf{T}(\mathbf{W^K} \left|\otimes\right| \mathbf{W^J})^\mathbf{T} + \underline{\mathbf{E}}_\mathbf{x} \tag{29}$$

De uma forma geral, o vetor **t** é equivalente aos escores do modelo PCA, e os vetores $\mathbf{w^J}$ e $\mathbf{w^k}$ equivalem aos pesos para os dois outros modos. A matriz **Y** (ou vetor **y**) é então relacionada com a matriz **T** (I,F) por meio da Equação (30), em que **Q** é a matriz dos pesos das variáveis dependentes **Y**, e esta matriz é determinada após a estimativa de cada fator do modelo (Equação (31)).

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^\mathbf{T} \tag{30}$$

$$\mathbf{Q} = (\mathbf{T}^\mathbf{T}\mathbf{T})^{-\mathbf{1}}\mathbf{T}^\mathbf{T}\mathbf{Y} \tag{31}$$

A estimativa de $\hat{\mathbf{Y}}$ para novas amostras é realizada como descrito na Equação (32), no qual $\mathbf{T}^*$ são os valores da matriz **T** para as novas amostras [Bro, 1996; Smilde, 1997; Bro, 1998]. A Figura I.13 contém a representação gráfica do modelo NPLS.

$$\hat{\mathbf{Y}} = \mathbf{T}^{*}\mathbf{Q}^{\mathbf{T}} \tag{32}$$

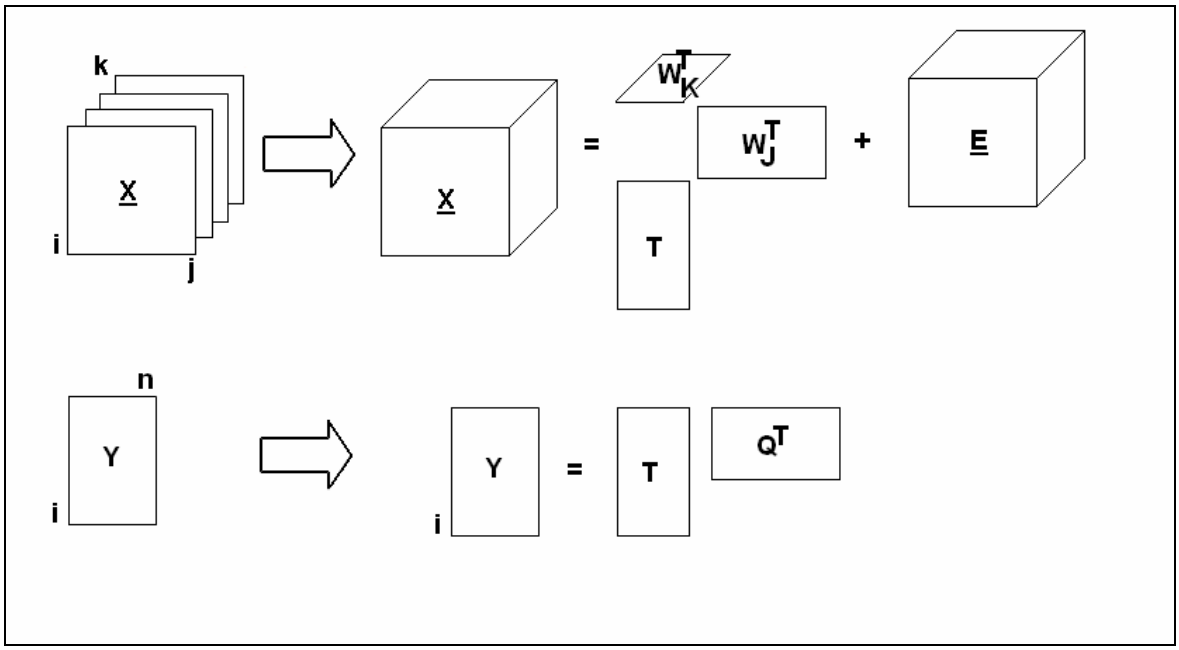


Figura I. 13: Representação gráfica do modelo NPLS.

As vantagens do NPLS em relação ao *unfold*-PLS é que o primeiro é mais simples e fácil de interpretar, menos sensível à presença de ruído, pois a decomposição é realizada através dos três modos simultaneamente [Bro, 1996; Smilde, 1997; Huang *et al.*, 2003].

Notação para modelos NPLS

Aos modelos PLS multilineares é atribuída a nomenclatura geral de "modelos NPLS". O uso de um prefixo indica a ordem do bloco **X** de variáveis independentes, enquanto o uso de um número após a sigla PLS indica a ordem do bloco **Y** de variáveis dependentes. A Tabela 15 ilustra a terminologia geral para os modelos PLS [Bro, 1996].

Tabela 15: Abreviações para diferentes modelos PLS dependendo das ordens dos blocos **X** e **Y** [Bro, 1996].

	Ordem de **X**		
Ordem de **Y**	2	3	4
1	Bi-PLS1	Tri-PLS1	Quadri-PLS1
2	Bi-PLS2	Tri-PLS2	Quadri-PLS2

O uso de uma variável dependente apenas (um vetor **y**) fornecerá um modelo PLS1, enquanto o uso de uma matriz de variáveis dependentes (**Y**) fornecerá um modelo PLS2. No modelo PLS2, todas as variáveis do bloco **Y** são modeladas simultaneamente [Bro, 1996; Smilde, 1997; Bro, 1998].

Para um arranjo **X** bidimensional, o modelo é denominado bi-PLS, enquanto para um arranjo $\underline{\mathbf{X}}$ tridimensional, o modelo é denominado tri-PLS etc. [Bro, 1996; Smilde, 1997; Bro, 1998].

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAMOWITZ, M., "*Microscope Basics and Beyond*", (Vol. 1), Olympus America Inc., Melville, New York, 2003.

ADAMSON, G. W. & BAWDEN, D., "Comparison of hierarchical cluster-analysis techniques for automatic classification of chemical structures", *J Chem Inf Comp Sci*, 21, 204-209, 1981.

ALESSANDRINI, A. & FACCI, P., "AFM: a versatile tool in biophysics", *Meas Sci Technol*, 16, R65-R92, 2005.

ANDERSSON C. A. & BRO R., "The N-way Toolbox for MATLAB", *Chemom Intell Lab Syst*, 52, 1-4, 2000, (http://www.models.kvl.dk/source/nwaytoolbox, consultado em 10/07/2007).

ANTONELLI, A.; COCCHI, M.; FAVA, P.; FOCA, G.; FRANCHINI, G. C.; MANZINI, D.; ULRICI, A., "Automated evaluation of food color by means of multivariate image analysis coupled to a wavelet-based classification algorithm", *Anal Chim Acta*, 515, 3-13, 2004.

ARTYUSHKOVA, K. & FULGHUM, J. E., "Multivariate image analysis methods applied to XPS imaging data sets", *Surf Interface Anal*, 33, 185-195, 2002.

ASTM 1655, Annual Book of ASTM Standards, E1655-00: Standard practise for infrared multivariate quantitative analysis, American Society for testing and materials international, PA, 2000.

BEEBE, K., "*Chemometrics: a practical guide*", John Wiley, 1998.

BHARATI, M. H.; LIU, J. J.; MACGREGOR, J. F., "Image texture analysis: methods and comparisons", *Chemom Intell Lab Syst*, 72, 57-71, 2004.

BONNET, N., "Some trends in microscope image processing", *Micron*, 35, 635-653, 2004.

BORIN, A.; FERRAO, M. F. MELLO, C.; CORDI, L.; PATACA L. C. M.; DURAN, N.; POPPI, R. J.; "Quantification of Lactobacillus in fermented milk by multivariate image analysis with least-squares support-vector machines", *Anal Bio Chem*, 387, 1105-1112, 2007.

BOZZOLA, J.J. & RUSSEL, L.D., "*Electron microscopy*", 2nd ed., Jones and Bartlett Publishers, Boston, 1999.

BRO, R., "*Multi-way Analysis in the Food Industry*", PhD Thesys, Denmark, 1998 (http://www.models.kvl.dk/users/rasmus/brothesis.pdf, consultado em 10/07/2007).

BRO, R., "PARAFAC. Tutorial and Applications", *Chemom Intell Lab Syst*, 38, 149-171, 1997.

BRO, R., "Multi-way calibration. Multi-linear PLS", *J Chemom*, 10, 47-62, 1996.

BRO, R. & KIERS, H. A. L., "A new efficient method for determining the number of components in PARAFAC models", *J Chemom*, 17, 274-286, 2003.

BRO, R. & SMILDE, A., "Centering and scaling in component analysis", *J Chemom*, 17, 16-33, 2003.

BUCHER, T.; DUNNWALD, M.; LINKE, I. M.; RABES H. M., "The profenitor concept in onogenesis – dendogram of 19 tissues derived from PGK-1B/-1A mosaic compositions in female mice", *Genet Res*, 45, 220-220 1985.

BURGER, J. & GELADI, P. "Hyperspectral NIR image regression part II: Dataset preprocessing diagnostics", *J Chemom*, 20, 106-119, 2006.

CHEMFINDER, www.chemfinder.com, acesso em 20/01/2007.

CHEVALLIER, S.; BERTRAND, D.; KOHLER; A.; COURCOUX, P; "Application of PLS-DA in multivariate image analysis", *J Chemom*, 20, 221-229, 2006.

COLOMBERA, K. M., "Efeitos de condicionadores comerciais nas propriedades mecânicas e nos processos de difusão de fibras capilares", Dissertação de Mestrado, Instituto de Química, UNICAMP, 2004.

COOKE, P.M., "Chemical Microscopy", *Anal Chem*, 72, R169-R188, 2000.

COOKE, P.M., "Chemical Microscopy", *Anal Chem*, 70, R385-R423, 1998.

COOKE, P.M., "Chemical Microscopy", *Anal Chem*, 68, R333-R378, 1996.

COOKE, P.M., "Chemical Microscopy", *Anal Chem*, 66, R558-R594, 1994.

COOKE, P.M., "Chemical Microscopy", *Anal Chem,* 64, R219-R243, 1992.

COOKE, P.M., "Chemical Microscopy", *Anal Chem*, 62, R423-R441, 1990.

COOKE, P.M., "Chemical Microscopy", *Anal Chem*, 60, R212-R226, 1988.

COOKE, P.M., "Chemical Microscopy", *Anal Chem*, 58, R1926-R1937, 1986.

COURCOUX, P.; DEVAUX, M.; BOUCHET, B., "Simultaneous decomposition of multivariate images using three-way data analysis. Application to the comparison of cereal grains by confocal laser scanning microscopy", *Chemom Intell Lab Syst*, 62, 103-113, 2002.

DE JUAN, A.; TAULER, R.; DYSON, R.; MARCOLLI, C.; RAULT, M.; MAEDER, M., "Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis", *Trends Anal Chem*, 23, 70-79, 2004.

DERDE, M. P. & MASSART, D. L., "Supervised pattern recognition: the ideal method?", *Anal Chim Acta*, 191, 1-16, 1986.

EIGENVECTOR RESEARCH, INC., "PLS Toolbox 3.0, for use with MATLAB™", USA, 2003.

ESBENSEN, K. & GELADI, P., "Strategy of multivariate image analisys (MIA)", *Chemom Intell Lab Syst*, 7, 67-86, 1989.

FERREIRA, M. M. C., "Multivariate QSAR", *J Braz Chem Soc*, 13, 742-753, 2002.

FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S., VOLPE, P. L. O., "Quimiometria I: Calibração Multivariada, um Tutorial", *Quím Nova*, 22, 724-731, 1999b.

FEUGHELMAN, M. & WILLIS, B. K., "Mechanical extension of human hair and the movement of the cuticle", *J Cosmet Sci*, 52, 185-193, 2001.

GELADI, P, "Notes of the History and Nature of Partial Least Squares (PLS) Modelling", *J Chemom*, 231-246, 1988.

GELADI, P., "Some special topics in multivariate image analysis", *Chemom Intell Lab Syst*, 14, 375-390, 1992.

GELADI, P.; BERGNER, H.; RINGQVIST, L., "From experimental design to images to particle size histograms to multiway analysys. An example of peat dewaring", *J Chemom*, 14, 197-211, 2000.

GELADI, P.& KOWALSKI, B. R., "Partial Least Square Regression: a Tutorial", *Anal Chim Acta*, 185, 1-17, 1986.

GELADI, P. & GRAHN, H., "*Multivariate image analysis*", John Wiley & Sons, England, 1996.

GELADI, P.; BENGTSSON, E.; ESBENSEN, K.; GRAHN, H., "Image analysis in chemistry I. Properties of images, greylevel operations, the multivariada image", *Trends Anal Chem*, 11, 41-53, 1992a.

GELADI, P.; GRAHN, H.; ESBENSEN, K.; BENGTSSON, E., "Image analysis in chemistry II. Multivaritate image analysis", *Trends Anal Chem*, 11, 121-131, 1992b.

GELADI, ISAKSSON H, LINDQSVIST, L, WOLD, S, ESBENSEN, K, "Principal Component Analysis of Multivariate Images", *Chemom Intell Lab Syst*, 5, 209-220, 1989.

GURDEN S. P.; LAGE E. M.; DE FARIA C. G.; JOEKES I.; FERREIRA M. M. C.; "Analysis of video images from a gas-liquid transfer experiment: a comparison of PCA and PARAFAC for multivariate image analysis", *J Chemom*, 17, 400-412, 2003.

GURDEN S. P.; MONTEIRO V. F.; LONGO E.; FERREIRA, M. M. C., "Quantitative analysis and classification of AFM images of human hair", *J Microsc-Oxford*, 215, 13-23, 2004.

GURDEN, S. P.; WESTERHUIS, J. A.; BRO, R.; SMILDE, A. K, "A comparison of multiway regression and scaling methods", *Chemom Intell Lab Syst*, 59, 121-136, 2001.

HADJUR, C.; DATY, G.; MADRY, G.; CORCUFF, P., "Cosmetic assessment of the human hair by confocal microscopy", *Scanning*, 24, 59-64, 2002.

HAMERS, R. J., "Scanned probe microscopies in chemistry", *J Phys Chem*, 100, 13103-13120, 1996.

HENRION, R.; HENRION, G.; ONUOHA, G., "Multi-way principal components analysis of a complex data array resulting from physicochemical characterization of natural waters", *Chemom Intell Lab Syst*,16, 87-94, 1992.

HENRION, R., "Tutorial: N-way principal component analysis – theory, algorithms ans applications", *Chemom Intell Lab Syst*, 25, 1-23, 1994.

HUANG, J.; WIUM, H.; QVIST, K. B.; ESBENSEN, K. H., "Multi-way methods in image analysis – relationships and applications", *Chemom Intell Lab Syst*, 66, 141-158, 2003.

HUANG J. & ESBENSEN K. H., "Applications of Angle Measure Technique (AMT) in image analysis Part I. A new methodology for in situ powder characterization" *Chemom Intell Lab Sys*, 54, 1-19, 2000.

HUANG J. & ESBENSEN K. H., "Applications of AMT (Angle Measure Technique) in image analysis Part II: Prediction of powder functional properties and mixing components using multivariate AMT regression (MAR)", *Chemom Intell Lab Syst*, 57, 37-56, 2001.

INFOMETRIX INC., Pirouette®, Multivariate Data Analysis, Version 4.0, 1990-2007.

JACKSON J. E. & MUDHOLKAR, G. S., "Control procedures for residuals associated with principal component analysis", *Technometrics*©, 21, 341-349, 1979.

JENSEN, D. R. & SOLOMON, H., "A gaussian approximation to the distribution of a quadratic form", J Amer Stat Assoc, 67, 898-902, 1972.

KOWALSKI, B. R., & SCHARTZKI, T. F., "Classification of archaeological artifacts by applying pattern recognition to trace element data", *Anal Chem*, 44, 2176-2180, 1972.

KUZUHARA, A., "Analysis of structural changes in bleached keratin fibers (black and white human hair) using Raman spectroscopy", *Biopolymers*, 81, 506-514, 2006.

LEHNINGER, A. L., "*Príncípios de Bioquímica*", Savier Editora Livros Médicos Ltda, São Paulo, 1989.

LEWIS, E. N.; SHOPPELREI, J.; LEE, E., "Near-infrared chemical imaging and the PAT iniciative", *Spectroscopy*, 19, 28-35, 2004.

LIED T. T.; GELADI P.; ESBENSEN K. H., "Multivariate image regression (MIR): implementation of image PLSR-first forays", *J Chemom*, 14, 585-598, 2000.

LIED T. T. & ESBENSEN K. H., "Principles of MIR, multivariate image regression I: regression typology and representative applications studies", *Chemom Intell Lab Syst*, 58, 213-226, 2001.

LIU J. J.; BHARATI M. H.; DUNN K. G.; MACGREGOR J. F., "Automatic masking in multivariate image analysis using support vector machines", *Chemom Intell Lab Syst*, 79, 42-54, 2005.

LYNCH, M. & WOOLGAR, S. (editores), "*Representation in Scientific Practice*", The MIT Press, London, 1990.

LYON, R. C. LESTER, D. S., LEWIS, E. N., LEE, E. N., YU, L. X., JEFFERSON, E. H., HUSSAIN, A. S., "Near-infrared spectral imagins for quality assurance of pharmaceutical products: analysis of tablets to assess poweder blend homogeneity", *AAPS Pharmatech*, 3, 1-15, 2002.

MARTENS, H. & NAES, T., "*Multivariate Calibration*", John Wilwy & Sons, Great Britain, 1993.

MATHWORKS INC., MatLab® for Windows, version 6.5, MathWorks Inc., 2002.

MCMULLEN, R. L.; JACHOWICZ, J.; KELTY, S. P., "Correlation of AFM/LFM with combing forces of human hair", *IFSCC Magazine*, 3, 39-45, 2000.

MCMULLEN, R. L. & KELTY, S. P., "Investigantion of human hair fibers using lateral force microscopy", *Scanning*, 23, 337-345, 2001.

MONTEIRO, V. F., "*Fibras Capilares: efeitos de diferentes agentes nas alterações físicas e químicas",* Tese de Doutorado, UFSCAR, 2003.

MONTEIRO, V. F., MACIEL, A. P., LONGO, E., "Thermal analysis of human hair", *J Therm Anal Calorim*, 79, 289-293, 2005.

NATTKEMPER, T. W., "Multivariate image analysis in biomedicine", *J Biom Inf*, 37, 380-391, 2004.

NIST/SEMATECH e-Handbook of Statistical Methods, Acesso em 20/07/2007, última atualização em 18/07/2006 - http://www.itl.nist.gov/div898/handbook/.

O'CONNOR, S. D.; KOMISAREK, K. L.; BALDESCHWIELER, J. D., "Atomic Force Microscopy of Human Hair Cuticles: a Microscopy Study of Environmental Effects on Hair Morphology", *J Invest Dermatol*, 105, 96-99, 1995.

OLIVIERI, A. C.; FABER, N. K.; FERRÉ, J.; BOQUÉ, R.; KALIVAS, J. H.; MARK, H., "Uncertainty estimation and figures of merit for multivariate calibration, IUPAC technical report", *Pure Appl Chem*, 78, 633-661, 2006.

OZEKI, H.; ITO, S.; WAKAMATSU, K.; THODY, A. J., "Spectrophotometric Characterization of Eumelanin and Pheomelanin in Hair", *Pigm Cell Res*, 9, 265–270, 1996.

PIMENTEL, M. F.; NETO, B. B., "Calibração: uma revisão para químicos analíticos", *Quim Nova*, 19, 268-277, 1996.

PHIPPS, J. B., "Dendrogram topololy", *Sys Zoology*, 20, 306-308, 1971.

REICH, G., "Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications", *Adv Drug Delivery Rev*, 57, 1109-1143, 2005.

REIS, M. M., "*Desenvolvimento e Aplicação de Métodos Quimiométricos de Ordem Superior*", Tese de Doutorado, IQ/UNICAMP, Brasil, 2002.

RIBEIRO, F. A. L. & FERREIRA. M. M. C., "QSAR Models of phototoxicity of polycyclic aromatic hydrocarbon" - *J Mol Struc –Theochem*, 719, 191-200, 2005.

ROBBINS, C. R., "*Chemical and Physical Behaviour of Human Hair*", 3rd Ed., Editora Springer, New York, 1994.

ROGGO, Y.; JENT, N.; EDMOND, A.; CHALUS, P.; ULMSCHNEIDER, M., "Characterizing process effects on pharmaceutical solid forms using near-infrared spectroscopy and infrared imagins", *Eur J Pharm Sci*, 61, 1000-110, 2005.

SANCHEZ, E. & KOWALSKI, B. R., "Tensorial resolution: a direct trilinear decomposition" *J Chemom*, 4, 29, 1990.

SANCHEZ, E. & KOWALSKI, B. R., "Tensorial calibration: I. First order calibration" *J Chemom*, 2, 247, 1988.

SANCHEZ, E. & KOWALSKI, B. R., "Tensorial calibration: II. Second order calibration *Anal Chem*, 58, 496, 1988.

SANT'ANNA, A. L. S., "*Estudo da deposição de ceramidas sobre a fibra capilar para o combate a danos cuticulares*", Dissertação de Mestrado, IQ/UNICAMP, 2000.

SAUERMANN, G.; HOPPE, U.; LUNDERSTADT, R.; SCHUBERT, B.; "Measurement of surface Profile of Human-Hair by Surface Profilometry", *J Soc Cosmet Chem*, 39, 27-42, 1988.

SAVITZKY, A., & GOLAY, M. J. E., "Smoothing and differentiation of data by simplified least squares procedures", *Anal Chem*, 36, 1627-1639, 1964.

SENA, M. M., "*Aplicação de Métodos Quimiométricos de Primeira e Segunda Ordem na Determinação Direta de Fármacos por Espectroscopia Molecular*", Tese de Doutorado, IQ/UNICAMP, Brasil, 2004.

SILVA, A. L. S. & JOEKES, I, "Rhodamine B diffusion in hair as a probe for structural integrity", *Colloids Surf B*, 40, 19-24, 2005.

SIMONCELLI, E. P. & OLSHAUSEN, B. A., "Natural image statistics and neural representation", *Annu Rev Neurosci*, 24, 193-216, 2001.

SMILDE, A. K., "Comments on multilinear PLS", *J Chemom*, 11, 367-377, 1997.

SMILDE, A. K., BRO, R. & GELADI, P., "*Multiway Analysis: Applications in the Chemical Sciences*", John Wiley & Sons, 2004.

SMITH J. A., "A quantitative method for analysing AFM images of the outer surfaces of human hair", *J Microsc*, 191, 223-228, 1998.

SWIFT, J. A., "Human hair cuticle: Biologically conspired to the owner´s advantage", *J Cosmet Sci*, 50, 23-47, 1999.

SWIFT, J. A. & BROWN, A. C., "The critical determination of fine changes in the surface architecture of human hair due to cosmetic treatment", *J Soc Cosmet Chem*, 23, 695-702, 1972.

SWIFT, J. A. & SMITH J. A., "Atomic force mycroscopy of human hair", *Scanning*, 22, 310-318, 2000.

VAN DEN BROEK, W. H. A.; DERCKS, E. P. P. A.; VAN DE VEN, E. W., WIENKE, D.; GELADI, P., BUYDENS, L. M. C., "Plastic identification by remote sensing spectroscopic NIR imaging using kernel partial least squares (KPLS)", *Chemom Intell Lab Syst*, 35, 187-197, 1996.

VAN ESPEN, P.; JANSSENS, G.; VANHOOLST, W.; GELADI, P., "Imaging and image processing in analytical chemistry", *Analusis*, 20, 81-90, 1992.

WATT, I., "*The principles and practice of electron microscopy*", 2nd. Ed., Cambridge University Press, 1997.

WISE, B. M.; GALLANGHER, N. B.; BUTLER, S. W.; WHITE, D. AND BARNA, G. G., "A comparison of principal components analysis, multi-way principal component analysis, tri-linear decomposition and parallel factor analysis for fault detection in a semiconductor etch process", *J Chemom*, 13, 379-396, 1999.

WISE, B.; GALLANGHER, N. B.; BRO, R.; SHAVER, J. M., "*PLS Toolbox 3.0, for use with MATLABTM - Manual*", Eigenvector Research, Inc., USA, 2005.

WOLD, S; "Pattern recognition by means of disjoint principal componentes models", *Pattern Recognit*, 8, 127-139, 1976.

WOLD, S.; ESBENSEN, K.; GELADI, P., "Principal Component Analysis", *Chemom Intell Lab Sys*, 2, 37-52, 1987.

WOLD, S.; GELADI, P., ESBENSEN, K. AND OHMAN, J., "Multi-way principal components and PLS analysis", *J Chemom*, 1, 41-56, 1987.

WOLD, S. & SJÖSTRÖM, M., "SIMCA: a method for analysing chemical data in terms of similarity and analogy", *Chemometrics*, Chapter 12, ACS Symposium Series, ed. R. F Gould, 1977, USA.

YOU, H. & YU, L., "Atomic force microscopy as a tool for study of human hair", *Scanning*, 19, 431-437, 1997.

YU, H.; MACGREGOR, J. F.; HAARSMAN, G.; BOURG, W., "Digital imaging for online monitoring and control of industrial snack food process", *Ind Eng Res*, 42, 3036-3044, 2003.