UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
DEPARTAMENTO DE COMPUTAÇÃO

EXTRAÇÃO DE ATRIBUTOS EM IMAGENS DE SENSORIAMENTO REMOTO UTILIZANDO *INDEPENDENT COMPONENT ANALYSIS* E COMBINAÇÃO DE MÉTODOS LINEARES

ALUNO: ALEXANDRE LUIS MAGALHÃES LEVADA
ORIENTADOR: PROF. DR. NELSON D. A. MASCARENHAS

SÃO CARLOS
FEVEREIRO/2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

L655ea

Levada, Alexandre Luís Magalhães.
Extração de atributos em Imagens de sensoriamento remoto utilizando Independent Component Analysis e combinação de métodos lineares / Alexandre Luís Magalhães Levada. -- São Carlos : UFSCar, 2006.
103 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Reconhecimento de padrões. 2. Sensoriamento remoto. 3. Seleção de atributos. I. Título.

CDD: 006.4 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"Extração de Atributos em Imagens de Sensoriamento Remoto Utilizando Independent Component Analysis e Combinação de Métodos Lineares"

ALEXANDRE LUÍS MAGALHÃES LEVADA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Membros da Banca:

Nelson Delfino d'Ávila Mascarenhas

Prof. Dr. Nelson Delfino d'Ávila Mascarenhas
(Orientador – DC/UFSCar)

Prof. Dr. Jander Moreira
(DC/UFSCar)

C. Freitas.

Profa. Dra. Corina da Costa Freitas
(INPE/São José dos Campos)

São Carlos
Fevereiro/2006

AGRADECIMENTOS

À toda minha família, em especial aos meus pais e irmã, pelo amor e apoio incondicionais, incentivos para que eu pudesse alcançar mais este objetivo, mas principalmente pela companhia ao longo de todos esses anos de vida.

Ao Prof. Dr. Nelson D. A. Mascarenhas, meu orientador, pela confiança e amizade, pelos ensinamentos, opiniões, correções, e pela oportunidade de poder compartilhar tanto conhecimento.

A todo pessoal do GAPIS, pela amizade, consideração e discussões sobre os mais variados temas e assuntos.

A todos do DC e PPG-CC, em especial pelo aprendizado através das aulas e dos professores durante o mestrado.

Ao INPE, em especial a Dra. Corina da Costa Freitas e ao Dr. Sidnei Sant'Anna, pelo fornecimento das imagens de satélite da região de Tapajós, Pará/Brasil, que foram extremamente importantes para o desenvolvimento do trabalho e contribuíram muito para os resultados experimentais.

Ao CNPq pelo apoio financeiro na realização do projeto.

Em suma, a todas aquelas pessoas que direta ou indiretamente contribuíram de alguma maneira para que tudo isso se tornasse realidade.

“The greatest challenge to any thinker is stating the problem in a way that will allow a solution.”

(Bertrand Russel)

RESUMO

Métodos para extração de atributos compõem uma etapa fundamental em aplicações na área de reconhecimentos de padrões. O presente trabalho apresenta uma metodologia para melhorar o desempenho da classificação criando modelos para fusão de atributos que combinam métodos estatísticos de segunda ordem com métodos de ordens superiores, superando limitações existentes nas abordagens tradicionais, como problemas de mal-condicionamento, o que pode provocar instabilidade na estimação dos componentes independentes, além de eventuais amplificações de ruídos. O esquema resultante é utilizado para combinar atributos obtidos através de diversos métodos num único vetor de padrões em duas abordagens: Fusão Concatenada e Fusão Hierárquica. A metodologia proposta é aplicada em diversos estudos de casos, incluindo imagens multiespectrais e hiperespectrais de sensoriamento remoto, classificadas utilizando-se a abordagem de máxima verossimilhança. Resultados indicam que essa metodologia supera métodos de segunda ordem tradicionais em alguns casos, constituindo um válido e interessante ferramental para análise e classificação de dados multivariados.

ABSTRACT

Methods for feature extraction represent an important stage in statistical pattern recognition applications. In this work we present how to improve classification performance creating a feature fusion framework to combine second and higher order statistical methods, avoiding existing limitations of the individual approaches and problems as ill-conditioned behavior, which may cause unstable results during the estimation of the independent components (whitening process) and eventual noise amplifications. The resulting scheme is used to combine features obtained from a variety of methods into a unique feature vector defining two approaches: Concatenated and Hierarquical Feature Fusion. The methods are tested on both multispectral and hyperspectral remote sensing images, which are classified using the maxver (maximum likelihood) approach. Results indicate that the technique outperforms the usual methods in some cases, providing a valid useful tool for multivariate data analysis and classification.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Atualmente, a análise de imagens multiespectrais é uma tarefa freqüentemente utilizada em aplicações de diversas áreas da ciência. Porém, não se trata de um processo trivial. A extração de informações específicas de um conjunto de dados multivariados, em especial quando a dimensionalidade se torna grande, é um problema complexo que requer a aplicação de técnicas baseadas em fundamentos da teoria de processamento de imagens e sinais.

Em sensoriamento remoto é possível adquirir informações sobre um objeto ou região sem a necessidade de contato físico, através da detecção da energia em pequenas faixas do espectro eletromagnético (ultravioleta, luz visível, infravermelho e microondas). A energia proveniente do sol incide sobre a superfície terrestre, que reflete, absorve e emite radiação eletromagnética, que pode ser captada e medida por sensores. Assim, os dados obtidos pelos sensores multiespectrais consistem em um conjunto de medidas, contendo informações de uma determinada região. As diferentes faixas de comprimento de onda (ou freqüências) caracterizam o mecanismo de interação entre a radiação eletromagnética e os materiais iluminados e as medidas dependem de propriedades dos materiais presentes na região, como por exemplo a pigmentação, a estrutura do terreno, a capacidade térmica dos materiais e a composição molecular, entre outras.

Partindo do princípio de que cada tipo de material tem uma resposta espectral diferente, pode-se classificar diferentes materiais que compõem uma região. Um dos propósitos de se adquirir dados de imagens de sensoriamento remoto é, então, classificar regiões da superfície terrestre presentes em uma determinada cena. O desenvolvimento de sensores multi/hiperespectrais, que permitem a produção de imagens com várias bandas, deve ser acompanhado também da criação de métodos mais eficientes para a extração de informações contidas nesses dados. Por isso, é importante visualizar a análise dos dados não isoladamente, mas como uma das etapas de todo um processo. A Figura 1, adaptada de (LANDGREBE, 1999), fornece uma visão geral das etapas desse processo, desde a obtenção dos dados até a classificação, sendo que este trabalho está restrito às duas últimas fases, extração de atributos e classificação, com maior ênfase à extração.

Os sensores medem a energia refletida (emitida) pelas áreas de interesse. Os dados são coletados e transmitidos para a etapa de processamento e extração de informações para posterior utilização.

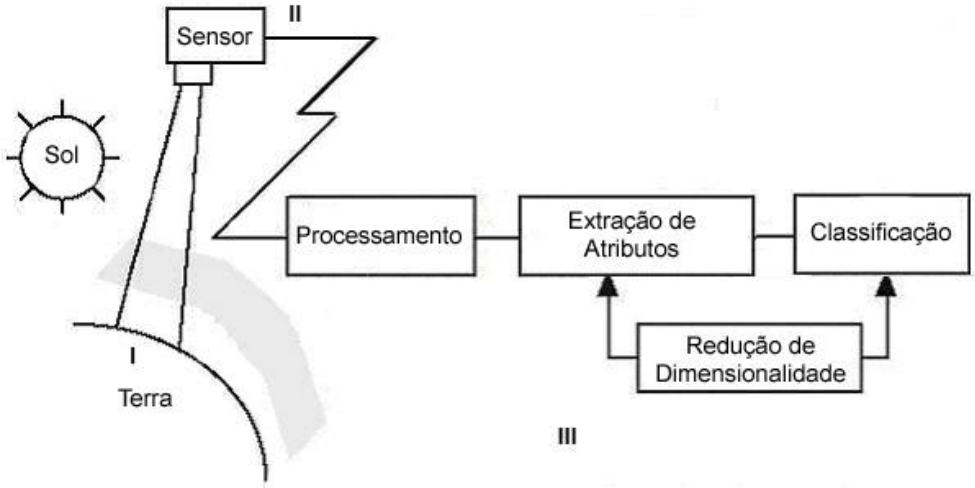


Figura 1*. Etapas envolvidas no processo de classificação de imagens multiespectrais.*

Pode-se dividir logicamente o sistema em três etapas: o cenário, os sensores e o processamento. O cenário se refere à componente do sistema que compreende a superfície terrestre, o sol e a atmosfera. Esta etapa possui duas características peculiares: não está, em momento algum, sob controle humano (nem na construção do sistema, nem na sua utilização) e consiste na parte mais dinâmica e complexa do sistema. Assim, obviamente, essa etapa não pode ser otimizada. Como exemplo, pode-se observar a variação da assinatura espectral de um determinado tipo de vegetação devido às condições climáticas ou às estações do ano.

A segunda parte do sistema é composta pelos sensores. Essa etapa é caracterizada pelo fato de que, apesar de estar sob desenvolvimento humano, não está sob controle do analista durante o período de aquisição dos dados. Isto é, o analista deve aceitar os dados em termos de parâmetros como a resolução espacial e espectral, a precisão da quantização da imagem, o campo de visão e ângulos dos sensores.

E, finalmente, a terceira parte do sistema é a etapa de processamento dos dados obtidos, sobre a qual o analista tem total controle sobre as operações. Nessa etapa são realizadas escolhas que dizem respeito à seleção dos algoritmos a serem utilizados e à classificação das imagens selecionadas. Uma das finalidades desse trabalho é combinar diferentes técnicas de extração de atributos, visando otimizar a representação dos dados na redução de dimensionalidade com o objetivo de melhorar o desempenho da classificação.

1.1 Dados Multivariados

Dados multivariados são representados em várias dimensões, e como conseqüência cada amostra individual é definida como um vetor. Uma forma de ocorrência particular desses dados pode ser obtida em imagens multiespectrais. Segundo (LANDGREBE, 1999), existem basicamente três formas de representação para imagens multiespectrais: a representação espacial, que mostra o relacionamento geométrico entre os pixels da imagem, a representação no espaço espectral, com a resposta espectral dos elementos em função do comprimento de onda, e a representação no espaço de atributos, no qual os pixels são considerados pontos ou vetores em um espaço *n*-dimensional.

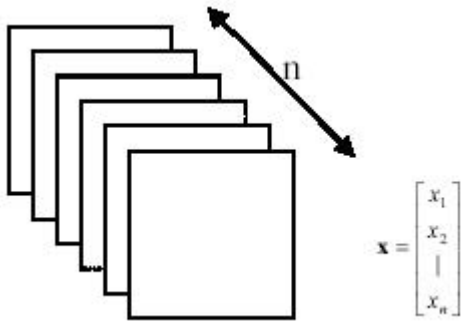


Figura 2*. Dados Multivariados em imagens multiespectrais.*

1.2 A Representação Espacial

A representação espacial de uma imagem multiespectral pode ser considerada como a mais natural para o sistema de percepção visual humano. A idéia é exibir amostras dos dados utilizando uma relação geométrica entre elas, fornecendo uma imagem do cenário observado. Pode-se observar a imagem em tons de cinza ou selecionar três bandas simultaneamente, atribuindo cada uma delas a uma componente RGB e assim obter uma melhor visualização.

Essa representação não carrega uma grande quantidade de informação relevante ao processamento dos dados multivariados. Além disso, relacionamentos entre bandas não são aparentes. Por outro lado, é extremamente útil para fornecer uma visão geral dos dados, definir quais propriedades espectrais pertencem a cada classe e como um meio para o analista associar pontos de dados multiespectrais (pixels) a localidades específicas no cenário terrestre, para seleção de amostras de treinamento, por exemplo. A Figura 3 mostra um exemplo de representação espacial de dados multivariados.

***Figura 3**. Representação espacial da imagem multiespectral*

1.3 O Espaço Espectral

Com o surgimento dos sensores multiespectrais, criou-se a possibilidade de verificar como a resposta espectral medida por um pixel varia como função do comprimento de onda. A resposta espectral possui a característica de fornecer ao analista informações diretamente interpretáveis, especialmente quando o número de bandas é consideravelmente elevado e cada classe/material presente na imagem contém uma assinatura espectral única. Um grande número de bandas representa uma melhor amostragem do intervalo espectral.

Porém, a resposta espectral de qualquer área da superfície terrestre tende a variar de maneira característica. Em uma plantação de soja, por exemplo, a resposta espectral não é uniforme para todos pixels, mas varia de maneira peculiar em torno de um valor médio, fato que pode ser utilizado como auxílio na identificação de diferentes materiais presentes na região. A Figura 4 mostra três exemplos de respostas espectrais. Nota-se que cada material possui uma assinatura espectral diferente (vegetação, água e solo).

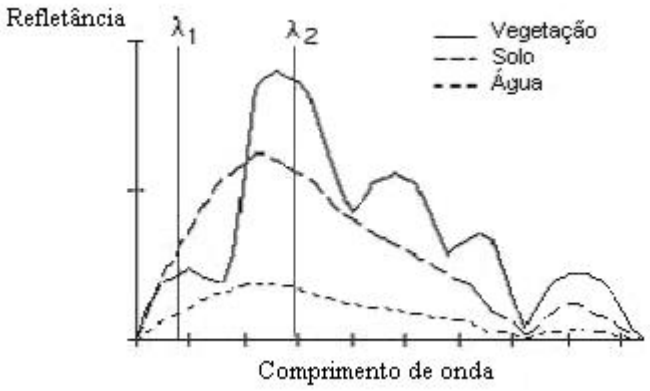


***Figura 4**. Resposta espectral para diferentes tipos de materiais*

1.4 O Espaço de Atributos

Para propósitos de reconhecimento de padrões e processamento por métodos computacionais, a representação no espaço de atributos se mostra como a mais conveniente a ser utilizada. Trata-se de uma representação matemática que amostra a resposta espectral de cada pixel em cada uma das N bandas, criando um vetor N-dimensional contendo toda a informação espectral disponível sobre os pixels da imagem. É uma maneira quantitativa de representar não apenas os valores numéricos associados a cada pixel, mas também como esses valores variam para diferentes classes de materiais. Um exemplo dessa representação utilizando três classes pode ser visualizado na Figura 5. Classes diferentes ocupam regiões distintas do espaço de atributos, possibilitando a partição do mesmo em regiões de decisão.

Figura 5. *Representação do espaço de atributos.*

1.5 Propriedades de Dados Hiperespectrais

Em problemas de otimização combinatórios, sabe-se que o esforço computacional cresce exponencialmente conforme o aumento do número de dimensões. Em estatística, vários problemas ocorrem na estimação de parâmetros e densidades devido à insuficiência de dados. Este efeito negativo é resultado de propriedades geométricas e estatísticas de espaços de atributos com alta dimensionalidade. Nessas situações a dimensionalidade torna-se uma complicação, pois impossibilita uma modelagem estatística adequada para o problema. Para verificar algumas características presentes em espaços hiperdimensionais, (JIMENES; LANDGREBE, 1998) mostram algumas propriedades não intuitivas de dados multivariados em espaços de grande dimensionalidade, relevantes na área de reconhecimento de padrões. Tais propriedades ajudam a esclarecer as razões da significativa diferença existente entre a análise de

dados multiespectrais e hiperespectrais. Algumas propriedades descritas a seguir, relacionam a geometria desses espaços com propriedades estatísticas presentes em densidades de probabilidade, revelando um comportamento nada trivial.

- **Conforme o aumento da dimensionalidade, o volume de um hiperelipsóide tende a se concentrar na concha elíptica externa (próximo da borda).**

  Isto é equivalente a dizer que em uma distribuição gaussiana os dados concentram-se nas caudas, ou seja, longe da média da distribuição. Pode ser mostrado (JIMENES; LANDGREBE, 1998) (LANDGREBE, 1999) que esse fato traz duas conseqüências importantes para vetores de padrões em espaços de atributos hiperdimensionais. A primeira delas é que um espaço de muitas dimensões é, em sua maior parte, vazio, o que implica dizer que dados multivariados de grande dimensionalidade geralmente se encontram em uma estrutura dimensional inferior. Como conseqüência, os dados podem ser projetados em um subespaço de menor dimensão, sem perda significativa de informação em termos de separabilidade entre classes. A segunda conclusão é que em densidades de probabilidade conhecidas (normal, uniforme etc.) os dados se concentram nas bordas, o que torna a estimação dessas densidades extremamente complicada, pois as vizinhanças tendem a ser muito esparsas. Isso faz com que seja obrigatória a utilização de um método para extração de atributos (redução de dimensionalidade), caso contrário o desempenho da classificação é severamente degradado.

- **Conforme a dimensionalidade aumenta, os vetores diagonais tendem a se tornar ortogonais aos vetores da base.**

  O co-seno de um ângulo entre qualquer vetor diagonal (que possui o mesmo valor em todas as componentes) e os vetores ortogonais da base é dado pela expressão

$$\cos\left(\theta_d\right) = \pm\frac{1}{\sqrt{d}} \tag{1.1}$$

  onde d representa a dimensionalidade do espaço. Pode-se notar que $\lim_{d\to\infty}\cos\left(\theta_d\right) = 0$, ou seja, em espaços com muitas dimensões, os vetores diagonais tornam-se praticamente ortogonais aos vetores da base. Neste caso, a

projeção de um agrupamento de dados em algum vetor diagonal poderia destruir completamente a informação contida nos dados multivariados. Portanto, deve-se projetar os dados utilizando critérios adequados, caso contrário o desempenho da classificação também é degradado. O gráfico da Figura 6, obtido em (LANDGREBE, 1999), mostra como o ângulo entre um vetor diagonal e os eixos coordenados se aproxima de 90º quando o número de dimensões aumenta.

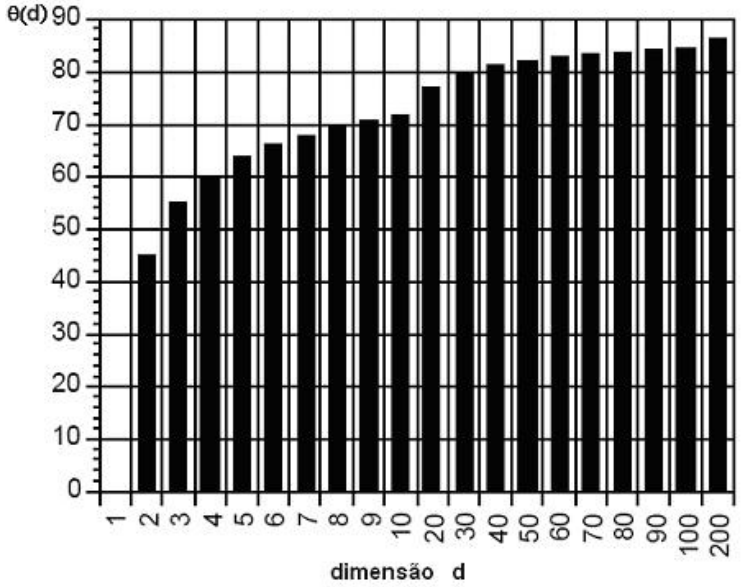


Figura 6*. Variação do ângulo entre um vetor diagonal e os vetores da base pela dimensionalidade do espaço.*

- **O número de amostras de treinamento para classificação supervisionada aumenta como uma função da dimensionalidade.**

É mostrado em (FUKUNAGA, 1990) que o número de amostras de treinamento necessárias é linearmente relacionado com a dimensionalidade para um classificador linear e ao quadrado da dimensionalidade no caso de um classificador quadrático. Em termos de classificadores não paramétricos a situação é ainda mais severa, pois resultados encontrados em (SCOTT, 1992), (HWANG; LAY; LIPPMAN, 1994) mostram que, conforme o aumento da dimensionalidade, o número de amostras deve aumentar exponencialmente para se ter uma estimativa efetiva das densidades multivariadas. Embora dados com alta dimensionalidade contenham mais informações, as características citadas anteriormente mostram que com as técnicas atuais é difícil extrair essas informações a menos que a quantidade de amostras disponíveis seja grande o suficiente.

Na teoria, o erro na classificação bayesiana decresce monotonicamente conforme o número de atributos aumenta. Entretanto, na prática, quando um número fixo de amostras é usado para desenvolver o classificador, o erro de classificação tende a aumentar conforme mostra o gráfico da Figura 7, obtida em (FUKUNAGA, 1990). Esse fato é conhecido como o fenômeno de Hughes. É provado em (HUGHES, 1968) que para um dado número finito de amostras *N*, existe uma dimensionalidade ótima *n*, acima da qual a acurácia média de classificação pode decrescer. Esse fenômeno é observado experimentalmente, sendo que em geral, é necessário que o número de amostras seja substancialmente maior que a dimensionalidade.

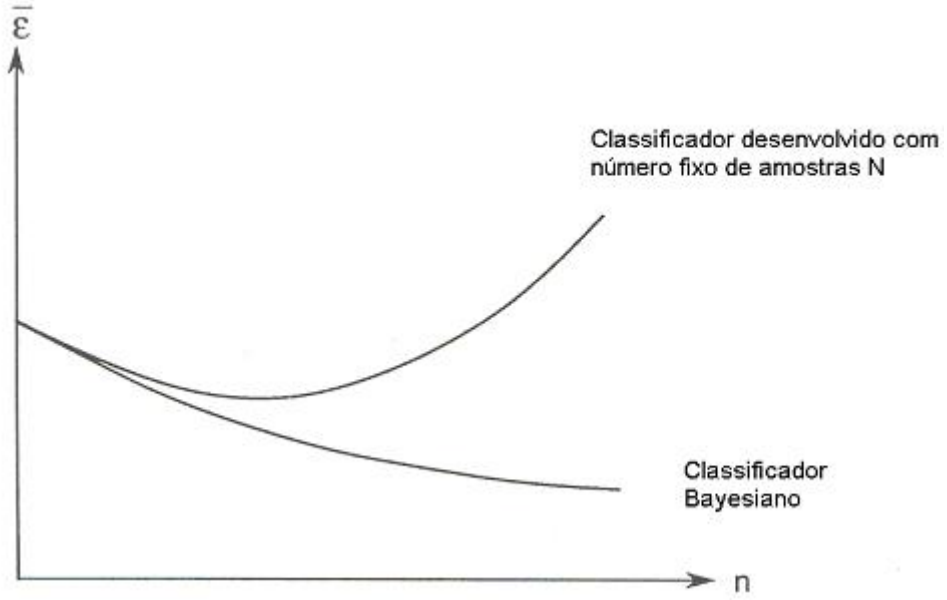


Figura 7*. Ilustração do fenômeno de Hughes.*

1.6 Motivação e Objetivos

A grande complexidade dos dados multivariados é, em muitos casos, um fator limitante para as fases de análise e classificação. Para solucionar esse problema, existem métodos clássicos de redução de dimensionalidade, como por exemplo, a transformação de Karhunen-Loève, que implementa a técnica PCA (*Principal Component Analysis*). Porém, em alguns casos, tais métodos não se mostram totalmente adequados para o problema de classificação. Por esse motivo, a abordagem ICA (*Independent Component Analysis)* se mostra interessante, uma vez que, nesse caso, a redução de dimensionalidade é realizada projetando-se os dados multivariados em um subespaço diferente do obtido com PCA.

Dessa forma, espera-se que atributos obtidos através de algoritmos ICA consigam prover uma melhor separabilidade entre as classes através de um mapeamento não supervisionado (sem informações a priori sobre as classes), utilizando dados reais extraídos de imagens multiespectrais de sensoriamento remoto. O objetivo desses métodos é utilizar apenas a informação presente nos dados multivariados através de estatísticas de ordem superiores. Um outro aspecto interessante é que atualmente existe um grande interesse em se estudar os princípios estatísticos referentes a codificação visual humana, pois como se sabe, os humanos são considerados excelentes reconhecedores de padrões. Segundo (BORGNE; GUÉRIN-DUGUÉ; ANTONIADIS, 2004), (BORGNE et al., 2003) e (INKI, 2004), estudos recentes sobre percepção visual sugerem que, em mamíferos, o córtex visual primário atua como um modelo de representação por redução de redundância. Essa idéia é compatível com a abordagem adotada em ICA, na qual os atributos produzidos são independentes, otimizando a representação para minimizar a redundância dos dados, uma vez que essa técnica não se restringe apenas ao não-correlacionamento (como em PCA), seguindo os princípios dos modelos de codificação. Além disso, recentemente, algoritmos ICA têm sido utilizados como ferramenta em diversos tipos de aplicações, dentre as quais reconhecimento de faces (DU; HU; SHYU, 2004) e definição de atributos espaço-temporais para classificação de sinais de ECG (HERRERO et al., 2005), dentre outras.

Portanto, um dos objetivos do trabalho consiste em buscar um entendimento de espaços de atributos multidimensionais no contexto de sensoriamento remoto e utilizar esse conhecimento para combinar métodos de extração de atributos visando a classificação de imagens multiespectrais utilizando a técnica ICA conjuntamente com métodos estatísticos de segunda ordem e, assim, superar limitações existentes nos métodos tradicionais para melhorar o desempenho do classificador.

Outro aspecto a ser discutido nesse trabalho é uma comparação entre as técnicas de extração de atributos tradicionais e métodos estatísticos de ordem superiores (ICA) para um mapeamento não supervisionado dos dados em imagens hiperespectrais. Dessa forma, os resultados podem ser verificados e o desempenho de cada técnica analisado para dados reais extraídos de imagens de sensoriamento remoto. Os esquemas de combinação propostos (hierárquico e concatenado) também podem contemplar etapas adicionais de pós-processamento, como seleção de atributos, com o intuito de melhorar ainda mais a capacidade discriminante do conjunto de atributos utilizado para a classificação.

1.7 Organização do Texto

O capítulo 2 contém uma visão geral de conceitos teóricos e objetivos da área de reconhecimento de padrões referentes ao tema do projeto.

No capítulo 3 há uma breve introdução sobre conceitos básicos de probabilidade e estatística utilizados na estimação ICA. São discutidos conceitos e definições como, independência e estatística de ordens superiores.

O capítulo 4 introduz os métodos estatísticos tradicionais (segunda ordem) de extração de atributos como PCA, através da Transformação de Karhunen-Loève, LDA (*Linear Discriminant Analysis*), suas características e limitações, além de critérios probabilísticos e medidas de distância entre classes para seleção de atributos.

O capítulo 5 apresenta a formulação teórica do modelo ICA, além de um estudo sobre diversos métodos de estimação ICA que descreve cada uma das principais abordagens, verificando que existe uma equivalência entre os critérios utilizados.

O capítulo 6 discute o desenvolvimento do projeto, através das seções: metodologia, medidas de desempenho, materiais e métodos, fusão de atributos e resultados. São apresentados detalhes sobre as imagens multiespectrais/hiperespectrais utilizadas no trabalho e todos os resultados obtidos nos experimentos, divididos em três estudos de casos distintos.

2 RECONHECIMENTO DE PADRÕES

Basicamente, o principal objetivo da área de reconhecimento de padrões é a classificação de objetos em classes. Nesse caso, objetos são representados em termos de seus atributos, através dos vetores de padrões. Em geral, a distinção entre padrões e atributos se deve à etapa de extração de atributos, ou seja, os atributos mais relevantes são extraídos do conjunto original de padrões.

O problema de reconhecimento de padrões pode ser visto como uma tarefa de categorização, ou seja, atribuir cada amostra a uma de N possíveis classes, sendo que as classes podem ser previamente definidas pelo analista/desenvolvedor do sistema (amostras com rótulos), caracterizando o modelo de classificação supervisionada, ou podem ser formadas criando-se aglomerados por meio de padrões de similaridade (amostras sem rótulos), no caso do modelo de classificação não-supervisionada.

Durante a especificação de uma aplicação em reconhecimento de padrões devem ser contemplados aspectos como definição das classes, representação dos padrões, extração e seleção de atributos, seleção de amostras de treinamento e teste, classificação e avaliação dos resultados. Em geral, o próprio domínio do problema obriga a escolha da técnica de extração de atributos, do esquema de representação e do modelo de classificação a serem utilizados.

Existem diversas abordagens utilizadas em reconhecimento de padrões, sendo algumas das mais relevantes: *template matching*, a abordagem estatística, a abordagem sintática ou estrutural e as redes neurais. Na literatura, a abordagem estatística se destaca principalmente devido ao seu forte embasamento teórico e matemático. As redes neurais também são amplamente utilizadas atualmente. A base teórica para o reconhecimento de padrões estatístico é a Teoria de Decisão Bayesiana, cujo principal objetivo é o particionamento do espaço de atributos em regiões ótimas de decisão, no sentido de minimizar o risco de Bayes. Pode ser mostrado que a minimização do risco de Bayes é equivalente a minimização da probabilidade de erro, no caso de uma matriz de perdas do tipo [0,1]. Uma extensa literatura sobre os princípios do reconhecimento de padrões estatístico pode ser encontrada em (YOUNG; CALVERT, 1974), (FUKUNAGA, 1990) e (DUDA; HART; STORK, 2001).

Ainda na abordagem estatística, cada padrão é representado em termos de seus d atributos e pode ser visualizado como um vetor num espaço d-dimensional. O objetivo é

escolher os atributos que permitam aos vetores de padrões pertencentes a diferentes classes ocupar regiões compactas e disjuntas do espaço de atributos. A eficiência da representação do conjunto de atributos é determinada através da separabilidade entre as classes.

Diversas estratégias podem ser utilizadas no desenvolvimento de um sistema de reconhecimento de padrões, dependendo do tipo de informação estatística disponível. Todas essas abordagens são mostradas em estrutura de árvore na Figura 8, adaptada de (JAIN; DUIN; MAO, 2000). Conforme se percorre a árvore do topo até as extremidades inferiores, no sentido da esquerda para a direita, menos informação se encontra disponível e a dificuldade no problema de classificação aumenta.

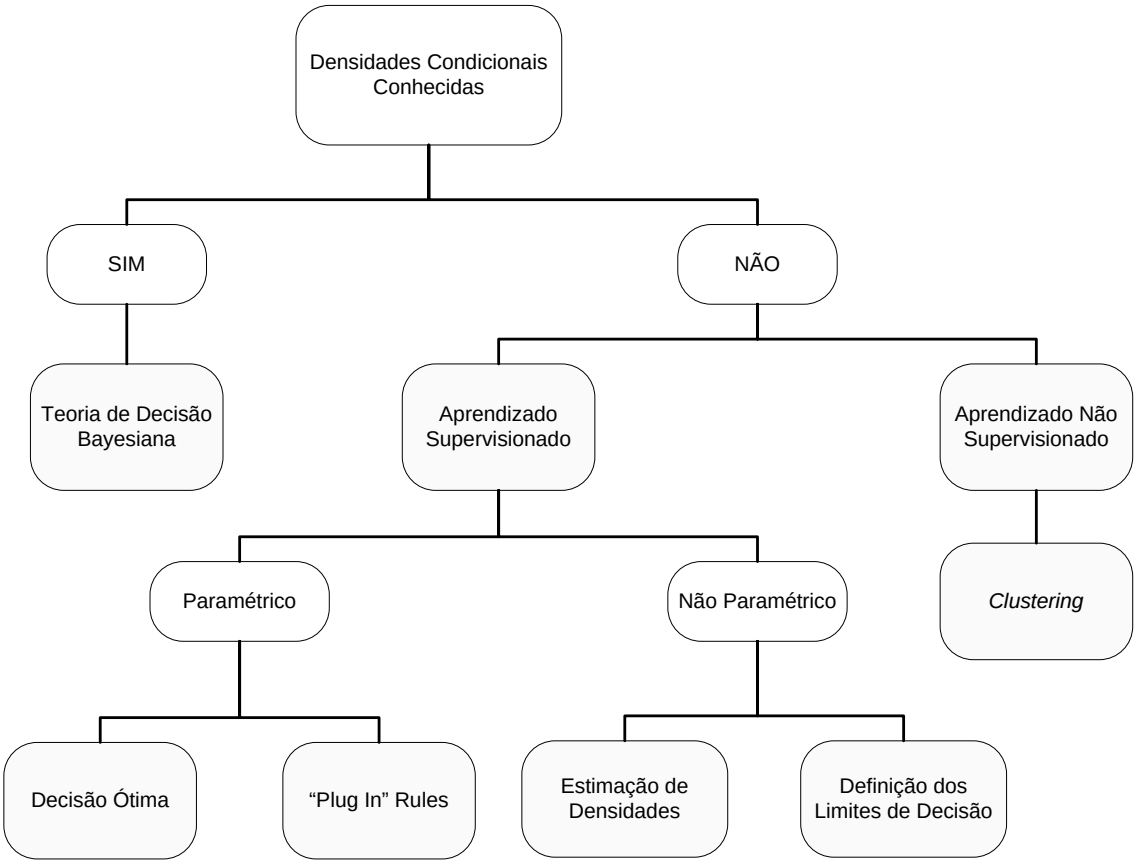


***Figura 8**. Abordagens para solução de problemas em reconhecimento de padrões estatístico*

2.1 O Problema da Dimensionalidade

O desempenho de uma classificação depende da relação existente entre o número de amostras, o número de atributos do vetor de padrões e a complexidade do classificador. É conhecido, na Teoria de Decisão Bayesiana, que a probabilidade de erro de classificação diminui conforme o aumento do número de atributos do vetor de

padrões (dimensionalidade), uma vez que as densidades condicionais são completamente conhecidas, o que equivale a dizer na prática que o número de amostras disponíveis é ilimitado.

Entretanto, na prática se observa que atributos adicionais podem degradar a performance da classificação se o número de amostras de treinamento que são usadas no desenvolvimento do classificador é relativamente pequeno em relação ao número de atributos. Em reconhecimento de padrões, esse fato é conhecido como o problema da dimensionalidade.

Na literatura, existem resultados que comprovam a existência do problema da dimensionalidade. Um deles, encontrado em (TRUNK, 1979), mostra este comportamento em um problema de classificação de duas classes equiprováveis com distribuição gaussiana multivariada e matriz de covariância igual à identidade. Foram considerados dois casos. No primeiro, o vetor média é conhecido, sendo possível a utilização da teoria de decisão bayesiana. Foi verificado que as duas classes são perfeitamente discriminadas aumentando-se arbitrariamente o número de atributos. No segundo caso, eram disponíveis n amostras de treinamento. O vetor média é estimado por máxima verossimilhança e a classificação é realizada substituindo-se a média real m por seu estimador $\hat{m}$ (*plug in rule*). É verificado que no caso limite a probabilidade de erro tende ao máximo valor possível.

Portanto, esses resultados indicam que, diferentemente do caso teórico, não se deve arbitrariamente aumentar o número de atributos quando os parâmetros das densidades condicionais das classes são estimados de um conjunto de amostras finito. Assim, uma implicação prática do problema da dimensionalidade, especificada em (JAIN; DUIN; MAO, 2000), é que o desenvolvedor da aplicação em reconhecimento de padrões deve tentar selecionar apenas um pequeno número de atributos relevantes quando se dispõe de um conjunto de treinamento limitado.

2.2 Redução de Dimensionalidade

Basicamente existem duas razões principais para manter a dimensionalidade da representação dos padrões a menor possível: o custo computacional, que resulta em um classificador mais rápido e com menos uso de memória, e a acurácia da classificação. Além disso, um número reduzido de atributos pode diminuir o problema da dimensionalidade quando o número de amostras de treinamento é limitado. Por outro lado, em alguns casos, uma redução no número de atributos pode levar a uma perda na

capacidade de discriminação dos dados, causando uma redução no desempenho da classificação.

O aspecto mais importante na redução de dimensionalidade é a escolha de um critério a ser otimizado. Um critério comum poderia ser o erro de classificação utilizando-se um determinado conjunto de atributos, mas esse erro não pode ser estimado confiavelmente quando a razão entre o número de amostras disponíveis e o número de atributos é pequena. Além da escolha de um critério adequado, é necessário determinar a dimensionalidade apropriada do espaço de atributos resultante.

2.3 Extração de Atributos

Métodos que criam novos atributos baseados em transformações ou combinações do conjunto de atributos original são denominados algoritmos de extração de atributos. Esses métodos determinam um subespaço apropriado de dimensionalidade m no espaço de atributos original de dimensão d ($m < d$). Transformações lineares utilizadas em métodos como PCA e LDA, uma abordagem que generaliza a função discriminante linear de Fischer, têm sido freqüentemente empregadas em reconhecimento de padrões para extração de atributos e redução de dimensionalidade. Basicamente, o método PCA representa a abordagem de aprendizado não supervisionado para o mapeamento dos dados, enquanto em LDA se tem a representação supervisionada, uma vez que são necessárias amostras de cada uma das classes para que seja possível a otimização do critério definido. A vantagem, nesse último caso, é que se consegue produzir projeções nas quais a separabilidade das classes é maior, pois a quantidade de informação disponível também é maior.

O método mais adotado na extração de atributos é o PCA, mais conhecido na literatura de reconhecimento de padrões como a transformação de Karhunen-Loève. Basicamente, esse método computa a expansão dos vetores de padrões nos m vetores próprios da matriz de covariância associados aos m maiores valores próprios. Como esse método utiliza os atributos mais significativos, a representação dos dados no subespaço linear minimiza o critério de erro médio quadrático.

Segundo a literatura, outros métodos como *Projection Pursuit* e ICA são mais apropriados, pois não são restritos apenas a propriedades estatísticas de segunda ordem. Dessa forma, espera-se conseguir um método não supervisionado de mapeamento (ICA) mais eficiente, através da utilização de informações adicionais relativas a estatística de ordens superiores. De acordo com (DUDA; HART; STORK, 2001), em geral, quando

utilizada como pré-processamento para classificação, a técnica ICA possui uma série de características que a tornam mais desejável em relação ao PCA linear ou não-linear. Técnicas ICA têm sido utilizadas com sucesso em BSS (*Blind Source Separation*) para extração de sinais individuais presentes em misturas, sem a necessidade de conhecimento prévio de cada um dos sinais originais.

A desvantagem da extração em relação à seleção de atributos é que, no primeiro caso, em se tratando de imagens, perde-se o sentido físico dos dados, pois os resultados das transformações aplicadas não são mais visualizados como imagens. Por outro lado, se a dimensionalidade dos dados é relativamente alta, existe um grande número de combinações possíveis de atributos, aumentando a complexidade dos métodos de seleção.

2.4 Classificação

Em reconhecimento de padrões, dificilmente se possui o conhecimento total da estrutura probabilística do problema. Em geral, tem-se um conjunto de amostras de cada classe e o objetivo é, a partir deste conhecimento limitado, desenvolver um classificador. Para esse problema existem basicamente duas abordagens: o método paramétrico e o não paramétrico de estimação.

2.4.1 Classificação por Máxima Verossimilhança

O método paramétrico de estimação assume uma forma pré-definida para a densidade de probabilidade, necessitando apenas da estimação dos parâmetros. O modelo paramétrico mais adotado na literatura é a distribuição normal. Desta forma, o problema é reduzido significativamente, pois estimação de parâmetros é um problema clássico da estatística que apresenta várias soluções, dentre elas a estimação por máxima verossimilhança (MV).

Na estimação por máxima verossimilhança, o objetivo é estimar os parâmetros desconhecidos utilizando um conjunto de vetores de atributos conhecidos em cada classe. Supondo $\vec{\theta}$ como o vetor de parâmetros e $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ o conjunto de amostras de uma classe, define-se a função de verossimilhança tomando-se $p\left(X; \vec{\theta}\right)$ como função de $\vec{\theta}$. Assim, o método de máxima verossimilhança estima o vetor de parâmetros maximizando a função de verossimilhança, através do gradiente da função. Considerando as densidades condicionais das classes como normais, ou seja:

$$p\left(\vec{x} \mid w_j\right) = \frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)\right\} \tag{2.1}$$

Pode-se mostrar que os componentes do vetor de parâmetros da *j*-ésima classe $\vec{\theta}_j = \left\{\vec{\mu}_j, \Sigma_j\right\}$, onde $\vec{\mu}_j$ é o vetor média da classe *j* e Σ_j é a matriz de covariância da classe *j*, são definidos como

$$\hat{\vec{\mu}}_j = \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i \tag{2.2}$$

$$\hat{\Sigma}_j = \frac{1}{n}\sum_{i=1}^{n} (\vec{x}_i - \vec{\mu}_j)(\vec{x}_i - \vec{\mu}_j)^T \ . \tag{2.3}$$

O processo de classificação pode, então, ser realizado através do cálculo de funções discriminantes g_j. Atribui-se o vetor de padrões observado $\vec{x}$ à classe w_j que fornece o máximo valor da função discriminante. Uma possível função discriminante é dada por

$$g_j\left(\vec{x}\right) = \ln p\left(\vec{x} \mid w_j\right) = \frac{d}{2}\ln\left(2\pi\right) - \frac{1}{2}\ln\left|\Sigma_j\right| - \frac{1}{2}\left(\vec{x}-\vec{\mu}_j\right)^T \Sigma^{-1}\left(\vec{x}-\vec{\mu}_j\right) \tag{2.4}$$

2.4.2 Clustering

No caso de classificação não supervisionada, não se possui conhecimento sobre a definição de cada classe. Desta forma, os métodos buscam definir classes em termos de aglomerações ou grupos de pontos no espaço de atributos através de medidas de similaridade. Em geral, a medida de similaridade é definida em função das distâncias entre os pontos, podendo ser isotrópica (independente da direção) no caso da distância euclidiana, ou não isotrópica (ponderação diferente para direções) como no caso da distância de Mahalanobis.

Basicamente, o algoritmo de *clustering k*-médias, além de adotar como medida de similaridade a distância euclidiana, define um índice de desempenho a ser minimizado, dado por:

$$S = \sum_{j=1}^{Nc} \sum_{x \in S_j} \left\|\vec{x} - \vec{m}_j\right\|^2 \tag{2.5}$$

com

$$\vec{m}_j = \frac{1}{N_j}\sum_{x \in S_j} \vec{x} \tag{2.6}$$

e N_c é o número de agrupamentos definido inicialmente, S_j é o j-ésimo agrupamento (*cluster*), $\vec{m}_j$ é o centro do *j*-ésimo agrupamento e N_j é o número de elementos do *j*-ésimo agrupamento. Uma motivação para a definição desse critério é a analogia com o caso probabilístico em que o valor c que minimiza $E\{x-c\}^2$ é $c = E\{x\}$.

3 CONCEITOS DE PROBABILIDADE E ESTATÍSTICA

Neste capítulo são apresentados brevemente alguns fundamentos importantes da teoria de probabilidade e estatística que estão profundamente relacionados às origens da teoria ICA. Dentre eles, figuram, por exemplo, conceitos como independência e estatística de ordens superiores. Essas definições são fundamentais na composição dos métodos e critérios utilizados nos algoritmos ICA e são freqüentemente referenciadas ao longo desse trabalho.

3.1 Independência Estatística

O principal conceito que constitui a base da teoria ICA é a independência estatística. Duas variáveis são independentes se o conhecimento do valor de uma delas não fornece nenhuma informação sobre o valor da outra, ou seja, não existe nenhuma redundância estatística. Matematicamente, independência estatística é definida em termos de funções densidade de probabilidade. As variáveis aleatórias *x* e *y* são independentes se, e somente se

$$p_{x,y}(x,y) = p_x(x)\,p_y(y) \tag{3.1}$$

onde $p_x(x)$ é a função densidade de probabilidade de x, $p_y(y)$ é a função densidade de probabilidade de y e $p_{xy}(x,y)$ é a função densidade de probabilidade conjunta.

Em palavras, a densidade conjunta deve ser fatorável no produto das densidades marginais. Assim, variáveis aleatórias independentes satisfazem a propriedade:

$$\begin{aligned} &E\{g(x)h(y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\,p_{x,y}(x,y)\,dxdy = \\ &\int_{-\infty}^{\infty} g(x)p_x(x)\,dx\int_{-\infty}^{\infty} h(y)p_y(y)\,dy = E\{g(x)\}\,E\{h(y)\} \end{aligned} \tag{3.2}$$

A equação 3.2 revela que a independência estatística é uma propriedade bem mais complexa que o não-correlacionamento. Observando a equação que define a condição de não-correlacionamento, $E\{xy\} = E\{x\}\,E\{y\}$, pode-se perceber que se trata de um caso particular de independência, em que ambas $g(x)$ e $h(y)$ são funções lineares, considerando-se apenas estatísticas até segunda ordem (correlações ou covariâncias). Apenas no caso gaussiano independência se reduz a não-correlacionamento, pois essa é

uma propriedade exclusiva dessa distribuição, uma vez que ela é completamente caracterizada apenas pelo primeiro e segundo momentos (média e variância). As definições de independência dadas acima se estendem de maneira natural a um número arbitrário de variáveis ou vetores aleatórios. Sejam $\vec{x},\vec{y},\ldots,\vec{z}$ vetores aleatórios. A condição de independência torna-se

$$p_{x,y,z,\ldots}(\vec{x},\vec{y},\vec{z},\ldots) = p_x(\vec{x})\, p_y(\vec{y}) p_z(\vec{z})\ldots \tag{3.3}$$

E a equação (3.2) generaliza-se para:

$$E\{g(\vec{x})h(\vec{y})k(\vec{z})\ldots\} = E\{g(\vec{x})\}E\{h(\vec{y})\}E\{k(\vec{z})\}\ldots \tag{3.4}$$

com $g(.),h(.),k(.)$ funções arbitrárias dos vetores $\vec{x},\vec{y},\vec{z}$.

3.2 Estatísticas de Ordens Superiores

Métodos padrões para processamento de sinais são baseados na utilização de informação estatística de primeira e segunda ordem. Essas teorias são bastante consistentes e bem desenvolvidas, se mostrando úteis em uma grande variedade de aplicações. Porém, quase sempre estão limitadas a hipóteses como gaussianidade, linearidade e estacionaridade.

Com o passar do tempo e o surgimento de novos paradigmas de aprendizado e redes neurais, o interesse em métodos estatísticos de ordens superiores começou a crescer na área de processamento de sinais.

A teoria ICA requer a utilização de estatísticas de ordens superiores, direta ou indiretamente, devido à necessidade de se obter não-linearidades e aproximações para funções e medidas de independência.

Dois conceitos importantes e freqüentemente utilizados em ICA são os *cumulants* e os momentos de variáveis aleatórias. Para a definição de *cumulant*, considere x uma variável aleatória contínua, real e escalar de média nula e com densidade de probabilidade $p_x(x)$. A primeira função característica $\varphi(w)$ é definida em (HYVÄRINEN; KARHUNEN; OJA, 2001) como a Transformada de Fourier de $p_x(x)$:

$$\varphi(w) = E\{\exp(jwx)\} = \int_{-\infty}^{\infty} p_x(x)\exp(jwx)dx \tag{3.5}$$

Pode ser verificado em (PAPOULIS, 1991) que toda distribuição de probabilidade é unicamente especificada por sua função característica, e vice-versa. A expansão de $\varphi(w)$ em série de Taylor fornece a expressão:

$$\varphi(w) = \int_{-\infty}^{\infty} p_x(x) \left[\sum_{k=0}^{\infty} x^k \frac{(jw)^k}{k!} \right] dx = \sum_{k=0}^{\infty} E\{x^k\} \frac{(jw)^k}{k!} \tag{3.6}$$

Portanto, os coeficientes da expansão são os momentos $E\{x^k\}$. Por esse motivo, a função característica $\varphi(w)$ também é conhecida como função geratriz de momentos. Analogamente, define-se a segunda função característica, ou função geratriz de *cumulants*, como o logaritmo natural da primeira função característica:

$$\phi(w) = \ln\left[\varphi(w)\right] = \ln\left\{E\left[\exp(jwx)\right]\right\} \tag{3.7}$$

Os *cumulants* k_k são definidos de maneira semelhante aos respectivos momentos, e são os coeficientes da expansão de Taylor de $\phi(w)$:

$$\phi(w) = \sum_{k=0}^{\infty} k_k \frac{(jw)^k}{k!} \tag{3.8}$$

onde o *k*-ésimo *cumulant* é obtido pela derivada:

$$k_k = (-j)^k \left. \frac{d^k \phi(w)}{dw^k} \right|_{w=0} \tag{3.9}$$

Existem algumas tabelas, encontradas na literatura (NIKIAS; PETROPULU, 1993), (ROSENBLATT, 1985) onde estão calculados os *cumulants* mais utilizados na prática. Por exemplo, para a variável x com média nula, os quatro primeiros *cumulants* são:

$$\begin{aligned} k_1 &= 0 \\ k_2 &= E\{x^2\} \\ k_3 &= E\{x^3\} \\ k_4 &= E\{x^4\} - 3E\{x^2\}^2 \end{aligned} \tag{3.10}$$

O quarto *cumulant*, k_4, é mais conhecido como curtose. Basicamente, é a medida mais simples de comportamento gaussiano de uma variável aleatória. Uma variável aleatória gaussiana possui curtose igual a zero. Em caso de valores negativos, a densidade da variável aleatória é chamada subgaussiana. Densidades subgaussianas são, em geral, multimodais. Por outro lado, se a curtose é positiva, a densidade, denominada supergaussiana, possui um pico mais agudo e alongado, como a densidade de Laplace.

No caso multivariado, $\vec{x}$ e $\vec{w}$ são vetores aleatórios e na equação (3.5) a integral é computada sobre todos os componentes de $\vec{x}$. Todas as expressões são análogas ao caso escalar. Nesse caso, os *cumulants* geralmente são denominados *cumulants* cruzados. Pode ser mostrado que os *cumulants* de segunda, terceira e quarta ordem para um vetor aleatório $\vec{x}$ são calculados por

$$\begin{aligned} &cum(x_i, x_j) = E\{x_i x_j\} \\ &cum(x_i, x_j, x_k) = E\{x_i x_j x_k\} \\ &cum(x_i, x_j, x_k, x_l) = E\{x_i x_j x_k x_l\} - E\{x_i x_j\} E\{x_k x_l\} - E\{x_i x_k\} E\{x_j x_l\} - E\{x_i x_l\} E\{x_j x_k\} \end{aligned} \tag{3.11}$$

Portanto, o *cumulant* de segunda ordem é igual ao segundo momento $E\{x_i x_j\}$, que é a correlação r_{ij} entre x_i e x_j. Analogamente, o *cumulant* de terceira ordem é igual ao terceiro momento. Porém, o *cumulant* de quarta ordem difere do quarto momento.

Em geral, percebe-se que momentos de ordens superiores possuem certa correspondência com as correlações utilizadas na estatística de segunda ordem e que os *cumulants*, por sua vez, correspondem às informações de ordens superiores relacionadas às covariâncias. Basicamente, momentos e *cumulants* contêm a mesma informação estatística, uma vez que *cumulants* podem ser expressos como somas de produtos de momentos. É preferível se trabalhar com *cumulants* porque eles conseguem representar de forma mais clara a informação adicional fornecida por estatística de ordens superiores, além de possuírem propriedades matemáticas únicas, não compartilhadas pelos momentos.

Uma das possíveis abordagens para ICA é uma generalização direta da técnica PCA. A transformação de Karhunen-Loéve é restrita a estatística de segunda ordem (matriz de covariância), com o objetivo de tornar os dados não correlacionados, ou seja, impõe a condição de que as covariâncias cruzadas $E\{x_i x_j\}$, para dados centralizados (média nula), sejam zero. Como em ICA se buscam componentes independentes, pode ser mostrado em (THEODORIDIS; KOUTROUMBAS, 2003) que essa condição é equivalente a impor que todos os *cumulants* cruzados de ordens superiores sejam nulos, como uma generalização da condição de não-correlacionamento. Em (COMON, 1994), é sugerido que aplicar tal restrição até *cumulants* de quarta ordem é suficiente para a maioria das aplicações. Existem diferentes métodos de estimação ICA que utilizam essa abordagem, dentre os quais pode-se citar as técnicas desenvolvidas em (CARDOSO, 1999) e (ZHANG; CHEN, 2004). Recentemente, métodos baseados na função

característica foram desenvolvidos. Em (CHEN; BICKEL, 2005) é proposto um algoritmo ICA, denominado CHFICA (*Characteristic Function Based ICA*) baseado no resultado de que a função característica $\varphi(\vec{w})$ pode ser fatorada no produto das marginais se e somente se $\vec{w}$ tem componentes mutuamente independentes. Assim, o algoritmo tenta encontrar uma matriz de transformação W tal que minimize a diferença entre a função característica e o produto das marginais.

3.3 Considerações Finais

Os conceitos sobre probabilidade e estatísticas introduzidos nesse capítulo fornecem uma base para o entendimento dos principais métodos ICA, sendo úteis para a compreensão das várias abordagens de estimação existentes. A característica principal dos métodos ICA é a incorporação de estatísticas de ordens superiores. Com a inclusão de tais métodos (informações estatísticas adicionais) no processo de extração de atributos, espera-se superar algumas limitações existentes nos métodos de segunda ordem, que serão discutidas no capítulo seguinte

4 MÉTODOS ESTATÍSTICOS DE SEGUNDA ORDEM

Este capítulo descreve de maneira breve os principais métodos lineares utilizados em extração de atributos: *Principal Component Analysis* (não-supervisionado) e *Linear Discriminant Analysis* (supervisionado). Tais métodos aplicam transformações geométricas no espaço de atributos, de modo a gerar novos atributos a partir de combinações lineares do conjunto de padrões originais com o intuito de reduzir a dimensionalidade. Pode ser verificado que um espaço de atributos de dimensionalidade reduzida pode melhorar a capacidade de generalização do classificador.

O objetivo é analisar os critérios adotados nos métodos e verificar que existem várias limitações em cada caso. Além disso, também são apresentados critérios para seleção de atributos baseados em distâncias probabilísticas e medidas de dispersão inter/intra classes. A utilização de métodos lineares pode ser motivada pelo uso de classificadores não lineares (i.e, bayesiano quadrático), pois assim o cálculo de funções não lineares é diminuído (custo computacional), especialmente em casos de dimensionalidade moderada e alta (hiperespectral), além de serem métodos mais simples e rápidos.

4.1 Principal Component Analysis

Principal Component Analysis é a técnica que implementa a Transformação de Karhunen-Loève, conhecida também como Transformação de Hotteling, um método clássico de análise estatística de dados multivariados que realiza a expansão de um vetor $\vec{x}$ em termos dos vetores próprios de sua matriz de covariância, bastante utilizado em áreas como extração de atributos e compressão de dados. Por essa razão, PCA é limitado a estatística de segunda ordem e nenhuma hipótese sobre densidades de probabilidade é necessária, uma vez que toda a informação necessária pode ser estimada diretamente das amostras. Dado um conjunto de dados multivariados, o objetivo é reduzir a dimensionalidade e a redundância existente nos dados para encontrar a melhor representação possível que minimize o erro médio quadrático. Em PCA, a redundância é medida por correlações entre os elementos dos dados, enquanto que em ICA o conceito mais amplo de independência é utilizado.

Este capítulo apresenta uma breve discussão sobre critérios adotados para redução de dimensionalidade em PCA e algumas limitações existentes no método, do ponto de vista de separabilidade dos dados.

4.1.1 Interpretação Geométrica

O requisito básico em PCA é a existência de um vetor aleatório $\vec{x}$ com n elementos. Considera-se $\vec{x}$ como um vetor coluna. Devem estar disponíveis amostras $x_1,\ldots,x_m$ desse vetor. Nenhuma hipótese sobre a densidade de probabilidade é feita, pois a estatística de primeira e segunda ordem pode ser estimada diretamente através das amostras. Além disso, nenhum modelo generativo é assumido para o vetor $\vec{x}$.

Tipicamente, em processamento de imagens, os elementos do vetor $\vec{x}$ são medidas como os níveis de cinza de cada pixel. Em PCA é necessário que os elementos do vetor sejam mutuamente correlacionados, ou seja, exista certa redundância, para que a compressão seja possível. Se os elementos forem não-correlacionados, nada pode ser feito com PCA.

Na transformação PCA, os dados são primeiramente centralizados subtraindo-lhes a média, que na prática é estimada através das amostras disponíveis. Em seguida, $\vec{x}$ é transformado linearmente em um outro vetor $\vec{y}$ contendo m elementos, com $m < n$, de maneira que a redundância introduzida pela correlação é eliminada. Geometricamente, tal condição é obtida através de uma rotação do sistema de coordenadas ortogonal, de modo que os componentes de $\vec{x}$ no novo sistema de coordenadas sejam não-correlacionados. Simultaneamente, as variâncias das projeções de $\vec{x}$ nos novos eixos são maximizadas, sendo que o primeiro eixo corresponde a maior variância, o segundo eixo corresponde a maior variância na direção ortogonal ao primeiro, e assim por diante, conforme a dimensionalidade dos dados.

No caso gaussiano multivariado, o espaço de coordenadas rotacionado corresponde ao espaço definido pelos eixos principais do hiperelipsóide que define a distribuição. A Figura 9, obtida em (FUKUNAGA, 1990), mostra um exemplo bidimensional. Os componentes principais são agora as projeções dos dados nos dois eixos principais, respectivamente ϕ_1 e ϕ_2. Além disso, as variâncias dos componentes, representadas pelos valores próprios, λ_i, são distintas na maioria das aplicações, sendo que um número considerável delas é tão pequeno que os componentes correspondentes podem ser descartados. Os componentes principais selecionados constituem o vetor $\vec{y}$. O

objetivo é encontrar quais são os vetores da nova base, através da otimização de alguns critérios.

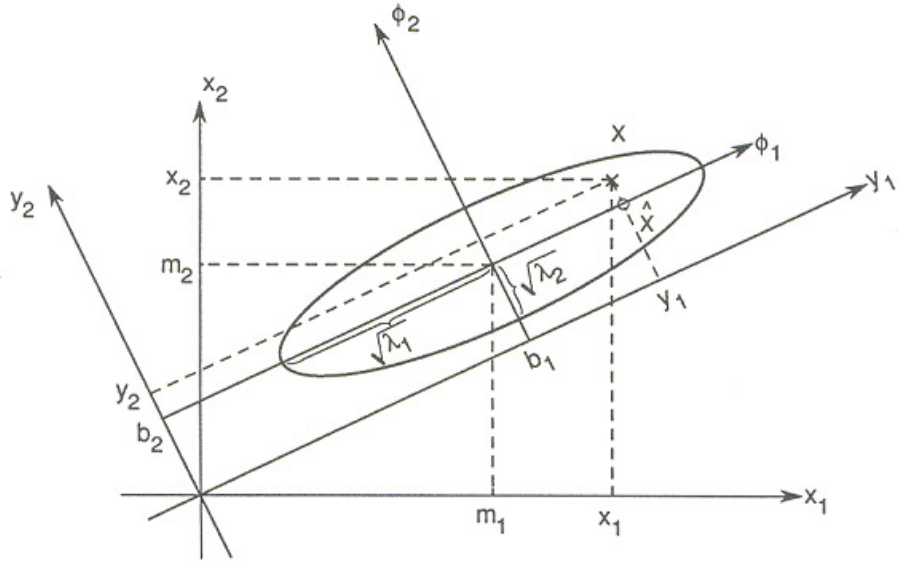


Figura 9*. Ilustração gráfica da Transformação de Karhunen-Loève para caso gaussiano bidimensional.*

4.1.2 PCA pela Maximização da Variância

Pode ser mostrado em (FUKUNAGA, 1990), que se λ_j e $\vec{u}_j$ são, respectivamente, o *j*-ésimo valor próprio e vetor próprio da matriz de covariância de $\vec{x}$, então

$$\begin{aligned} \lambda_j &\geq 0 \\ \vec{u}_j \cdot \vec{u}_k &= 0, \quad \text{para } j \neq k \end{aligned} \tag{4.1}$$

Ou seja, todos os valores próprios são positivos e os vetores próprios são mutuamente ortogonais entre si. Como conseqüência, para uma matriz de rank *n*, tem-se *n* vetores próprios ortonormais, assumindo que $\left\|\vec{u}_j\right\| = 1$ para $j = 1,...,n$, associados aos valores próprios $\lambda_1, \lambda_2, ..., \lambda_n$. Matematicamente, pode-se expressar a rotação do sistema de coordenadas definida pela Transformação de Karhunen-Loève como uma matriz ortonormal $Z = \left[T^T, S^T\right]$, de dimensões $n \times n$, com $T^T = \left[\vec{w}_1, \vec{w}_2, ..., \vec{w}_m\right]_{N \times M}$ representando os eixos do novo sistema de coordenadas, e $S^T = \left[\vec{w}_{m+1}, \vec{w}_{m+2}, ..., \vec{w}_{n,}\right]_{N \times (N-M)}$ como os eixos referentes as componentes eliminadas durante a redução da dimensionalidade. A condição de ortonormalidade implica que $\vec{w}_j \cdot \vec{w}_k = 0$ para $j \neq k$, e $\vec{w}_j \cdot \vec{w}_k = 1$ para $j = k$.

Pode-se escrever o vetor *n*-dimensional $\vec{x}$ através de sua expansão nos vetores da base:

$$\vec{x} = \sum_{j=1}^{n} \left(\vec{x}^T \vec{w}_j \right) \vec{w}_j = \sum_{j=1}^{n} c_j \vec{w}_j \tag{4.2}$$

onde c_j é o produto interno entre $\vec{x}$ e $\vec{w}_j$.

Então, o novo vetor *m*-dimensional $\vec{y}$ é obtido pela transformação:

$$\vec{y}^T = \vec{x}^T T^T = \sum_{j=1}^{n} c_j \vec{w}_j^T \left[\vec{w}_1, \vec{w}_2, ..., \vec{w}_m \right] = \left[c_1, c_2, ..., c_m \right] \tag{4.3}$$

Desta forma, busca-se uma transformação T que maximize a variância dos dados, ou seja, otimize o critério PCA a seguir, com C_X sendo a matriz de covariância do vetor centralizado $\vec{x}$:

$$J_1^{PCA}\left(\vec{w}_j \right) = E\left[\left\| \vec{y} \right\|^2 \right] = E\left[\vec{y}^T \vec{y} \right] = \sum_{j=1}^{m} E\left[c_j^2 \right] \tag{4.4}$$

Porém, sabe-se que $c_j = \vec{x}^T \vec{w}_j$, e portanto:

$$J_1^{PCA}\left(\vec{w}_j \right) = \sum_{j=1}^{m} E\left[\vec{w}_j^T \vec{x} \vec{x}^T \vec{w}_j \right] = \sum_{j=1}^{m} \vec{w}_j^T E\left[\vec{x} \vec{x}^T \right] \vec{w}_j = \sum_{j=1}^{m} \vec{w}_j^T C_X \vec{w}_j \tag{4.5}$$

sujeito a restrição $\left\| \vec{w}_j \right\| = 1$.

Trata-se de um problema de otimização com restrição de igualdade. É conhecido que a solução é encontrada através de multiplicadores de Lagrange. Nesse caso, tem-se:

$$J_1^{PCA}\left(\vec{w}_j, \gamma_j \right) = \sum_{j=1}^{m} \vec{w}_j^T C_X \vec{w}_j - \sum_{j=1}^{m} \gamma_j \left(\vec{w}_j^T \vec{w}_j - 1 \right) \tag{4.6}$$

Derivando a expressão acima em relação a cada componente de $\vec{w}_j$ e igualando a zero, chega-se ao seguinte resultado, encontrado em (YOUNG; CALVERT, 1974):

$$C_X \vec{w}_j = \lambda_j \vec{w}_j \tag{4.7}$$

Portanto, tem-se um problema de vetor próprio, ou seja, os vetores $\vec{w}_j$ da nova base que maximizam a variância dos dados transformados são os vetores próprios da matriz de covariância C_X. Porém, informações a respeito de como os *m* vetores próprios devem ser selecionados, tornam-se mais claras na abordagem apresentada a seguir.

É importante notar que após essa transformação os dados se encontram descorrelacionados, ou seja, a matriz de covariância C_Y, onde $\vec{y}$ é o resultado da aplicação da transformação T em $\vec{x}$, é diagonal (pela decomposição em valores próprios (EVD) da matriz C_X):

$$C_Y = T^T C_X T = T^T T D T^T T = D \tag{4.8}$$

onde $D = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ é a matriz diagonal dos valores próprios de C_X .

4.1.3 PCA pela Minimização do Erro Médio Quadrático

Uma outra possível abordagem para o problema PCA é obtida através da minimização do erro médio quadrático durante a redução de dimensionalidade do vetor $\vec{x}$. Nessa visão, busca-se por um conjunto de *m* vetores de base, ortonormais, (*m* < *n*), que gerem um subespaço *m*-dimensional tal que o erro médio quadrático entre o vetor original $\vec{x}$ e sua projeção nesse subespaço seja mínima. Denotando os vetores da base por $\vec{w}_1, \vec{w}_2, ..., \vec{w}_m$, pela condição de ortonormalidade tem-se:

$$\vec{w}_i^T \vec{w}_j = \delta_{ij} \quad , \text{ onde } \delta_{ij} = 1 \text{ se } i = j \qquad \delta_{ij} = 0 \text{ se } i \neq j \tag{4.9}$$

A projeção de $\vec{x}$ no subespaço gerado pelos vetores $\vec{w}_j$, com $j = 1, ..., m$, é dada pela equação (4.2) e, portanto o critério de erro médio quadrático a ser minimizado torna-se:

$$J_{MSE}^{PCA}\left(\vec{w}_j\right) = E\left[\left\|\vec{x} - \sum_{j=1}^{m}\left(\vec{x}^T \vec{w}_j\right)\vec{w}_j\right\|^2\right] \tag{4.10}$$

Devido às propriedades de ortonormalidade e considerando o vetor média nulo, esse critério pode ser simplificado para:

$$\begin{aligned} J_{MSE}^{PCA}\left(\vec{w}_j\right) &= E\left[\left\|\vec{x}\right\|^2\right] - E\left[\sum_{j=1}^{m}\left(\vec{x}^T \vec{w}_j\right)^2\right] = \\ E\left[\left\|\vec{x}\right\|^2\right] - \sum_{j=1}^{m} E\left[\vec{w}_j^T \vec{x}\vec{x}^T \vec{w}_j\right] &= E\left[\left\|\vec{x}\right\|^2\right] - \sum_{j=1}^{m} \vec{w}_j^T C_X \vec{w}_j \end{aligned} \tag{4.11}$$

Como o primeiro termo não depende de $\vec{w}_j$, para minimizar o critério MSE, basta maximizar $\sum_{j=1}^{m} \vec{w}_j^T C_X \vec{w}_j$. Porém, da equação (4.5) na seção anterior, esse mesmo problema de otimização foi resolvido através de multiplicadores de Lagrange, e o resultado obtido é que os vetores $\vec{w}_j$ devem ser os vetores próprios de C_X . Então, substituindo-se a equação (4.7) em (4.11), tem-se:

$$J_{MSE}^{PCA}\left(\vec{w}_j\right) = E\left[\left\|x\right\|^2\right] - \sum_{j=1}^{m} \gamma_j \tag{4.12}$$

Esse resultado mostra que para minimizar o erro médio quadrático, deve-se escolher os *m* vetores próprios associados aos m maiores valores próprios da matriz de

covariância. Pode ser mostrado em (FUKUNAGA, 1990) que o valor do mínimo erro médio quadrático é:

$$J_{MSE}^{PCA}\left(\vec{w}_j\right) = \sum_{j=m+1}^{n} \gamma_j \tag{4.13}$$

4.1.4 Limitações da Transformação de Karhunen-Loève

Os critérios adotados na transformação de Karhunen-Loève, apesar de bastante efetivos em termos de compactação dos dados e redução de dimensionalidade, não são necessariamente adequados a alguns tipos de aplicações. Geralmente, quando existe a necessidade de separabilidade dos dados em classes, o PCA não garante que a projeção dos dados no subespaço maximize a separação, pois a redução de dimensionalidade não é desenvolvida para essa finalidade, mas para otimizar a representação dos dados da mistura em termos do erro médio quadrático. Essa situação pode ser visualizada na Figura 10, adaptada de (THEODORIDIS; KOUTROUMBAS, 2003). Os vetores de padrões das duas classes seguem uma distribuição gaussiana com mesma matriz de covariância. As elipses indicam as curvas de nível. Os vetores próprios da matriz são $\vec{a}_1$, associado ao maior valor próprio, e $\vec{a}_2$ associado ao menor valor próprio. A projeção dos dados em $\vec{a}_1$ é desfavorável em termos de separabilidade entre as classes, enquanto que em $\vec{a}_2$, a separação é máxima. Tal fato é uma motivação a mais para se incorporar métodos estatísticos de ordens superiores durante a etapa de extração de atributos.

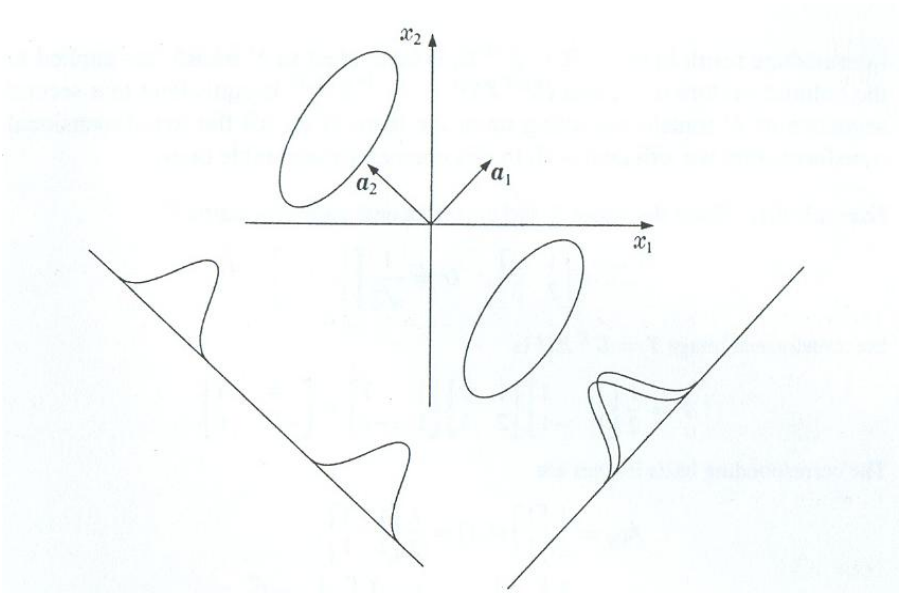


Figura 10*. Limitação da transformação de Karhunen-Loève, do ponto de vista de separabilidade entre as classes.*

4.2 Linear Discriminant Analysis

Linear Discriminant Analysis é uma generalização da Função Discriminante Linear de Fischer. Trata-se de um mapeamento supervisionado que encontra vetores em que a projeção dos dados maximiza um critério de separabilidade entre as classes. Embora sejam produzidos atributos ótimos do ponto de vista de classificação, esse método apresenta diversas limitações. O critério utilizado por Fischer para o problema de classificação em duas classes encontra a projeção que maximiza a distância entre as médias das classes e minimiza o uma medida de espalhamento dos dados:

$$J(\vec{w}) = \frac{\left|m_1^{'} - m_2^{'}\right|}{\left(s_1^2 + s_2^2\right)} \tag{4.14}$$

onde os escalares $m_i^{'}$ e s_i^2 são, respectivamente, a média das projeções da classe w_i e o "espalhamento" (variância) das projeções da classe w_i.

É possível utilizar um critério equivalente, obtido reescrevendo-se a equação (4.14) em função de um vetor $\vec{w}$ e generalizá-lo para *m* projeções:

$$J_{LDA}(W) = \frac{W^T S_B W}{W^T S_W W} \tag{4.15}$$

onde S_B denota a matriz de espalhamento entre classes e S_W a matriz de espalhamento intra classe.

Portanto, esse método encontra uma matriz $d \times m$ contendo *m* vetores de projeção ótimos (colunas da matriz ortonormal) $W_{LDA}^T = [\vec{w}_1, \vec{w}_2, \ldots \vec{w}_m]$ que determinam a projeção $\vec{y}^T = \vec{x}^T W_{LDA}^T$ e maximiza a relação entre a matriz de espalhamento inter-classes S_B e a matriz de espalhamento intra-classe S_W. Considerando um problema de c classes com *N* amostras, essas matrizes são definidas como:

$$S_B = \sum_{i=1}^{c} N_i \left(\vec{\mu}_i - \vec{\mu}\right)\left(\vec{\mu}_i - \vec{\mu}\right)^T \tag{4.16}$$

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in w_i} \left(\vec{x}_k - \vec{\mu}_i\right)\left(\vec{x}_k - \vec{\mu}_i\right)^T = \sum_{i=1}^{c} S_{W_i} \tag{4.17}$$

onde $\vec{\mu}$ é a média de todas as amostras (global), $\vec{\mu}_i$ é a média da classe w_i, S_{W_i} é a matriz de covariância da classe w_i e N_i é o número de amostras na classe w_i.

Pode ser mostrado que a maximização da equação (4.15) conduz à definição de um problema de vetor próprio generalizado, onde os vetores colunas da matriz

$W_{LDA}^{T} = \left[\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_m\right]$ correspondem aos m vetores próprios associados aos m maiores valores próprios da matriz definida por $\left(S_W^{-1} S_B\right)$, ou seja:

$$\left(S_W^{-1} S_B\right)\vec{w}_j = \lambda_j \vec{w}_j \ , \ j = 1,...,m \ , \ \lambda_1 \geq ... \geq \lambda_m \tag{4.18}$$

Porém, o *rank* da matriz de espalhamento intra-classes S_B é, no máximo, $c-1$, onde c representa o número de classes, porque essa matriz é definida como a soma de c matrizes de *rank* um ou menos, o que segundo (DUDA; HART; STORK, 2001) implica dizer que podem existir apenas $c-1$ valores próprios não nulos, e portanto, na existência da severa restrição de que o número de atributos obtidos d deva ser igual ou inferior ao número de classes do problema, ou seja, $d \leq c-1$. Outro problema dessa abordagem, discutido em (YANG; AHUJA, 2001), é relacionado a dados com dimensionalidade muito alta, onde para um conjunto contendo N amostras de treinamento com d atributos, em geral, quando N é menor que d a matriz de espalhamento entre classes, $S_W \in \Re^{n \times n}$, é singular, ou seja, não admite inversa.

4.3 Critérios para seleção de atributos

Segundo (DEVIJVER; KITTLER, 1982), existem basicamente dois tipos de critérios utilizados para seleção de atributos: aqueles baseados em medidas de distância entre classes, que utilizam o conceito de matrizes de espalhamento e aqueles baseados em medidas probabilísticas de separabilidades, que podem ser associados ao limite superior do erro na classificação bayesiana. A primeira classe de critérios fornece algoritmos simples e eficientes, uma vez que mede a separabilidade entre c classes simultaneamente, mas não tem relação com o erro de Bayes. Já no segundo caso, trata-se de medidas entre duas classes apenas, podendo ser generalizado para várias classes. Porém, na prática, sua utilização se baseia em hipótese de classes gaussianas.

Todas as matrizes de espalhamento utilizadas como medidas de distância entre classes, S_W, S_B e $S_T = S_W + S_B$, são invariantes a deslocamentos de coordenadas (i.e, centralização dos dados). Para usar essa abordagem é necessário converter essas matrizes em um número que seja grande quando os elementos da matriz de espalhamento entre-classes forem altos ou quando os elementos da matriz de espalhamento intra-classes forem pequenos. Como exemplo, (FUKUNAGA, 1990) considera os seguintes critérios:

$$J_1 = tr\left(S_B^{-1} S_W\right) \tag{4.19}$$

$$J_2 = \ln\left|S_B^{-1} S_W\right| = \ln\left|S_W\right| - \ln\left|S_B\right| \tag{4.20}$$

$$J_3 = tr\left(S_W\right) - \mu\left[tr\left(S_B\right) - c\right] \tag{4.21}$$

As medidas probabilísticas também são bastante úteis na prática, pois o erro de classificação é o critério ideal para a avaliação de um conjunto de atributos. Algumas das medidas mais importantes são a distância de Bhattacharrya $\left(J_B\right)$ e a distância de Jeffreys-Matusita ou distância J-M $\left(J_M\right)$, baseadas no limite de Chernoff, que define o limite superior do erro de Bayes.

$$J_B = -\ln \int \sqrt{p\left(\vec{x}\middle|w_i\right) p\left(\vec{x}\middle|w_j\right)} dx \tag{4.22}$$

$$J_M = \left\{ \int \left[\sqrt{p\left(\vec{x}\middle|w_i\right)} - \sqrt{p\left(\vec{x}\middle|w_j\right)} \right]^2 dx \right\}^{1/2} \tag{4.23}$$

onde $p\left(\vec{x}\middle|w_i\right)$ é a função densidade de probabilidade condicional da classe w_i, $0 \le J_B < \infty$ e $0 \le J_M \le \sqrt{2}$.

Um aspecto importante dos critérios probabilísticos é que quando as funções densidade de probabilidade condicionais pertencem a uma família específica (i.e, normal), e as expressões podem ser simplificadas como mostra (DEVIJVER; KITTLER, 1982):

$$J_B = \frac{1}{8}\left(\vec{\mu}_i - \vec{\mu}_j\right)^T \left[\frac{\Sigma_i + \Sigma_j}{2}\right]^{-1} \left(\vec{\mu}_i - \vec{\mu}_j\right) + \frac{1}{2}\ln\left[\frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{\left(\left|\Sigma_i\right|\left|\Sigma_j\right|\right)}}\right] \tag{4.24}$$

$$J_M = \left\{2\left[1 - e^{-J_B}\right]\right\}^{1/2} \tag{4.25}$$

onde $\vec{\mu}_i$ é o vetor media para a classe w_i e Σ_i é a matriz de covariância da classe w_i.

Para problemas de c classes, a idéia é selecionar o melhor subconjunto de atributos possíveis (i.e, componentes independentes) de acordo com a distância J-M média ou maximizando a mínima distância entre classes:

$$J_{M_{AVG}} = \sum_{i=1}^{c} \sum_{j=1}^{c} p\left(w_i\right) p\left(w_j\right) J_{M_{ij}} \tag{4.26}$$

$$\max_{\{M\}} J_{M_{\min}} = \max_{\{M\}} \min_{\{i,j\}} J_{M_{ij}} \tag{4.27}$$

4.4 Considerações Finais

Os métodos estatísticos de segunda ordem para extração de atributos possuem limitações que podem afetar negativamente o desempenho da classificação. O PCA é um método sub-ótimo do ponto de vista de classificação (ótimo do ponto de vista de representação), enquanto o LDA possui sérias restrições quanto ao número de atributos, ou seja, o número de atributos está condicionado ao número de classes do problema. No próximo capítulo serão discutidas algumas abordagens para algoritmos ICA, visando a definição de um esquema proposto para combinação de métodos de extração (fusão de atributos) para tentar superar essas dificuldades.

5 *INDEPENDENT COMPONENT ANALYSIS*

Nesse capítulo são apresentados os conceitos básicos de *Independent Component Analysis*, através da formulação matemática do modelo ICA. Além disso, são definidos os princípios fundamentais que regem a estimação e as condições necessárias para que isso seja possível, bem como algumas considerações sobre restrições existentes no modelo teórico.

5.1 Modelo Matemático

Independent Component Analysis pode ser rigorosamente definida como um modelo estatístico de variáveis latentes, no qual as observações $x_1, x_2, ..., x_n$ são modeladas como combinações lineares das variáveis $s_1, s_2, ..., s_n$, ou seja:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + ... + a_{in}s_n \tag{5.1}$$

para todo $i = 1, ..., n$ e com a_{ij} coeficientes reais. Por definição, as variáveis s_i são estatisticamente independentes. O modelo ICA é generativo, o que significa que ele descreve como os dados observados são gerados por um processo de combinação dos componentes s_j. Os componentes independentes s_j (ICs) são variáveis latentes, pois não são diretamente observadas. A matriz de mistura composta pelos coeficientes a_{ij} também é desconhecida. Toda a observação se restringe exclusivamente às variáveis aleatórias x_i, sendo que devem ser estimados tanto a matriz de mistura quanto os componentes independentes. Usualmente, é mais conveniente se utilizar a notação vetorial ao invés de somatórios. Seja $\vec{x}$ o vetor aleatório cujos elementos são as observações $x_1, x_2, ..., x_n$ e $\vec{s}$ o vetor aleatório com os elementos $s_1, s_2, ..., s_n$. Considere ainda A como a matriz que contém os coeficientes a_{ij}. O modelo ICA pode ser escrito da seguinte maneira:

$$\vec{x} = A\vec{s} \tag{5.2}$$

Detalhes matemáticos mais profundos sobre condições necessárias e suficientes para garantir a identificabilidade (estimação da matriz de mistura), a separabilidade (recuperação dos componentes independentes) e a unicidade (em casos subdeterminados) de modelos ICA são definidos formalmente em (ERIKSSON; KOIVUNEN, 2004).

5.1.1 Restrições do Modelo ICA

De acordo com (HYVÄRINEN; KARHUNEN; OJA, 2001), a definição teórica do modelo ICA possui algumas restrições e hipóteses para que a estimação dos componentes independentes seja possível. Algumas delas são consideradas apenas para a simplificação das definições.

1. **Os componentes s_j são estatisticamente independentes.**
2. **Os componentes independentes devem possuir distribuição não-gaussiana.**
   O motivo para essa restrição vem do fato de que na distribuição gaussiana os *cumulants* de ordem superiores são nulos. Porém, essas informações referentes a estatística de altas ordens são essenciais para a estimação ICA, o que impossibilita o caso gaussiano. Embora se tenha conhecimento sobre o comportamento não-gaussiano, a distribuição dos componentes ainda é completamente desconhecida. Caso contrário, o problema seria bastante simplificado.
3. **Para simplificação, em teoria, assume-se que a matriz de mistura A é quadrada.**
   Essa restrição não é obrigatória, ou seja, é possível que o número de componentes independentes seja menor que o número de observações. A utilidade dessa restrição é facilitar a estimação e a definição do modelo teórico. Dessa forma, é necessário estimar a matriz A, encontrar sua inversa W e multiplicar diretamente por $\vec{x}$ para obter $\vec{s}$, ou seja, $\vec{s} = W\vec{x}$. Porém, essa hipótese pode ser generalizada em casos específicos.
4. **Não se pode determinar a ordem dos componentes independentes.**
   Esse fato também decorre do desconhecimento de ambos $\vec{s}$ e A. Formalmente, é como se existisse uma matriz de permutação P e sua inversa no modelo dado pela equação (5.2), ou seja, $\vec{x} = AP^{-1}P\vec{s}$. Nesse caso, $P\vec{s}$ seriam os componentes originais, mas em outra ordem, e AP^{-1} seria apenas uma nova matriz de mistura a ser estimada. Na prática, significa que os componentes independentes estimados não seguem um tipo de ordenação como acontece em PCA.

5.1.2 Modelo ICA Estendido

Em muitos casos, as observações são vetores aleatórios, como por exemplo, no caso de vários sinais temporais ou imagens multiespectrais. Para esses casos, denota-se por X a matriz que contém em suas colunas os vetores observados $x_1, x_2, ..., x_n$ e

similarmente para a matriz S. Então, o modelo ICA pode ser representado por:

$$X = AS \tag{5.3}$$

A figura 11 ilustra esta situação. Observando a figura, pode-se verificar, por exemplo, que o primeiro elemento do primeiro sinal observado, o componente um, é definido como uma combinação linear dos primeiros componentes de cada um dos sinais independentes, ou seja, da primeira coluna da matriz S. Porém, como citado anteriormente, na prática toda observação se restringe a matriz X, sendo necessária a estimação tanto de A quanto de S.

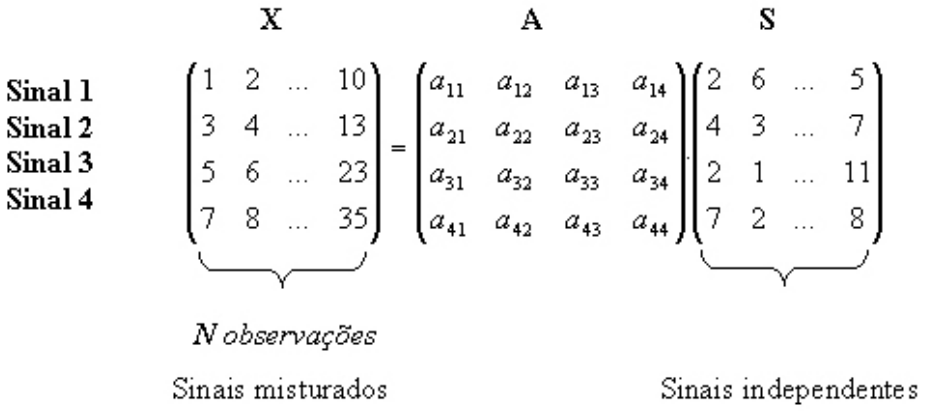


Figura 11*. Esquema para modelo ICA estendido.*

5.2 Princípios da estimação ICA

O objetivo deste tópico é descrever as principais abordagens utilizadas na estimação ICA, de modo a oferecer uma visão geral de cada uma delas e verificar a equivalência existente entre as diversas abordagens.

Existem vários métodos para estimação ICA, sendo que suas propriedades estatísticas dependem basicamente da escolha dos critérios a serem otimizados. Essas variações diferem principalmente em questões como a estabilidade do método, a velocidade de convergência, e requisitos de memória necessários para sua execução (custo computacional). As principais abordagens para a estimação ICA existentes na literatura são: a maximização da não-gaussianidade (FastICA) (HYVÄRINEN, 1999), a estimação por máxima verossimilhança (infomax, ICAML) (BELL; SEJNOWSKI, 1995), (MACKAY, 1999), (HANSEN; LARSEN; KOLENDA, 2001), a minimização da informação mútua (AMARI; CICHOCKI; YANG, 1996), (THEODORIDIS; KOUTROUMBAS, 2003), (ROBILA; VARSHNEY, 2002a), (DINH-TUAN PHAM, 2004), métodos para PCA não linear (OJA, 1997), e métodos baseados em *cumulants*,

ou *cumulant-based methods*, (CARDOSO,1997), (CARDOSO, 1999), (ZHANG; CHEN, 2004), (BLASCHKE; WISKOTT, 2004).

5.2.1 ICA pela maximização da não-gaussianidade

Uma das possíveis abordagens para a estimação ICA consiste na maximização de uma medida de comportamento não gaussiano da distribuição. Essa abordagem é motivada por resultados do Teorema Central do Limite e constitui a base para o desenvolvimento do algoritmo FastICA. Os conceitos fundamentais utilizados dessa abordagem foram desenvolvidos em (HYVÄRINEN, 1999) e (HYVÄRINEN; KARHUNEN; OJA, 2001) e tomam como base a utilização do comportamento não-gaussiano como medida de independência. Em resumo, a idéia é que a soma de variáveis aleatórias independentes é mais gaussiana que cada uma delas (a densidade de probabilidade da soma f_{x+y} é igual à convolução das densidades originais f_x e f_x).

5.2.1.1 Medidas de não-gaussianidade

Para utilizar a não-gaussianidade na estimação ICA, deve-se obter uma medida quantitativa desse comportamento para variáveis aleatórias. Como visto anteriormente, uma medida clássica, conhecida como curtose, ou *cumulant* de quarta ordem, é bastante empregada para essa finalidade. A curtose possui como principal vantagem a fácil tratabilidade e ser computacionalmente simples, pois é considerada uma versão normalizada do quarto momento, que pode ser estimado diretamente dos dados amostrais. Além disso, a curtose possui propriedades matemáticas únicas, simplificando as análises teóricas. Porém, a curtose tem sérios problemas na prática, sendo o principal deles a de não ser uma medida robusta, uma vez que é extremamente sensível a *outliers*, ou seja, poucos pontos isolados da distribuição afetam demais o resultado, prejudicando a estimação.

Portanto, uma outra medida de gaussianidade, denominada entropia relativa (*negentropy*), baseada em conceitos da Teoria da Informação, é utilizada. Um outro termo geral para essa medida é entropia relativa. Ela possui propriedades opostas à curtose, pois é robusta, mas computacionalmente complexa. Existe uma forte conexão entre ICA e Teoria da Informação, principalmente devido à definição dos critérios de otimização utilizados na estimação ICA. Existem boas aproximações para a entropia relativa que são muito úteis no desenvolvimento de métodos ICA e são discutidas a seguir.

5.2.1.2 *Negentropy* como medida de comportamento não-gaussiano

A *negentropy* é uma medida utilizada na Teoria da Informação e baseada na definição de entropia diferencial. A entropia de uma variável aleatória está relacionada com a informação fornecida pela sua observação. Quanto mais imprevisível, maior o valor da entropia. A entropia diferencial H de um vetor aleatório $\vec{y}$ com densidade de probabilidade $p_y(y)$ é definida como:

$$H(\vec{y}) = -\int p_y(\vec{y}) \log p_y(\vec{y})\, d\vec{y} \tag{5.4}$$

Pode ser mostrado, em teoria da informação, que uma variável aleatória gaussiana possui a maior entropia dentre todas as outras variáveis aleatórias de igual variância, o que caracteriza a entropia como uma medida de comportamento não-gaussiano. Uma propriedade interessante diz respeito a entropia de uma transformação linear. Se $\vec{z} = M\vec{y}$, a entropia de $\vec{z}$ é:

$$H(\vec{z}) = H(\vec{y}) + \log\left|\det M\right| \tag{5.5}$$

Em geral, a entropia é pequena para distribuições em que os dados são altamente agrupados ao redor de certos valores. Para se obter uma medida de gaussianidade que é zero para variáveis gaussianas e sempre não negativa caso contrário, define-se uma versão normalizada da entropia diferencial, denominada *negentropy*:

$$J(\vec{y}) = H(\vec{y}_{gauss}) - H(\vec{y}) \tag{5.6}$$

onde $\vec{y}_{gauss}$ é uma variável aleatória gaussiana com mesma matriz de covariância de $\vec{y}$. Sua entropia pode ser calculada através de algumas manipulações algébricas, sabendo-se que $H(\vec{y}) = -E\left[\log p_y(\vec{y})\right]$:

$$H(y_{gauss}) = \frac{1}{2}\log\left|\det \Sigma\right| + \frac{n}{2}(\log 2\pi) + \frac{n}{2} \tag{5.7}$$

sendo que n é a dimensão e S a matriz de covariância de $\vec{y}$.

A vantagem de se utilizar a *negentropy* é sua forte consolidação na teoria estatística, sendo considerada como o estimador ótimo para medida de não-gaussianidade. Entretanto, ela é computacionalmente complexa, e aproximá-la usando a definição dada acima requer a estimação (possivelmente não paramétrica) da densidade de probabilidade.

5.2.1.3 Aproximações para *Negentropy*

O método clássico de aproximação para a *negentropy* é através dos *cumulants* de

ordens superiores, utilizando expansões em funções polinomiais baseadas em idéias similares à expansão de Taylor. Em geral, duas expansões são usualmente utilizadas nessa situação: a expansão de *Gram-Charlier* e a expansão de *Edgeworth*. Tais expansões utilizam os chamados polinômios de *Chebyshev-Hermite*, H_i, definidos como a *i*-ésima derivada da densidade gaussiana padrão, com os respectivos *cumulants* de ordem i (k_i) sendo os coeficientes da expansão. Pode-se mostrar que esses polinômios formam um sistema ortonormal, uma propriedade desejável em expansão de funções em séries. Após cálculos algébricos e simplificações, chega-se ao seguinte resultado:

$$J(y) = \frac{1}{12} E\left[y^3\right]^2 + \frac{1}{48} kurt(y)^2 \tag{5.8}$$

com $kurt(y)$ sendo a curtose, ou *cumulant* de quarta ordem de y. Porém, essa aproximação ainda sofre de não-robustez, uma vez que o segundo termo da expressão indica a presença da curtose. Assim, uma aproximação mais sofisticada, através de funções não polinomiais, e baseada no princípio da máxima entropia é utilizada. Basicamente, o princípio da máxima entropia postula o seguinte: suponha que o todo conhecimento que se tem sobre uma densidade p se encontra disponível na forma da estimação de valores esperados do tipo $E\left[F_i(x)\right]$, onde F_i são m funções diferentes de x.

Então, a questão que surge é qual seria a função densidade de probabilidade que satisfaz as condições acima e possui a máxima entropia dentre todas. Esse questionamento é motivado notando-se que um número finito de observações não diz exatamente como é p. Nesse caso, a entropia pode ser considerada como uma medida de regularização para encontrar a densidade menos estruturada compatível com o conhecimento estimado. Em outras palavras, o princípio da máxima entropia pode ser interpretado como o mecanismo de se obter a densidade p compatível com as medidas estimadas e que faz o menor número possível de suposições sobre os dados, pois a entropia pode ser interpretada como uma medida de aleatoriedade, e portanto, a densidade obtida é a mais aleatória possível que satisfaz as observações. Maiores detalhes sobre a utilização da entropia como regularização são encontrados em (PAPOULIS, 1991) e (COVER; THOMAS, 1991). Essa abordagem generaliza a aproximação por *cumulants* de ordens superiores para que sejam utilizados valores

esperados de funções não-quadráticas, ou seja, momentos não polinomiais. Em geral, pode-se substituir as funções polinomiais y^3 e y^4 por quaisquer outras funções G_i. O método fornece então, um modo de aproximar a *negentropy* baseado no cálculo dos valores esperados $E\left[G_i\left(x\right)\right]$. Detalhes matemáticos mais profundos sobre o desenvolvimento das aproximações usando funções polinomiais e não polinomiais podem ser encontrados em (HAYKIN, 1998) e (HYVÄRINEN; KARHUNEN; OJA, 2001):

$$J\left(y\right) \approx \left\{E\left[G\left(y\right)\right] - E\left[G\left(v\right)\right]\right\}^2 \tag{5.9}$$

em que y é assumida ter média zero e variância unitária, e v é uma variável aleatória gaussiana normalizada, ou seja, com média zero e variância unitária. Dessa forma, tem-se uma medida de comportamento não-gaussiano consistente, pois é sempre não negativa e com valor mínimo igual a zero, somente quando y é uma variável gaussiana. A escolha da função G é importante, pois é possível conseguir melhores estimativas, sendo que diversas indicações podem ser obtidas na literatura.

5.2.1.4 FastICA: Um algoritmo iterativo

O objetivo do algoritmo FastICA é encontrar uma direção, ou seja, um vetor unitário $\vec{w}$, tal que a projeção dos dados nessa direção, $\vec{w}^T\vec{z}$, maximize a não-gaussianidade, medida pela função objetivo $J\left(\vec{w}^T\vec{z}\right)$, que é a aproximação da *negentropy* obtida anteriormente. Segundo (LEE; GIROLAMI; SEJNOWSKI, 1999), maximizar a *negentropy* é equivalente a maximizar a soma das *negentropies* marginais, o que possibilita a estimação de diversos componentes independentes de modo seqüencial. Nesse caso, restringir a variância de $\vec{w}^T\vec{z}$ a 1 é equivalente a restringir que $\left\|\vec{w}\right\| = 1$, supondo que os dados de entrada passaram por um processo de branqueamento. O algoritmo é baseado num esquema iterativo e pode ser derivado por meio do método de Newton.

Para a derivação do algoritmo iterativo através do método de Newton, deve-se primeiramente observar que o máximo de $J\left(\vec{w}^T\vec{z}\right)$ é obtido otimizando $E\left[G\left(y\right)\right]$ na equação (5.9), sujeito à restrição $E\left[\vec{w}^T\vec{z}\right] = \left\|\vec{w}\right\|^2 = 1$. Utilizando-se multiplicadores de Lagrange para escrever uma única expressão, derivando-se a expressão em relação a $\vec{w}$, e igualando-se o resultado a zero, tem-se:

$$E\left[\vec{z}g\left(\vec{w}^T\vec{z}\right)\right]+\beta\vec{w}=0 \tag{5.10}$$

sendo que $g(.)$ é a derivada da função G escolhida, e β o multiplicador de Lagrange.

O próximo passo é resolver a equação (5.10) em relação a $\vec{w}$ utilizando o método de Newton, o que é equivalente a encontrar o ponto ótimo da expressão lagrangiana. Chamando o lado esquerdo da equação acima de F, a derivada em relação a $\vec{w}$ fica:

$$\frac{\partial F}{\partial\vec{w}}=E\left[\vec{z}\vec{z}^T g'\left(\vec{w}^T\vec{z}\right)\right]+\beta I \tag{5.11}$$

onde $g'(.)$ é a derivada da função $g(.)$.

Como a matriz derivada na equação (5.11) precisa ser invertida devido ao método de Newton (pois aparece no denominador), pode-se adotar uma simplificação para substituir o primeiro termo da equação. Considerando que os dados passaram por um processo de branqueamento anteriormente, é razoável realizar a seguinte aproximação:

$$E\left[\vec{z}\vec{z}^T g'\left(\vec{w}^T\vec{z}\right)\right]\approx E\left[\vec{z}\vec{z}^T\right]E\left[g'\left(\vec{w}^T\vec{z}\right)\right]=E\left[g'\left(\vec{w}^T\vec{z}\right)\right]I \tag{5.12}$$

Assim, a matriz torna-se diagonal, sendo facilmente invertida. Portanto, aplicando o método de Newton, obtém-se a seguinte regra de iteração:

$$\vec{w}\leftarrow\vec{w}-\frac{E\left[\vec{z}g\left(\vec{w}^T\vec{z}\right)\right]+\beta\vec{w}}{E\left[g'\left(\vec{w}^T\vec{z}\right)\right]+\beta} \tag{5.13}$$

A iteração anterior pode ser simplificada ainda mais, multiplicando-se ambos os lados por $E\left[g'\left(\vec{w}^T\vec{z}\right)\right]+\beta$ para finalmente obter-se:

$$\vec{w}\leftarrow E\left[\vec{z}g\left(\vec{w}^T\vec{z}\right)\right]-E\left[g'\left(\vec{w}^T\vec{z}\right)\right]\vec{w} \tag{5.14}$$

onde os valores esperados são estimados na prática como as médias amostrais.

Essa é, basicamente, a iteração principal do algoritmo FastICA para encontrar uma direção, dada pelo vetor $\vec{w}$, tal que a projeção dos dados maximize a não-gaussianidade, e conseqüentemente, seja um componente independente. A seguir trata-se do caso onde se deseja estimar diversos componentes independentes, ou seja, várias direções para projetar os dados multivariados. A prova de convergência completa do algoritmo FastICA pode ser encontrada com maiores detalhes em (HYVÄRINEN; KARHUNEN; OJA, 2001).

5.2.1.5 Estimação de diversos componentes

A chave para a estimação de diversos componentes independentes baseia-se no

princípio de que os vetores $\vec{w}_i$ correspondentes as várias direções são ortogonais. Portanto, nesse caso, é necessário executar o algoritmo para uma única direção diversas vezes, com a restrição de que os vetores $\vec{w}_1, \vec{w}_2, ..., \vec{w}_n$ sejam ortogonalizados a cada passo, para que acelere a convergência. Uma versão completa do algoritmo FastICA para a estimação de diversos componentes independentes, com ortogonalização através do método de Gram-Schmidt pode ser encontrada em (HYVÄRINEN; KARHUNEN; OJA, 2001).

5.2.1.6 ICA e Projection Pursuit

É interessante notar a semelhança existente entre essa abordagem ICA e a técnica denominada *Projection Pursuit*, um método desenvolvido em estatística que busca encontrar direções interessantes para a projeção de dados multidimensionais. Tais projeções podem ser utilizadas, por exemplo, para otimizar a visualização dos dados, estimação de densidades, ou mesmo em técnicas de regressão.

Em geral, os dados são projetados em subespaços 1-D ou 2-D e as direções mais interessantes se têm mostrado como aquelas que possuem as distribuições menos gaussianas possíveis. Uma motivação para esse fato é que essas distribuições tendem a ser multimodais, e assim, mostram algumas estruturas de aglomeração dos dados.

Dessa forma, medidas de comportamento não-gaussiano utilizadas na estimação ICA podem ser consideradas, nessa abordagem, como índices de projeção. Deve-se notar que na formulação de P*rojection Pursuit*, não há modelo de dados nem hipótese sobre os componentes independentes. Em ICA, existe a definição do modelo, e a otimização das medidas de não-gaussianidade fornece os componentes independentes. Na ausência desse modelo, tais otimizações fornecem as direções ideais em *Projection Pursuit*, estabelecendo, assim, uma estreita relação entre as duas técnicas.

5.2.2 Estimação ICA por máxima verossimilhança

Uma abordagem muito popular para estimar os componentes independentes é um modelo de estimação de máxima verossimilhança, pois se trata de uma técnica fundamental na teoria estatística utilizada na resolução de diversos problemas.

5.2.2.1 A Função de verossimilhança em ICA

A derivação da função de verossimilhança do modelo ICA é baseada no resultado da densidade de uma transformação linear. Considerando o modelo ICA:

$$\vec{x} = A\vec{s} \tag{5.15}$$

pode-se escrever:

$$p_x(\vec{x}) = |\det B|\, p_s(\vec{s}) = |\det B| \prod_i p_i(s_i) \tag{5.16}$$

onde $B = A^{-1}$, e p_i denota as densidades dos componentes independentes. Essa equação pode ser expressa como uma função de $B = \left(\vec{b}_1, \vec{b}_2, ..., \vec{b}_n\right)^T$ e $\vec{x}$:

$$p_x(\vec{x}) = |\det B| \prod_i p_i\left(\vec{b}_i^T \vec{x}\right) \tag{5.17}$$

Assumindo-se que existem T observações de $\vec{x}$, a função de verossimilhança $L(B)$ pode ser obtida como o produto da densidade calculado nos T pontos:

$$L(B) = \prod_{t=1}^{T} \prod_{i=1}^{n} p_i\left(\vec{b}_i^T \vec{x}_t\right) |\det B| \tag{5.18}$$

Em geral é mais fácil utilizar o logaritmo da razão de verossimilhança, devido a duas propriedades algébricas. Assim, a equação acima torna-se:

$$\log L(B) = \sum_{t=1}^{T} \sum_{i=1}^{n} \log p_i\left(\vec{b}_i^T \vec{x}_t\right) + T \log |\det B| \tag{5.19}$$

Para simplificar ainda mais, pode-se denotar a soma sobre as amostras $\vec{x}_t$ como um operador de valor esperado e dividir a expressão por T:

$$\frac{1}{T} \log L(B) = E\left[\sum_{i=1}^{n} \log p_i\left(\vec{b}_i^T \vec{x}_t\right)\right] + \log |\det B| \tag{5.20}$$

Na prática, pode-se substituir o valor esperado pela média amostral. Se as densidades dos componentes independentes não são conhecidas, essa abordagem torna-se mais complicada, pois será preciso utilizar aproximações. Porém, pode ser mostrado em ICA, que na estimação de máxima verossimilhança é suficiente utilizar apenas duas opções na aproximação para a densidade de um componente independente. Para cada componente independente, basta determinar qual das duas aproximações é mais adequada (subgaussiana ou supergaussiana) através do cálculo de momentos não polinomiais e escolher aquela que melhor satisfaz um critério de estabilidade adotado. Um exemplo é quando os logaritmos das aproximações das densidades são dados por:

$$\begin{aligned} \log \hat{p}_i^+(s) &= \alpha_1 - 2\log\left[\cosh(s)\right] \\ \log \hat{p}_i^-(s) &= \alpha_2 - \left\{\frac{s^2}{2} - \log\left[\cosh(s)\right]\right\} \end{aligned} \tag{5.21}$$

onde a_1 e a_2 são constantes positivas. A justificativa para esse resultado é que $\hat{p}_i^+$ é

uma densidade supergaussiana enquanto $\hat{p}_i^-$ é subgaussiana, ou seja, ambas tendem a maximizar a nãogaussianidade, que é justamente um dos critérios utilizados em ICA para encontrar os componentes independentes.

5.2.2.2 Algoritmos ICA de máxima verossimilhança

Para realizar a estimação de máxima verossimilhança, são necessários algoritmos para maximizar a função de verossimilhança. As técnicas existentes na literatura utilizam o método do gradiente na otimização do problema, como é o caso do algoritmo de *Bell-Sejnowski*, inicialmente proposto em (BELL; SEJNOWSKI, 1995). A idéia desse método consiste basicamente em calcular o gradiente do logaritmo da função de verossimilhança dada pela equação (5.20). Porém, resultados experimentais indicam que sua convergência é lenta em alguns casos, devido a necessidade de inversão da matriz B a cada passo do algoritmo. Calculando o gradiente do logaritmo da função de verossimilhança (5.20), tem-se:

$$\frac{1}{T}\frac{\partial \log L(B)}{\partial B} = \left(B^T\right)^{-1} + E\left[g\left(B\vec{x}\right)\vec{x}^T\right] \tag{5.22}$$

onde $g(\vec{y}) = \left[g_i(y_i),...,g_n(y_n)\right]$ é um vetor contendo as derivadas das aproximações das distribuições dos componentes independentes $g_i(y_i) = \left(\log \hat{p}_i\right)'$. Assim, a iteração para o algoritmo de estimação por máxima verossimilhança fica:

$$\Delta B \propto \left(B^T\right)^{-1} + E\left[g\left(B\vec{x}\right)\vec{x}^T\right] \tag{5.23}$$

Maiores detalhes sobre essa abordagem podem ser encontrados em (BELL; SEJNOWSKI, 1995). Outra abordagem de máxima verossimilhança para estimação ICA é o princípio *Infomax*, que se baseia na idéia de maximizar a entropia de saída de uma rede neural com entradas não-lineares. Em (LEE; GIROLAMI; SEJNOWSKI, 1999) é proposto um algoritmo que estende o algoritmo *Infomax* (*Extended Infomax*), pela da utilização do gradiente natural na maximização da função de verossimilhança. Esse método é considerado como uma versão otimizada do algoritmo de Bell & Sejnowski, pois é equivalente, e aumenta a velocidade de convergência consideravelmente. Além disso, (CARDOSO, 1997) mostra que o princípio *Infomax* de estimação ICA coincide com a estimação de máxima verossimilhança, pois ambos os casos consistem na minimização da divergência de Kullback-Leibler, um critério bastante utilizado em Teoria da Informação.

5.2.3 ICA pela minimização da informação mútua

Uma abordagem natural para ICA é a minimização da informação mútua, que é uma medida freqüentemente utilizada em Teoria da Informação. A informação mútua é uma medida de dependência entre variáveis aleatórias, que é sempre não negativa, e zero, se e somente se, as variáveis são estatisticamente independentes, ou seja, pode ser utilizada como função objetivo na estimação ICA. A informação mútua pode ser definida através da entropia como:

$$I\left(x_1, x_2, ..., x_n\right) = \sum_{i=1}^{n} H\left(x_i\right) - H\left(\vec{x}\right) \tag{5.24}$$

Essa definição pode ser interpretada como o uso da entropia em codificação. De acordo com a Teoria da Informação, a entropia está relacionada ao tamanho do código necessário para representar uma variável aleatória. Os termos $H\left(x_i\right)$ definem o tamanho dos códigos para x_i quando cada elemento é codificado separadamente, e $H\left(\vec{x}\right)$ fornece o tamanho do código quando $\vec{x}$ é codificado como um vetor. A informação mútua mostra a redução no código obtido codificando o vetor inteiro ao invés de cada componente separado. Assim, se x_i são independentes, não fornecem informações adicionais uns aos outros, e o tamanho do código é o mesmo em ambas as situações.

No modelo ICA pode-se definir os componentes independentes como uma transformação dos dados observados, ou seja:

$$\vec{y} = B\vec{x} \tag{5.25}$$

Nessa abordagem busca-se encontrar a matriz B de forma que a informação mútua entre os componentes de $\vec{y}$ seja minimizada. Alternativamente, a informação mútua pode ser interpretada como uma distância, ou seja, como a divergência de *Kullback-Leibler* entre a densidade de probabilidade de $\vec{x}$, $p_x\left(\vec{x}\right)$, e o produto das densidades marginais $p_i\left(x_i\right)$:

$$I\left(\vec{x}_1,\vec{x}_2,...,\vec{x}_n\right)=-H\left(\vec{x}\right)+\sum_{i=1}^{n}H\left(x_i\right)=$$

$$E\left[\log p_x\left(\vec{x}\right)\right]-\sum_{i=1}^{n}E\left[\log p_i\left(x_i\right)\right]=E\left[\log p_x\left(\vec{x}\right)-\log\prod_{i=1}^{n}p_i\left(x_i\right)\right]= \quad (5.26)$$

$$\int_{-\infty}^{\infty}p_x\left(\vec{x}\right)\log\left(\frac{p_x\left(\vec{x}\right)}{\prod_{i=1}^{n}p_i\left(x_i\right)}\right)d\vec{x}=D_{KL}\left[p_x\left(\vec{x}\right),\prod_{i=1}^{n}p_i\left(x_i\right)\right]$$

No caso de independência, tem-se que a divergência de *Kullback-Leibler* é zero, ou seja, a densidade conjunta é exatamente igual ao produto das marginais.

Combinando as equações (5.24) e (5.25), além da formula que relaciona a função densidade de probabilidade de $\vec{x}$ e $\vec{y}$, pode-se escrever a informação mútua como:

$$I\left(\vec{y}\right)=-H\left(\vec{x}\right)-\ln\left|\det W\right|-\sum_{i=1}^{n}\int p_i\left(y_i\right)\ln p_i\left(y_i\right)dy_i \quad (5.27)$$

Uma vez que $H\left(\vec{x}\right)$ não depende de W, segundo (THEODORIDIS; KOUTROUMBAS,2003), minimizar $I\left(\vec{y}\right)$ é equivalente a maximização de:

$$J\left(W\right)=\ln\left|\det W\right|+E\left[\sum_{i=1}^{n}\ln p_i\left(y_i\right)\right] \quad (5.28)$$

Aplicando o gradiente na função de custo $J\left(W\right)$ definida acima em relação a W:

$$\frac{\partial J\left(W\right)}{\partial W}=\left(W^T\right)^{-1}-E\left[\phi\left(\vec{y}\right)x^T\right] \quad (5.29)$$

onde $\phi\left(\vec{y}\right)$ é um vetor de *n* elementos definido por:

$$\phi\left(\vec{y}\right)=\left[-\frac{p_1^{'}\left(y_1\right)}{p_1\left(y_1\right)},...,-\frac{p_n^{'}\left(y_n\right)}{p_n\left(y_n\right)}\right]^T \quad (5.30)$$

sendo que $p_i^{'}\left(y_i\right)$ representam as derivadas das densidades marginais, ou seja:

$$p_i^{'}\left(y_i\right)=\frac{dp_i\left(y_i\right)}{dy_i} \quad (5.31)$$

As derivadas das densidades marginais dependem do tipo de aproximação adotada para cada caso. Dessa forma, utilizando o método do gradiente, pode-se escrever o *i*-ésimo passo de iteração como:

$$\begin{aligned} W_i &= W_{i-1} + \mu\left(\left(W_{i-1}^T\right)^{-1} - E\left[\phi\left(\vec{y}\right)\vec{x}^T\right]\right) \\ W_i &= W_{i-1} + \mu\left(I - E\left[\phi\left(\vec{y}\right)\vec{y}^T\right]\right)\left(W_{i-1}^T\right)^{-1} \end{aligned} \tag{5.32}$$

Algumas observações importantes devem ser feitas em relação a essa abordagem. Em primeiro lugar, a necessidade de inversão de matriz a cada passo do algoritmo afeta negativamente o desempenho do método. Além disso, a partir da equação (5.29), é verificado que no ponto estacionário (gradiente igual a zero), a seguinte relação é verdadeira (multiplicando ambos os lados da equação por W^T):

$$\frac{\partial J\left(W\right)}{\partial W} W^T = E\left[I - \phi\left(\vec{y}\right)\vec{y}^T\right] = 0 \tag{5.33}$$

Em outras palavras, o que se consegue com ICA é uma generalização não linear do PCA. Ainda, segundo (THEODORIDIS; KOUTROUMBAS,2003), ao considerar o não correlacionamento, que pode ser definido como $E\left[I - \vec{y}\vec{y}^T\right] = 0$, verifica-se que a presença da função não linear $\phi(.)$ faz com que se consiga ir além dessa condição. Foi a partir dessa idéia que foram inspirados os trabalhos pioneiros em ICA como uma generalização direta do PCA, desenvolvidos em (JUTTEN; HERAULT, 1991), e a relação entre a introdução de estatística de ordens superiores pela utilização de não linearidades.

Outros métodos mais otimizados de estimação ICA baseados no critério de minimização da informação mútua foram propostos na literatura. Em (ROBILA; VARSHNEY, 2002a) é desenvolvido um algoritmo ICA para detecção de alvos em imagens hiperespectrais mais eficaz que o método descrito anteriormente, pois elimina a necessidade de inversão da matriz W a cada passo. Nesse caso, as observações são denotadas pelo vetor $\vec{x}$ e o algoritmo tenta encontrar uma transformação, representada por uma matriz W, tal que nos dados obtidos $\vec{u} = W\vec{x}$, a informação mútua entre os componentes do vetor seja mínima. O algoritmo é desenvolvido através da derivação da expressão da informação mútua em relação aos componentes de W. Da equação (5.26), sabe-se que, para componentes independentes:

$$I\left(u_1, u_2, ..., u_n\right) = E\left[\log p\left(\vec{u}\right)\right] - E\left[\log \prod_{i=1}^{n} p_i\left(u_i\right)\right] = 0 \tag{5.34}$$

Uma das abordagens para minimização da informação mútua é calcular a derivada da expressão anterior em relação aos componentes de W:

$$\frac{\partial I(\mathrm{u}_1, \mathrm{u}_2, ..., \mathrm{u}_n)}{\partial W} = \frac{\partial E\left[\log p(\vec{u})\right]}{\partial W} - \sum_{i=1}^{n} \frac{\partial E\left[\log p_i(u_i)\right]}{\partial W} \tag{5.35}$$

O primeiro termo da equação acima pode ser simplificado através de algumas manipulações algébricas:

$$\frac{\partial E\left[\log p(\vec{u})\right]}{\partial W} = \frac{\partial E\left[\log\left(\frac{1}{|\det W|} p(\vec{x})\right)\right]}{\partial W} = \frac{\partial E\left[\log p(\vec{x})\right]}{\partial W} - \frac{\partial E\left[\log|\det W|\right]}{\partial W} \approx \left(W^T\right)^{-1} \tag{5.36}$$

Para o segundo termo, quando a função densidade de probabilidade dos componentes de $\vec{u}$ é aproximada por funções $g_i(.)$, tem-se:

$$\begin{gathered}\sum_{i=1}^{n} \frac{\partial E\left[\log p_i(u_i)\right]}{\partial W} \approx \sum_{i=1}^{n} \frac{\partial E\left[\log g_i(u_i)\right]}{\partial W} = \sum_{i=1}^{n} \frac{1}{g_i(u_i)} \frac{\partial g_i(u_i)}{\partial W} = \\ \sum_{i=1}^{n} \frac{1}{g_i(u_i)} \frac{\partial g_i(u_i)}{\partial u_i} \frac{\partial u_i}{\partial W} = \left(\frac{1}{g(\vec{u})} \frac{\partial g(\vec{u})}{\partial \vec{u}}\right) \vec{x}^T\end{gathered} \tag{5.37}$$

O passo iterativo pode ser obtido através da regra do gradiente. O gradiente de uma função J aponta para a direção mais íngreme da função num espaço euclidiano ortogonal. Porém, quando o espaço tem uma estrutura métrica de *Riemann*, que é o caso do espaço das matrizes $m \times m$ não singulares utilizadas em ICA, deve-se utilizar um método mais geral, conhecido como gradiente natural. O gradiente natural é calculado a partir do gradiente comum, através da expressão:

$$\frac{\partial J}{\partial W_{Nat}} = \frac{\partial J}{\partial W} W^T W \tag{5.38}$$

Dessa forma, chamando $W_k - W_{k-1} = \Delta W$, e aplicando o gradiente natural, o passo de iteração torna-se:

$$\begin{gathered}\Delta W \propto k \frac{\partial I(\vec{u})}{\partial W} W^T W = k\left[\left(W^T\right)^{-1} + \left(\frac{1}{g(\vec{u})} \frac{\partial g(\vec{u})}{\partial \vec{u}}\right) \vec{x}^T\right] W^T W = \\ kW + k\left(\frac{1}{g(\vec{u})} \frac{\partial g(\vec{u})}{\partial \vec{u}}\right) \vec{u}^T W\end{gathered} \tag{5.39}$$

sendo que k é um escalar e $g(.)$ é uma função exponencial, que é definida em (ROBILA; VARSHNEY, 2002b) como $g(u) = \frac{1}{(1+e^{-u})}$.

Resultados experimentais obtidos no trabalho desenvolvido por (ROBILA; VARSHNEY, 2002b) mostram que o método é capaz de identificar possíveis alvos em imagens hiperespectrais, sendo que eles são representados por áreas retangulares,

contendo materiais previamente escolhidos, representando alvos a serem detectados, como mostra a Figura 12:

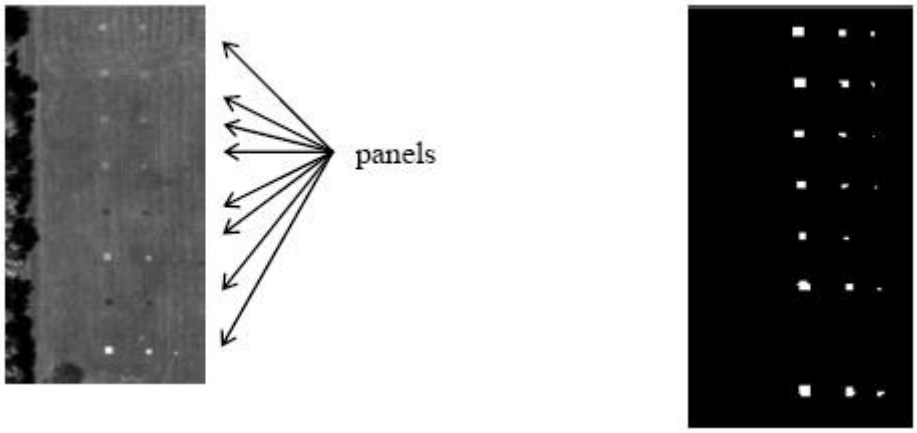


Figura 12*. Aplicação de algoritmo ICA para detecção de alvos em imagens hiperespectrais.*
Fonte: (ROBILA; VARSHNEY, 2002b).

Em (AMARI; CICHOCKI; YANG, 1996) é proposto um algoritmo ICA baseado na minimização da Informação Mútua, utilizando a divergência de *Kullback-Leibler* como função objetivo e o método do gradiente natural para a otimização. Resultados de simulações obtidos também com dados sintéticos verificam a validade do método. Um novo algoritmo proposto em (DINH-TUAN PHAM, 2004), realiza a estimação ICA através da minimização da informação mútua com a utilização da verdadeira densidade dos componentes independentes $g(u)$ estimada não parametricamente. O grande problema é que nesse tipo de estimação (não paramétrica), a dimensionalidade se torna um empecilho, dado que para se obter um resultado razoável é necessário uma quantidade muito grande de amostras, um requisito que raras vezes é satisfeito em aplicações reais (amostras de treinamento limitadas). Por esse motivo, não se tem resultados práticos da viabilidade do método em casos reais, apenas em simulações. Uma vantagem desse algoritmo é a utilização de um método de otimização quasi-Newton para acelerar a convergência (em relação ao método do gradiente).

5.2.4 ICA através de PCA Não Linear

Descorrelação não linear pode ser vista como uma extensão de métodos estatísticos de segunda ordem, como branqueamento e *Principal Component Analysis*. As funções não lineares utilizadas nessa abordagem introduzem estatísticas de ordens superiores, tornando a estimação ICA possível. Assim, em certos casos, a abordagem de PCA não

linear fornece os componentes independentes. O objetivo é mostrar que essa abordagem é equivalente a outros critérios como a utilização de *cumulants* e estimação por máxima verossimilhança, por exemplo.

5.2.4.1 PCA não linear

Os critérios existentes para o PCA linear foram citados anteriormente no capítulo 4. O objetivo agora é verificar como o critério e sua solução são modificados na presença de não-linearidades. A maneira mais utilizada para introduzir não linearidades em PCA é assumir um conjunto de funções não lineares $g_1(.), g_2(.), ..., g_n(.)$ e considerar o seguinte critério:

$$J(\vec{w}_1, ...\vec{w}_n) = E\left[\left\|\vec{x} - \sum_{i=1}^{n} g_i\left(\vec{w}_i^T \vec{x}\right)\vec{w}_i\right\|^2\right] \tag{5.40}$$

Esse critério foi primeiramente considerado por (XU, 1993) e é denominado *Least Mean Square Error Reconstruction* (LMSER), ou Reconstrução por Mínimo Erro Médio Quadrático. A diferença é que nesse caso têm-se funções não lineares dos fatores lineares $\vec{w}_i^T \vec{x}$. Portanto, a técnica de encontrar vetores de base $\vec{w}_i$ que minimizam o erro médio quadrático da representação do vetor na nova base é denominada PCA não linear (NLPCA). Deve ser notado que minimizar o critério não linear não garante um erro médio quadrático menor que o PCA padrão. A vantagem desse critério é que ele introduz estatísticas de ordem superiores de maneira relativamente simples via não-linearidades.

5.2.4.2 Relação entre ICA e PCA não-linear

É interessante notar a relação existente entre o critério PCA não-linear e abordagens ICA como a otimização da curtose, e a máxima verossimilhança. Para dados brancos, pode-se mostrar a equivalência entre esses critérios. Considerando $\vec{z}$ como o vetor de entrada $\vec{x}$ que já passou por um processo de branqueamento e com dimensionalidade *n* igual ao vetor $\vec{s}$ dos componentes independentes, e denotando por $W = (\vec{w}_1, ..., \vec{w}_n)$ a matriz que tem como linhas os vetores da base $\vec{w}_i$, pode-se escrever a equação (5.40) em notação matricial da seguinte maneira:

$$J(\vec{w}_1, ..., \vec{w}_n) = J(W) = E\left[\left\|\vec{z} - W^T g(W\vec{z})\right\|^2\right] \tag{5.41}$$

sendo que o resultado da função $g(W\vec{z})$ é um vetor coluna. Após algumas

manipulações algébricas tem-se:

$$\begin{gathered}\left\|\vec{z}-W^T g\left(W\vec{z}\right)\right\|^2=\left(\vec{z}-W^T g\left(W\vec{z}\right)\right)^T\left(\vec{z}-W^T g\left(W\vec{z}\right)\right)= \\ \left(\vec{z}-W^T g\left(W\vec{z}\right)\right)^T W^T W\left(\vec{z}-W^T g\left(W\vec{z}\right)\right)= \\ \left\|W\vec{z}-WW^T g\left(W\vec{z}\right)\right\|^2=\left\|\vec{y}-g\left(\vec{y}\right)\right\|^2= \\ \sum_{i=1}^{n}\left(y_i-g_i\left(y_i\right)\right)^2\end{gathered} \tag{5.42}$$

com $\vec{y}=W\vec{z}$. Assim, o critério $J\left(W\right)$ torna-se:

$$J_{NLPCA}\left(W\right)=\sum_{i=1}^{n}E\left[\left(y_i-g_i\left(y_i\right)\right)^2\right] \tag{5.43}$$

Em um caso particular, quando as funções g_i são definidas como as seguintes funções quadráticas:

$$y=\begin{cases}y^2+y & \text{se } y\geq 0\\ -y^2+y & \text{se } y<0\end{cases} \tag{5.44}$$

O critério definido pela equação acima torna-se:

$$J_{kurt}\left(W\right)=\sum_{i=1}^{n}E\left[\left(y_i-y_i\pm y_i^2\right)^2\right]=\sum_{i=1}^{n}E\left[y_i^4\right] \tag{5.45}$$

que é igual à curtose a menos de uma constante (3), para o caso de dados brancos. Ou seja, minimizar o critério de erro médio quadrático no PCA não-linear é equivalente a minimizar a curtose, que é um dos possíveis critérios para a estimação ICA, uma vez que a curtose é zero para distribuição gaussiana, e minimizá-la ou maximizá-la fornece as projeções menos gaussianas possíveis (densidades subgaussianas se curtose negativa ou supergaussianas se positiva), estimando os componentes independentes.

No caso da estimação por máxima verossimilhança, parte-se do logaritmo da função de verossimilhança L definida anteriormente. Considerando os dados de entrada brancos como $\vec{z}$, e L como função de W, tem-se:

$$\log L\left(W\right)=\sum_{t=1}^{T}\sum_{i=1}^{n}\log p_i\left(\vec{w}_i^T\vec{z}_t\right) \tag{5.46}$$

É mostrado que quando $t\to\infty$ essa função tende a:

$$J_{MV}\left(W\right)=\sum_{i=1}^{n}E\left[\log p_i\left(\vec{w}_i^T\vec{z}_t\right)\right] \tag{5.47}$$

A partir deste ponto, pode-se derivar a conexão entre o PCA não linear e o critério ICA de estimação por máxima verossimilhança. Na minimização do somatório em

(5.43), não há maiores problemas em se adicionar um termo multiplicativo β. Dessa forma, para que haja a equivalência dos critérios, e como o critério de PCA não linear é um problema de minimização (5.43) e o critério ICA de máxima verossimilhança (5.47) é um problema de maximização, troca-se o sinal de um dos argumentos (as expressões diferem apenas pelo argumento do valor esperado). Na condição em que os argumentos são iguais, tem-se:

$$\log p_i(y_i) = -\beta\left(y_i - g_i(y_i)\right)^2 \tag{5.48}$$

que leva ao resultado:

$$p_i(y_i) \propto \exp\left\{-\beta\left[y_i - g_i(y_i)\right]^2\right\} \tag{5.49}$$

mostrando a relação entre os dois critérios. Essa expressão ajuda na escolha da função $g_i(y_i)$ para uma dada densidade $p_i(y_i)$ observando-se a forma de sua expressão, quando se pretende utilizar o critério de PCA não linear para estimar os componentes independentes de maneira equivalente à estimação ICA por máxima verossimilhança. Detalhes e informações adicionais sobre essa abordagem podem ser encontrados em (OJA et al., 1995) e (OJA, 1997).

5.2.5 Métodos baseados em *cumulants* (*Cumulant-based Methods*)

Outra possível abordagem para estimação ICA são os métodos baseados em *cumulants* (*cumulant-based methods*). A idéia dessa abordagem é generalizar o conceito utilizado na Transformação de Karhunen-Loève de diagonalizar a matriz de covariância (descorrelação) utilizando estatísticas que não se limitam a segunda ordem apenas.

O algoritmo JADE, ou *Joint Approximate Diagonalization of Eigenmatrices*, proposto em (CARDOSO, 1999), baseia-se na generalização da abordagem PCA pela utilização de estatística de ordens superiores através do tensor *cumulant* de quarta ordem (generalização de matriz de covariância), representado por uma matriz de quatro dimensões (tensor), onde as entradas são os *cumulants* cruzados de quarta ordem dos dados. Pode ser mostrado que um tensor pode ser decomposto em termos de suas matrizes próprias. O método busca, então, implementar a diagonalização simultânea das matrizes próprias do tensor como forma de aproximar a condição de independência, onde todos os *cumulants* cruzados são nulos e o tensor *cumulant* é diagonal. Recentemente, otimizações para algoritmos ICA que utilizam essa idéia (diagonalização do tensor *cumulant* de quarta ordem) foram desenvolvidos em (BLASCHKE;

WISKOTT, 2004). Tais otimizações são baseadas na utilização de Rotações de Givens (*Givens Rotations*) com o objetivo de alcançar a condição de independência.

Outro método de estimação ICA que utiliza estatística de ordens superiores através de *cumulants* de terceira e quarta ordens é proposto em (ZHANG; CHEN, 2004). Esse método busca minimizar uma função de custo definida como um somatório de *cumulants* cruzados. Resultados da aplicação desse método em imagens *SAR* (*Synthetic Aperture Radar*) utilizadas em sensoriamento remoto são comparados com resultados obtidos por outras abordagens ICA e com a técnica PCA.

5.3 Considerações sobre a estimação ICA

Métodos estatísticos de ordens superiores (i.e, algoritmos ICA) tornam todo o processo mais robusto a ruídos aditivos e mais apropriados a dados com alta dimensionalidade. Entretanto, o uso inadequado de métodos ICA para extração de atributos pode degradar severamente o desempenho obtido na etapa de classificação. Basicamente, isso pode ocorrer em alguns casos, devido à presença de valores próprios pequenos da matriz de covariância, o que causa um problema de mal-condicionamento durante o processo de branquemento dos dados (*whitening*), um processo muito comum utilizado para simplificar o desenvolvimento e acelerar a convergência de diversos algoritmos ICA. Em alguns casos, trata-se de um pré-requisito para estimação de componentes independentes. Portanto, para evitar problemas como amplificações de eventuais ruídos, mal-condicionamento e estatísticas de segunda ordem, um esquema hierárquico com PCA e ICA é proposto.

Além disso, em se tratando de dados de sensoriamento remoto, é freqüente a disponibilidade de informações adicionais (i.e, verdade terrestre), sendo interessante incluí-las no processo de extração de atributos. Portanto, os resultados podem ser melhorados ainda mais com a utilização de etapas adicionais, que pode ser um método linear supervisionado (LDA) e/ou algoritmos para seleção de atributos no caso de limitações quanto ao número de atributos, como será discutido em detalhe no capítulo seguinte na seção sobre a metodologia proposta. Métodos que combinam a transformação de Karhunen-Loève e LDA têm sido aplicados com sucesso em problemas de reconhecimento de faces como em (YANG; YE; ZHANG, 2004) e mostram bons resultados.

5.3.1 Branqueamento dos dados

O branqueamento dos dados (*whitening*) é freqüentemente utilizado como uma etapa de pré-processamento em ICA, com o objetivo de simplificar as deduções e também acelerar a convergência dos algoritmos. Como esse processo é, em essência, decorrelação seguida de escala, a técnica PCA pode ser utilizada. Isso implica que o branqueamento pode ser realizado como uma transformação linear. O problema é, então, dado um vetor aleatório $\vec{x}$, encontrar uma transformação V que produz um vetor $\vec{z}$ que seja branco, ou seja:

$$\vec{z} = V\vec{x} \tag{5.50}$$

Este problema admite uma solução direta, proveniente da expansão de Karhunen-Loève. Seja $E = \left[\vec{e}_1, \vec{e}_2, ..., \vec{e}_n\right]$ a matriz cujas colunas são os vetores próprios normalizados da matriz de covariância $C_X = E\left[\vec{x}\vec{x}^T\right]$, se $\mu = 0$, que pode ser computada diretamente através de amostras do vetor $\vec{x}$. Seja também $D = diag\left(d_1, d_2, ..., d_n\right)$ a matriz diagonal dos valores próprios de C_X. Então, uma transformação de branqueamento pode ser definida por:

$$V = D^{-1/2}E^T \tag{5.51}$$

Para verificar que a matriz V é uma transformação de branquemento, basta escrever a matriz de covariância pela decomposição em valores próprios (EVD), em função de E e D, ou seja, $C_X = EDE^T$, sendo que E é uma matriz ortogonal. Portanto, a matriz de covariância de $\vec{z}$, que possui média zero, é a identidade, ou seja, a variância de todos os componentes é unitária e os dados se encontram descorrelacionados (correlação nula):

$$E\left[\vec{z}\vec{z}^T\right] = VE\left[\vec{x}\vec{x}^T\right]V^T = D^{-1/2}E^T EDE^T ED^{-1/2} = I \tag{5.52}$$

Porém, o operador linear V não é único. Qualquer UV, onde U é uma matriz ortogonal também é uma transformação de branqueamento. Isso é válido, pois para $\vec{z} = UV\vec{x}$:

$$E\left[\vec{z}\vec{z}^T\right] = UVE\left[\vec{x}\vec{x}^T\right]V^T U^T = I \tag{5.53}$$

Uma matriz de branqueamento de particular interesse é $ED^{-1/2}E^T$. Ela é obtida multiplicando-se o lado esquerdo da equação (5.51) pela matriz ortogonal E. Essa

matriz geralmente é chamada de raiz quadrada inversa de C_X , e denotada por $C_X^{-1/2}$.

5.3.1.1 O problema do branqueamento

Alguns problemas podem ocorrer ao se utilizar esse tipo de transformação nos dados multivariados extraídos de imagens de sensoriamento remoto. Nesses casos, a relação entre o maior e o menor valor próprio da matriz de covariância é enorme e o número de condição $\lambda_{\max}/\lambda_{\min}$ assume valores extremamente altos. Dessa forma, quaisquer pequenas perturbações no vetor de entrada $\vec{x}$ (ruídos, incerteza ou até aproximações numéricas) produzem uma solução dominada por grandes termos oscilatórios, caracterizando um comportamento semelhante ao de problemas mal-condicionados. Uma solução para esse problema consiste em encontrar um subespaço PCA adequado (redução de dimensionalidade) onde esses efeitos indesejáveis sejam atenuados e os atributos extraídos pela estimação ICA produzam um melhor desempenho na classificação.

5.3.2 Ortogonalização

A partir da teoria de PCA e ICA, sabe-se que os vetores solução são ortogonais, ou ortonormais, formando uma base. Mas, os algoritmos iterativos não produzem automaticamente um conjunto de vetores ortogonais, ou seja, é necessária a utilização de métodos de ortogonalização. Esse problema pode ser descrito da seguinte maneira: Dado um conjunto de *m* vetores *n*-dimensionais linearmente independentes $\vec{a}_1, \vec{a}_2, ..., \vec{a}_m$, com $m \leq n$, computar um outro conjunto de vetores $\vec{w}_1, \vec{w}_2, ..., \vec{w}_m$ ortonormais que gere o mesmo subespaço gerado pelos vetores originais.

Basicamente, existem duas classes de métodos de ortogonalização: os métodos seqüenciais e simétricos. Nos métodos simétricos, parte-se de um vetor original inicial e a cada iteração obtém-se um vetor ortogonal aos demais. Como exemplo, tem-se o método de ortogonalização de Gram-Schmidt (GSO). Já nos métodos simétricos, nenhum vetor original é privilegiado em relação aos demais, ou seja, são métodos de ortogonalização de matrizes.

5.4 Considerações Finais

Resultados obtidos na literatura mostram que a estimação ICA pela minimização da informação mutua é equivalente a encontrar direções em que a entropia relativa é

maximizada, ou seja, equivale a maximizar a não-gaussianidade. Além disso, pode-se mostrar que a informação mútua e a verossimilhança estão diretamente ligadas, pois o logaritmo da função de verossimilhança é igual, além de um termo constante e aditivo, ao negativo da expressão da informação mútua. Portanto, verifica-se que todas as abordagens são, na realidade, equivalentes.

Em (DUDA; HART; STORK, 2001) é proposta uma abordagem para a estimação ICA onde o objetivo é encontrar um parâmetro W que torne os dados transformados o mais independentes possível utilizando como critério a maximização da entropia conjunta. Em (HAYKIN, 1998) são propostos modelos neurais para o problema de *Blind Source Separation* que estimam os componentes independentes da mistura, utilizando critérios como a minimização da divergência de *Kullback-Leibler* da saída, através da determinação da soma das entropias marginais, pela definição da equação (5.26).

Além dos algoritmos citados anteriormente, existem vários métodos baseados em redes neurais utilizados para estimação ICA utilizados com sucesso em diversos tipos de aplicações. Alguns dos trabalhos pioneiros usando essa abordagem podem ser encontrados em (BELL; SEJNOWSKI, 1995), (CICHOCKI; UNBERHAUEN, 1996) e (KARHUNEN et al., 1997). Finalmente, é interessante perceber que embora existam dezenas de métodos para estimação ICA com diferentes funções e critérios a serem otimizados, cada um deles tem o mesmo objetivo: fornecer como resultado aproximações para os componentes independentes.

6 DESENVOLVIMENTO DO PROJETO

O presente trabalho apresenta um modelo de combinação/fusão de atributos com o intuito de melhorar o desempenho da classificação, combinando métodos estatísticos de segunda ordem com métodos de ordens superiores, para assim, superar limitações existentes nas abordagens tradicionais. O esquema resultante é utilizado para combinar atributos obtidos através de diversos métodos num único vetor de padrões em duas abordagens: Fusão Hierárquica e Fusão Concatenada. A metodologia proposta é aplicada em diferentes estudos de casos, representando problemas com dimensionalidade baixa (3 bandas), moderada (12 bandas) e alta (220 bandas). As imagens são classificadas utilizando-se a abordagem bayesiana. Resultados indicam que essa metodologia supera métodos tradicionais, constituindo uma interessante ferramenta para extração de atributos visando a classificação.

6.1 Metodologia

A metodologia proposta consiste em, a partir da obtenção de imagens de sensoriamento remoto multiespectrais e hiperespectrais, combinar técnicas de extração (fusão de atributos) e seleção de atributos para classificação supervisionada, através de abordagem bayesiana.

6.1.1 Imagens Multiespectrais

Imagens multiespectrais contém dados em uma estrutura lógica tridimensional. Para armazenar esses dados em um arquivo, de maneira seqüencial, é necessário um mapeamento adequado. Em sensoriamento remoto, é comum a utilização de formatos *Band Interleaved*. Existem basicamente três variações: *Band Interleaved by Pixel* (BIP), *Band Interleaved by Line* (BIL) e *Band Sequential* (BSQ). A definição do formato da imagem é essencial para que a leitura dos dados do arquivo seja realizada corretamente. A Figura 13 juntamente com a Tabela 1 mostra a organização para cada um dos formatos.

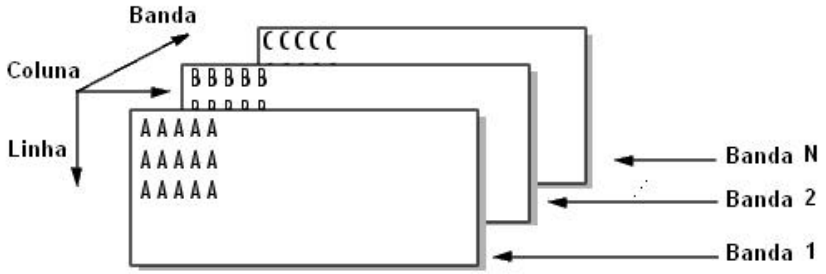


Figura 13*. Estrutura tridimensional de imagens multiespectrais.*

Tabela 1*. Formatos padrões para imagens multiespectrais de sensoriamento remoto.*

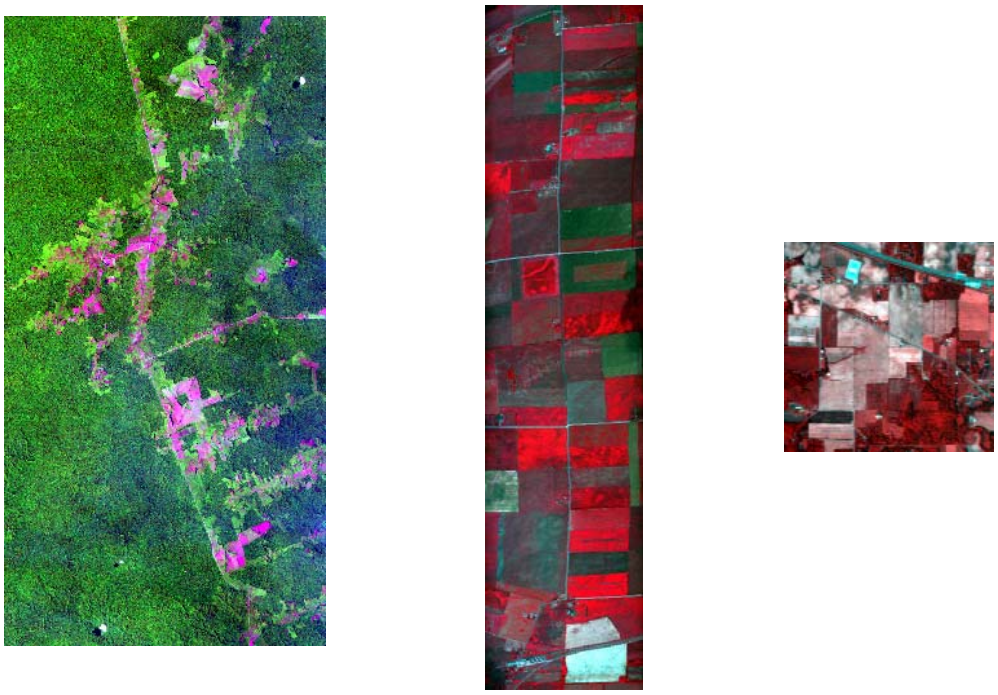
Método	Sigla	Descrição	Representação
Band Interleaved by Line	BIL	Leitura/Escrita uma linha de cada banda. Método intermediário para acessar tanto as informações espaciais quanto as espectrais.	AAAABBBBCCCC AAAABBBBCCCC AAAABBBBCCCC
Band Interleaved by Pixel	BIP	Leitura/Escrita de um pixel de cada banda. Otimizado para acessar informações espectrais.	ABCABCABCABC ABCABCABCABC ABCABCABCABC
Band Sequential	BSQ	Leitura/Escrita de uma banda por completo. Otimizado para acessar informações espaciais.	AAAAAAAAAAAA BBBBBBBBBBBB CCCCCCCCCCCC

Durante o desenvolvimento do trabalho, para avaliar os métodos desenvolvidos, os algoritmos foram aplicados a três estudos de casos distintos: duas imagens multiespectrais com dimensionalidade baixa (3 bandas) e moderada (12 bandas), além de uma imagem hiperespectral com dimensionalidade alta (220 bandas). Abaixo segue uma breve descrição sobre cada uma das imagens utilizadas:

- *Tapajos.img* : Imagem multiespectral da região da Floresta Nacional de Tapajós (Figura 14a), Pará, Brasil, contendo 3 bandas, com amostras de treinamento e teste divididas em 4 classes. As dimensões da imagem são de 3900×1900 pixels digitalizados com 8 bits, *IEEE Little Endian* (bits mais significativos a esquerda), e formato BSQ. Estima-se que a cena equivale a uma área aproximada de $25 \times 50\ \text{km}^2$. Mais informações sobre a região podem ser encontradas em (FREITAS; SANT'ANNA; RENNÓ, 1999).

- *Flc1.lan* : Imagem multiespectral da região do estado norte americano de Indiana (Figura 14b), contendo 12 bandas, com verdade terrestre (8 classes de vegetação), dimensões espaciais de 949×220 pixels digitalizado com precisão de 8 bits, *IEEE Little Endian* (bits mais significativos a esquerda), cabeçalho de 128 bytes e formato BIL.

- *92AV3C* : Imagem hiperespectral (sensor AVIRIS) com 220 bandas (Figura 14c), amostras de treinamento e teste divididas em 10 classes, dimensões de 145×145 pixels digitalizados utilizando-se 16 bits, *IEEE Big Endian* (bits mais significativos a direita), cabeçalho de 128 bytes e formato BIL.



a) ***b)*** ***c)***

Figura 14*. Imagens multiespectrais de sensoriamento remoto utilizadas no trabalho.*

6.1.2 Avaliação dos resultados

A fim de se avaliar os resultados obtidos na classificação supervisionada e realizar uma comparação objetiva entre as técnicas de extração de atributos e o método proposto, é necessário se estimar a probabilidade de erro de classificação em cada caso, a partir do desempenho do classificador. Obviamente, a melhor técnica de extração de atributos é aquela que minimiza o erro de classificação.

Para isso, é adotado o procedimento de separar o conjunto de amostras em um conjunto de treinamento e outro de teste. O desempenho do classificador é medido na classificação das amostras presentes no conjunto de teste. Para a definição da divisão das amostras é utilizado tanto o método *holdout*, quando o número de amostras é grande, quanto o método *leave-one-out cross-validation*, no caso em que o número de amostras é pequeno. No primeiro caso, definem-se dois conjuntos disjuntos, um para

treinamento e outro para teste, sendo assim considerada uma abordagem pessimista. Um critério comum adotado é separar 50% das amostras para teste e 50% para treinamento. No segundo caso, as N amostras são divididas em um conjunto de treinamento de $N-1$ elementos e um conjunto de teste constituído do único elemento restante. Esse processo é repetido N vezes, de modo que cada possível elemento do conjunto de amostras seja utilizado uma vez como conjunto de teste (validação). Esse método tenta evitar possíveis polarizações no estimador introduzidas por uma divisão particular do conjunto de amostras. Porém as desvantagens são o elevado custo computacional e o tempo de processamento.

Além disso, é comum os resultados serem expressos na forma da Matriz de Classificação, também conhecida como *Matriz de Confusão*, que indica o resultado da classificação, fornecendo a distribuição dos percentuais de erros e acertos nas diversas classes. A partir desta matriz, estima-se o desempenho médio (DM), a abstenção média (AM) e a confusão média (CM) do classificador: o primeiro oferece a percentagem de amostras corretamente classificadas, sendo obtido pela média da percentagem de classificação correta para cada classe, ponderada pelo número de amostras de cada classe; o segundo fornece a percentagem de amostras que não foram classificadas e é obtido pela média das percentagens de amostras não classificadas por classe, ponderada pelo número de amostras de cada classe. A confusão média oferece a percentagem de amostras de cada classe que são classificadas como sendo de outra, definida em (GOSE; JOHNSONBAUGH; JOST, 1996) como $1-\left(DM+AM\right)$.

6.1.3 Medidas de desempenho

Para se avaliar a classificação, pode-se calcular uma medida de desempenho a partir da matriz de classificação (ou confusão), denominada Coeficiente *Kappa*, originalmente proposto por (COHEN, 1960). Esse coeficiente é definido como:

$$K = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i+}x_{+i}}{N^2 - \sum_{i=1}^{r} x_{i+}x_{+i}} \tag{5.54}$$

onde r é o número de linhas ou colunas da matriz de confusão, x_{ii} é o número de observações dos elementos da diagonal da matriz de confusão, $x_{i+} = \sum_{j} x_{ij}$ é a soma dos

valores da linha i, $x_{+i} = \sum_j x_{ji}$ é a soma dos valores da coluna i e N é o número total de observações. A Tabela 2 apresenta uma sugestão para uma possível interpretação do desempenho obtido na classificação em função do valor assumido pelo coeficiente Kappa (CONGALTON, 1991). Porém, como cada aplicação possui um dado nível de acerto, devido às características dos dados disponíveis, essa interpretação pode variar.

***Tabela 2**. Possível interpretação da classificação em função do coeficiente Kappa.*

Resultado	Interpretação do desempenho do classificador
$K <= 0$	Péssimo
$0 < K <= 0,2$	Mau
$0,2 < K <= 0,4$	Razoável
$0,4 < K <= 0,6$	Bom
$0,6 < K <= 0,8$	Muito Bom
$0,8 < K <= 1,0$	Excelente

Fonte: (Congalton, 1991).

6.1.4 Materiais e Métodos

O desenvolvimento do trabalho foi composto pela definição de uma base teórica contendo revisões da literatura e discussões sobre temas pertinentes a área de pesquisa, além da implementação prática da solução proposta, de onde serão gerados e analisados os resultados e conclusões da pesquisa.

A implementação dos métodos e algoritmos propostos foi realizada através do software Matlab, devido a sua facilidade em tratar e operar dados multidimensionais como vetores e matrizes, sendo ideal para trabalhar com imagens multiespectrais, além de permitir a integração de diversos pacotes especializados em programação matemática, estatística, processamento de imagens, ente outros, como o pacote FastICA desenvolvido e disponibilizado em (HYVÄRINEN et al., 2004) pela Universidade de Helsinki, Finlândia, usado como referência para implementação de algortimos ICA para maximização da não-gaussianidade, e o pacote PRTOOLS (*Pattern Recognition Tools*), desenvolvido por (DUIN, 2003) na Universidade de Delft, Holanda, e que contém diversas rotinas úteis para aplicações em reconhecimento de padrões. Referências adicionais sobre programação Matlab em reconhecimento de padrões estatístico podem ser encontradas em (MARTINEZ; MARTINEZ, 2002).

Como ferramentas de auxílio ao desenvolvimento foram utilizados softwares para análise e processamento de imagens de sensoriamento remoto, dentre os quais, o *Multispec*, desenvolvido pela Universidade de Purdue e o ambiente *ENVI*, principalmente para visualização das imagens multiespectrais, além da obtenção dos conjuntos de amostras de treinamento e identificação das características referentes a cada imagem (formato, dimensão, codificação, etc).

6.2 Fusão de Atributos

Basicamente, a idéia da metodologia proposta consiste em combinar métodos de extração, juntamente com métodos de seleção de atributos, para se obter melhores conjuntos de atributos do ponto de vista de classificação (separabilidade dos dados), numa tentativa de superar limitações encontradas nos métodos individuais, através de um esquema denominado fusão de atributos, uma vez que o subconjunto de atributos resultante depende de diversos conjuntos de atributos (obtidos através de cada um dos métodos de extração utilizados). O objetivo é combinar atributos obtidos a partir de uma variedade de métodos de extração para formar o vetor de atributos resultante (subconjunto de atributos utilizado na classificação). Para isso, foram propostas duas possibilidades: a fusão hierárquica e a fusão concatenada.

Na abordagem hierárquica, a idéia consiste em, a partir do conjunto original de atributos, aplicar um método de extração para obter um novo subconjunto de atributos com dimensionalidade igual ou menor. Em seguida um novo método de extração é aplicado sobre o novo subconjunto de atributos obtido na primeira etapa para gerar um outro subconjunto, e assim por diante de maneira seqüencial. Nesse caso, o objetivo é gerar cada novo subconjunto de atributos em função do conjunto obtido anteriormente. Dessa forma, o esquema resultante impõe uma seqüência de transformações geométricas no espaço de atributos, de forma que a dimensionalidade do subespaço resultante é sempre menor (ou igual) que a dimensionalidade do espaço de atributos original. A presença ou não de algum método de extração de atributos supervisionado durante o processo, define se a fusão é supervisionada, caso em que são necessárias amostras representativas de cada classe (nível maior de informação) ou não supervisionada, considerando apenas a mistura das classes. Essa abordagem pode ser particularmente interessante quando se deseja incorporar algoritmos ICA no processo de extração, devido a alguns problemas encontrados durante a estimação dos componentes independentes. Um diagrama de blocos ilustrativo pode ser visualizado na Figura 15.

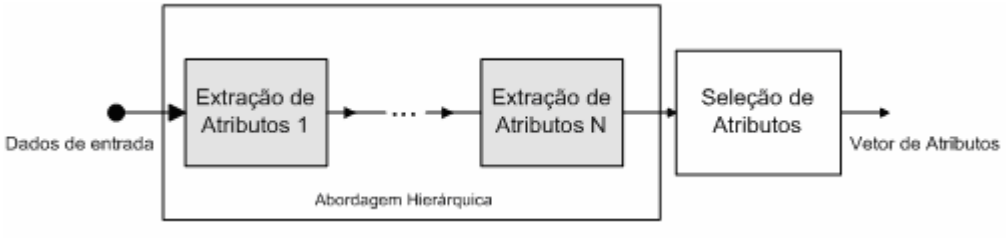


Figura 15*. Esquema para fusão de atributos hierárquica.*

Por exemplo, supondo $\Im$ como o conjunto original de atributos de dimensionalidade n, e T_A, T_B e T_C respectivamente como primeiro, segundo e terceiro métodos de extração, o subconjunto de atributos resultante $\Im_R$ é obtido através da seqüência de aplicações $T_C\left(T_B\left(T_A\left(\Im\right)\right)\right)$, ou seja, a aplicação de T_A em $\Im$ gera um novo subconjunto $\Im_A$ de dimensionalidade $m \le n$ sobre o qual é aplicado o método T_B que obtém o novo subconjunto $\Im_B$ de dimensionalidade $l \le m$, que por sua vez é utilizado por T_C para se obter o subconjunto resultante $\Im_R$ de dimensionalidade $d \le l$, sendo que no caso de métodos lineares, os operadores transformam-se em matrizes.

Além disso, ao término do processo, quando se obtém o subconjunto de atributos resultante, pode ser interessante a aplicação de um método de seleção de atributos (maior nível de informação – disponibilidade de amostras representativas das classes), caso seja necessário a definição de um subconjunto mais restrito. Diversos critérios podem ser utilizados para avaliar a capacidade discriminante dos atributos, o que juntamente com algum algoritmo de busca, como por exemplo, *Best Individual Features*, *Exhaustive Search*, *Sequential Forward Selection*, *"Plus L-take away R" Selection*, dentre outros (JAIN; DUIN; MAO, 2000), provê um modelo adequado para seleção de atributos. Vale ressaltar que nem todos os algoritmos de busca existentes garantem a solução ótima (subconjunto ótimo de atributos). Particularmente, o algoritmo adotado (*Best Individual Features*) seleciona o melhor atributo de todos como o primeiro do subconjunto solução e, a partir daí, adiciona a esse subconjunto o melhor atributo dentre os restantes, e assim por diante. Porém, os k melhores atributos individuais não implicam no subconjunto ótimo de k atributos. Assim, para garantir os subconjuntos ótimos em cada dimensionalidade é necessário, para cada valor de d (dimensionalidade), utilizar um método de busca ótimo, o que significa um custo computacional extremamente alto.

A abordagem concatenada consiste em aplicar diversos métodos de extração no conjunto original de atributos para se obter vários subconjuntos de atributos. Cada método de extração aplicado ao conjunto original, fornece um subconjunto de atributos específico. Por exemplo, ao se aplicar um método de extração T_A no conjunto original $\Im$ de n atributos, obtém-se um subconjunto de atributos $\Im_A$ com m atributos $(m \leq n)$. Da mesma forma, ao se aplicar um método T_B em $\Im$, obtém-se um subconjunto $\Im_B$ com d atributos $(d \leq n)$. A idéia dessa abordagem é concatenar os subconjuntos de atributos obtidos $(\Im_A, \Im_B)$ para definir um único conjunto de atributos $\Im_{AB}$, com dimensionalidade $m+d$, sendo que o processo pode ser generalizado para mais de dois métodos de extração. Na seqüência, um método de seleção de atributos é utilizado para se obter o subconjunto de atributos resultante $\Im_R$ com dimensionalidade $p \leq m+d$, contendo os atributos mais discriminantes de $\Im_{AB}$. Vale ressaltar que, nessa abordagem, a etapa de seleção de atributos é fundamental, pois a concatenação dos conjuntos de atributos leva a um aumento na dimensionalidade do problema, o que possivelmente introduz algumas dificuldades. A Figura 16 ilustra a abordagem concatenada.

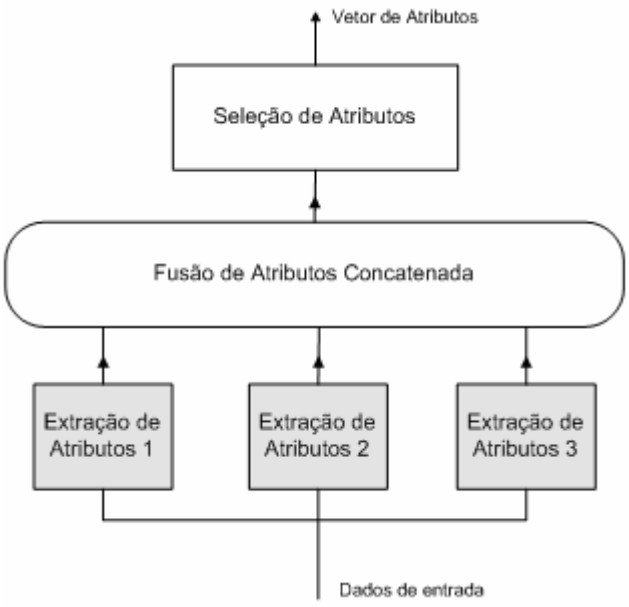


***Figura 16**. Esquema para fusão de atributos concatenada.*

6.3 Experimentos e resultados

A fim de avaliar o desempenho dos métodos propostos e seu efeito na posterior etapa de classificação, experimentos foram realizados em três estudos de caso envolvendo diferentes imagens de sensoriamento remoto com dimensionalidades baixa, moderada e alta. A seguir encontram-se as descrições de cada experimento, bem como a análise dos resultados obtidos.

6.3.1 Caso I: Dimensionalidade baixa

Para os experimentos desse estudo de caso, foi utilizada a imagem de Tapajós, no Pará, descrita anteriormente e mostrada na Figura 14a. Foram consideradas 4 classes: Florestas primárias, Florestas secundárias novas, Florestas secundárias antigas e Atividades recentes. Nesse caso, os resultados da extração de atributos foram avaliados utilizando o método *holdout* para estimar o erro na classificação supervisionada (o conjunto de treinamento contendo 223104 amostras e o conjunto de teste contendo 666995 amostras são disjuntos), além do coeficiente Kappa, calculado a partir da matriz de confusão. A definição teórica desse coeficiente assume a independência dos pixels, porém é conhecido que pixels selecionados como amostras possuem certa dependência espacial, de modo que nas aplicações é usual desconsiderar tal condição. Uma visualização da imagem completa e das regiões de interesse (ROI's) consideradas para obtenção das amostras pode ser vista na Figura 17.

***Tabela 3**. Definições sobre os conjuntos de amostras de treinamento e testes.*

Classes	*Amostras Treinamento*	*Amostras Teste*
Floresta primária	(Vermelho) – 197905 pixels	(Rosa) – 619750 pixels
Floresta Secundária Antiga	(Azul Escuro) – 5085 pixels	(Azul Claro) – 10384 pixels
Floresta Secundária Nova	(Amarelo) – 10705 pixels	(Laranja) – 21249 pixels
Atividades Recentes	(Verde Claro) – 9409 pixels	(Verde Escuro) – 15612 pixels

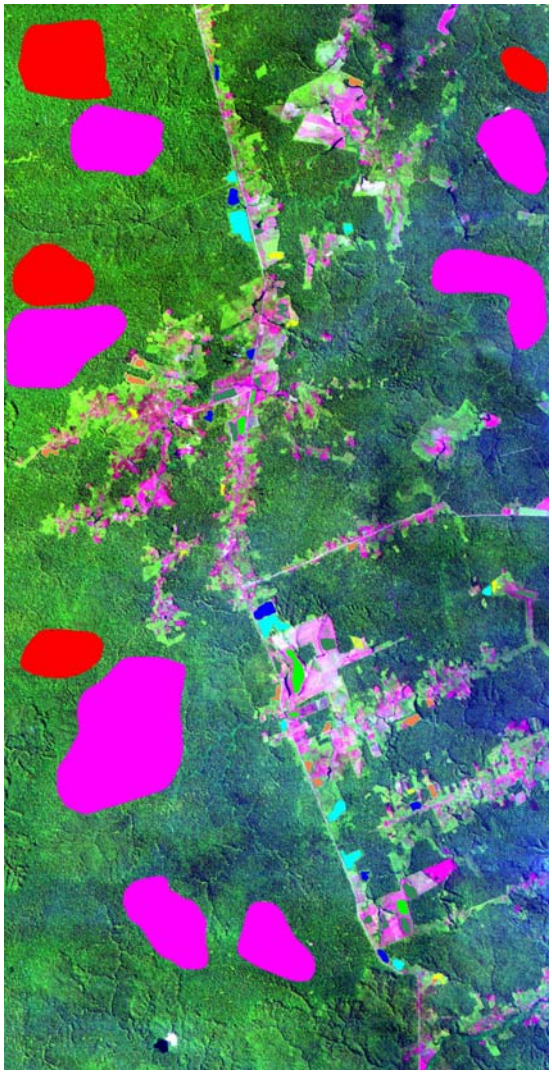


Figura 17. *Regiões de interesse para obtenção das amostras de teste e treinamento.*

Note que o número de amostras disponíveis da classe Floresta primária é muito maior que o número de todas as outras classes juntas, correspondendo a aproximadamente 88% do total das amostras de treinamento e 93% do total das amostras de teste.

A idéia desse experimento consiste basicamente em verificar o desempenho da classificação para cada um dos métodos tradicionais (PCA e LDA) e comparar com o resultado obtido pela utilização do esquema de fusão hierárquica supervisionada e não supervisionada. Mais especificamente, o objetivo é comparar as abordagens não supervisionadas (PCA versus PCA+ICA) juntamente com as supervisionadas (LDA versus PCA+ICA+LDA) tanto para o caso unidimensional quanto bidimensional.

Após a aplicação de cada método de extração de atributos individual, foi testado o desempenho da fusão de atributos, através da fusão hierárquica supervisionada, envolvendo todos os tipos de métodos (PCA, ICA e LDA) e não supervisionada (apenas

PCA e ICA). O algoritmo ICA utilizado foi implementado através da abordagem de máxima verossimilhança, pois foi o método que apresentou melhor desempenho e convergência, além de resultados mais estáveis, no que diz respeito a diferentes subespaços-PCA (tamanho do subconjunto de atributos obtidos com PCA). Para a obtenção dos resultados, mostrados nas Tabelas 4 e 5, a metodologia adotada foi definida da seguinte maneira: nos casos tradicionais (apenas PCA ou LDA) é trivial, ou seja, os métodos são diretamente aplicados e os atributos são obtidos. Já no caso unidimensional, para o esquema não supervisionado, a partir do conjunto original de três atributos, foi aplicado PCA para se obter um novo subconjunto de dois atributos, sobre o qual foi aplicado ICA para gerar o atributo resultante. No esquema supervisionado, primeiramente foi aplicado PCA no conjunto de atributos original para se obter três novos atributos, sobre os quais foram aplicados ICA para obter um novo subconjunto com dois atributos e finalmente LDA para gerar o conjunto resultante com apenas um atributo.

No caso bidimensional, para o esquema não supervisionado a partir do conjunto original de três atributos, foi aplicado PCA para se obter um novo subconjunto com três atributos, sobre o qual foi aplicado ICA para gerar outro subconjunto com dois atributos. Analogamente, no esquema supervisionado, a partir do subconjunto de atributos ICA é aplicado LDA para se obter o subconjunto resultante com dois atributos. Deve-se notar que no caso 2-D, a fusão supervisionada teve desempenho igual à não supervisionada, pois como os dois atributos obtidos através de LDA são ortogonais, embora sejam diferentes dos atributos ICA, geram o mesmo espaço de atributos. Além disso, devido ao reduzido número de atributos, não é possível a realização de diferentes combinações, ou seja, o número de etapas envolvidas na combinação é maior que o numero de atributos e não há possibilidade de redução de dimensionalidade a cada etapa.

Finalmente, o procedimento adotado para a classificação supervisionada utilizou um classificador *maxver*, com diferentes vetores média e matrizes de covariância para cada classe (estimação de parâmetros por máxima verossimilhança), sob hipótese de classes gaussianas multivariadas. Para o cálculo do erro, cada amostra do conjunto de teste é classificada e o número de amostras erroneamente classificadas é normalizado pelo total de amostras, fornecendo uma estimativa do erro de classificação. Os erros estimados de classificação (*holdout*) e os coefficientes Kappa obtidos são apresentados nas Tabelas 3 e 4. Para o caso 3-D, tanto os erros quanto os coeficientes Kappa são

idênticos para todos os métodos (mesmo espaço de atributos) com $\varepsilon_{3-D} = 0.0789$ e $k_{3-D} = 0.5390$.

***Tabela 4**. Comparação entre erros de classificação para métodos de extração de atributos.*

Método	Dimensionalidade	
	1-D	2-D
PCA	0,0665	0,0446
LDA	0,1062	0,0871
Fusão Hierárquica Não Supervisionada (PCA, ICA)	0,0531	0,0436
Fusão Hierárquica Supervisionada (PCA,ICA,LDA)	0,0496	0,0436

***Tabela 5**. Comparação entre coeficientes Kappa para métodos de extração de atributos.*

Método	Dimensionalidade	
	1-D	2-D
PCA	0,3217	0,6501
LDA	0,4423	0,5096
Fusão Hierárqiuca Não Supervisionada (PCA, ICA)	0,5370	0,6555
Fusão Hierárquica Supervisionada (PCA,ICA,LDA)	0,6081	0,6555

É interessante notar que nesse caso o coeficiente Kappa é mais sensível às alterações no espaço de atributos do que o erro estimado (*holdout*), fornecendo diferenças um pouco mais significativas entre os métodos de extração, sendo que representa um índice mais compatível com os resultados obtidos nas imagens temáticas. Provavelmente, isso ocorre devido ao fato de que o coeficiente Kappa penaliza erros pequenos porém críticos, por causa da presença de zeros na diagonal da matriz de confusão. Por exemplo, no caso PCA unidimensional, o classificador apenas detecta duas das quatro classes existentes na imagem, mas como o número de amostras dessas classes não identificadas corresponde a uma pequena quantidade (menos de 10% do total de amostras), o erro é pequeno, ou seja, mesmo que todas as amostras sejam classificadas como floresta, ainda assim o erro estimado é pequeno (se acertar apenas

floresta, corresponde a 93% dos pixels de teste).

Imagens temáticas do resultado da classificação no caso unidimensional para as abordagens supervisionadas (LDA e fusão hierárquica supervisionada) e não supervisionada (PCA e fusão hierárquica não supervisionada) são apresentadas nas Figuras 18, 19, 20 e 21.

Uma comparação da capacidade discriminante dos atributos obtidos no pior (PCA) e melhor caso (fusão hierárquica supervisionada) é mostrada nas Figuras 22a e 22b, através dos respectivos espaços de atributos unidimensionais (densidades condicionais). As densidades condicionais de cada classe foram estimadas através de um método não paramétrico (parzen-window) e suavizadas pela convolução com uma função gaussiana padrão. Pode-se verificar claramente que a separabilidade entre classes é bem melhor no segundo caso.

Portanto, observando os resultados obtidos, é verificado que houve uma certa melhora no desempenho da classificação tanto no caso supervisionado (LDA versus fusão hierárquica supervisionada) quanto no caso não supervisionado (PCA versus fusão hierárquica não supervisionada). Assim, espera-se que a abordagem proposta possa ser útil em processos de extração de atributos, produzindo atributos mais discriminantes.

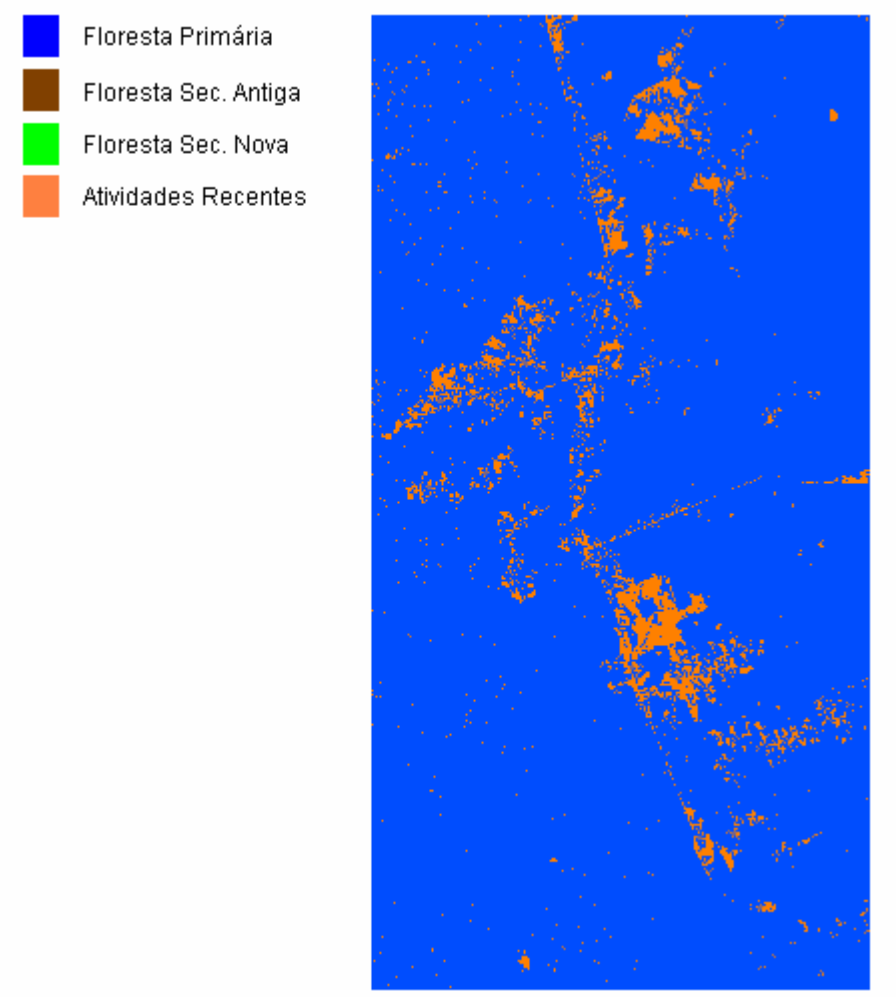


***Figura 18**. Resultado da classificação da imagem de Tapajós utilizando PCA durante a extração de atributos (Método não supervisionado) para o caso 1-D.*

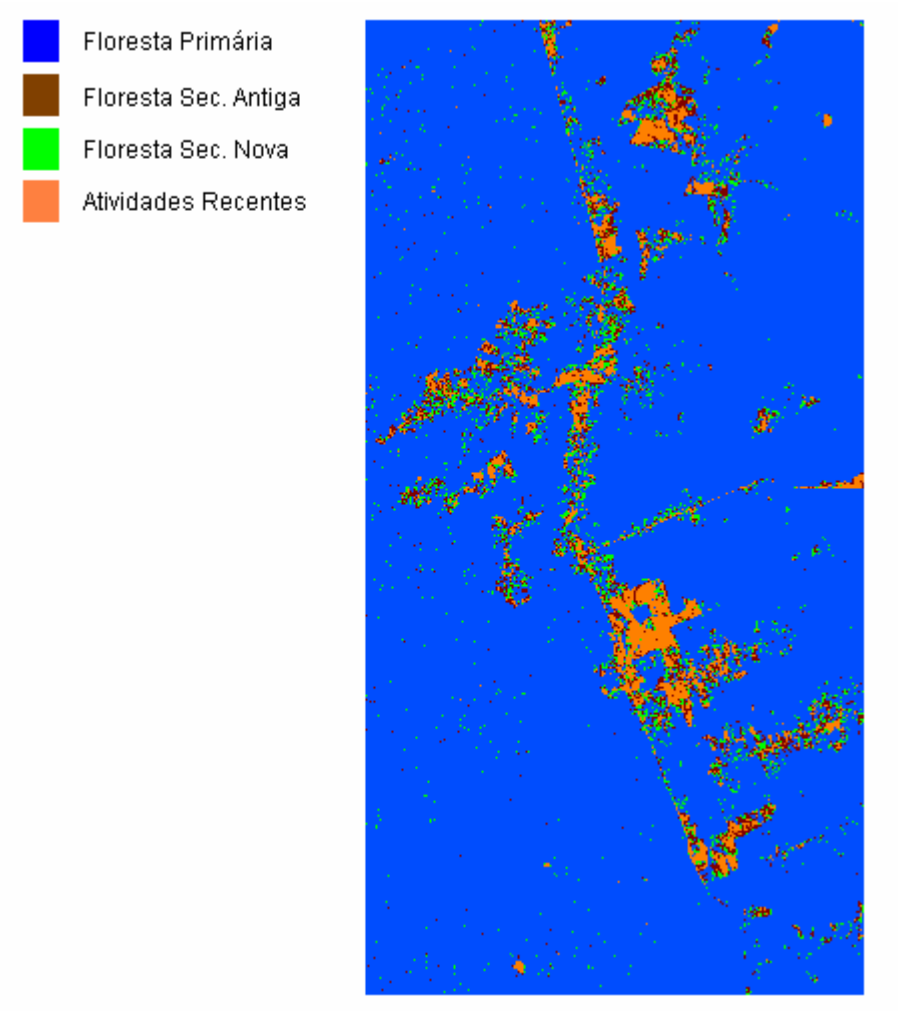


***Figura 19**. Resultado da classificação da imagem de Tapajós utilizando esquema hierárquico não supervisionado com PCA e ICA durante a extração de atributos para o caso 1-D.*

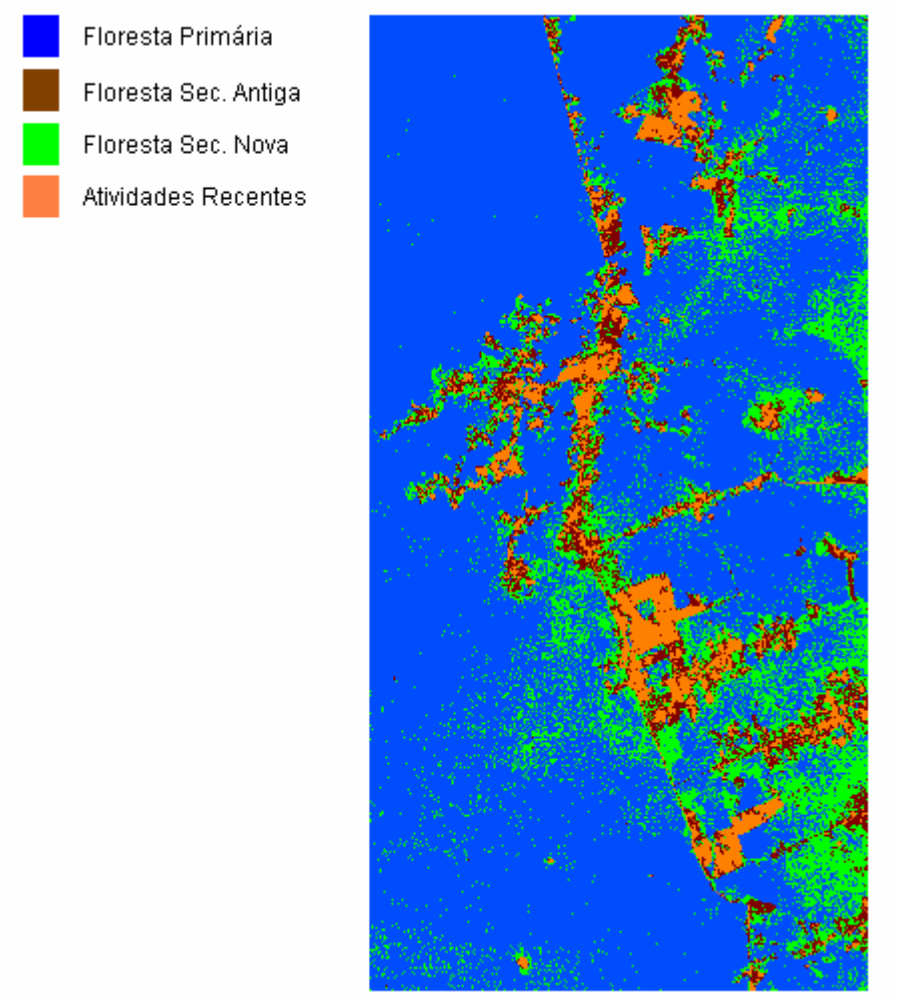


***Figura 20**. Resultado da classificação da imagem de Tapajós utilizando LDA durante a extração de atributos (Método supervisionado) para o caso 1-D.*

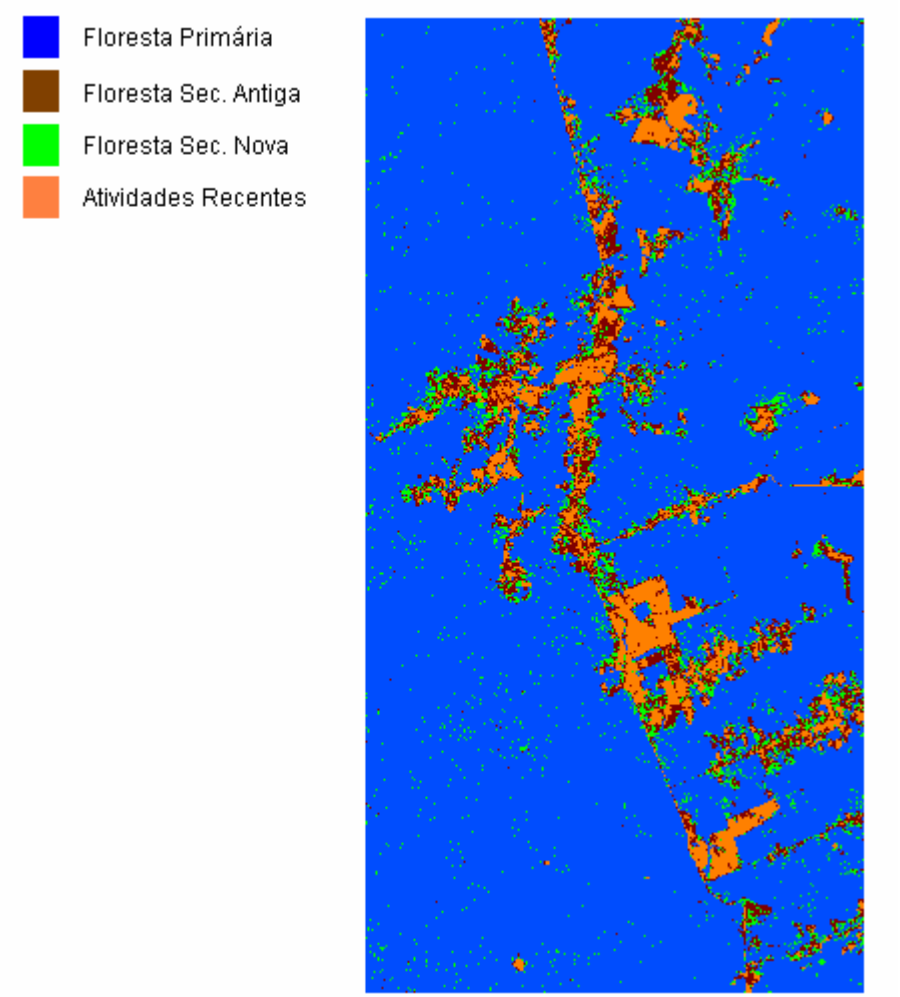
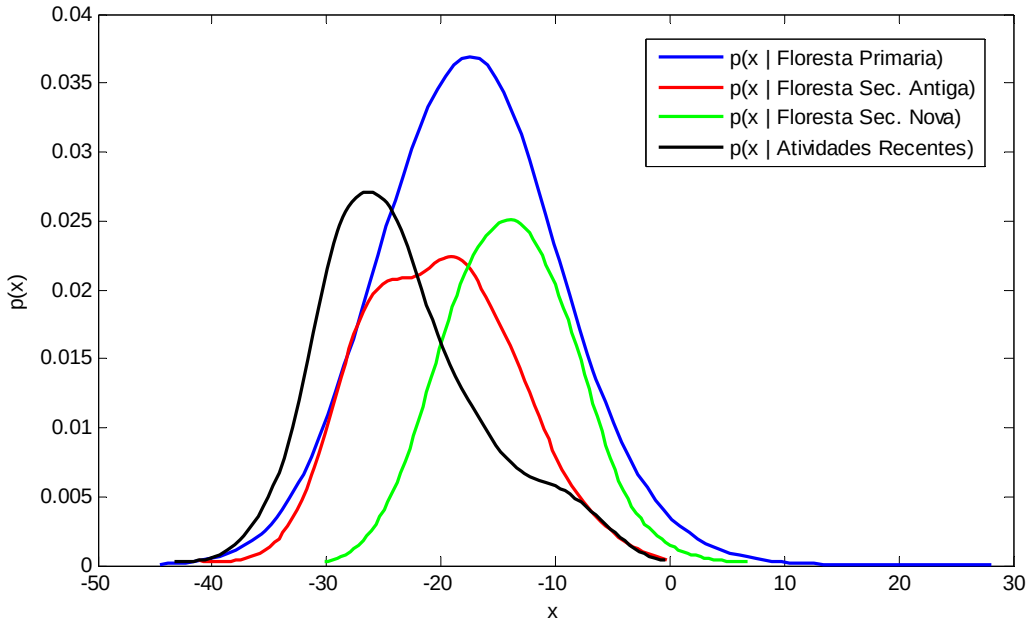
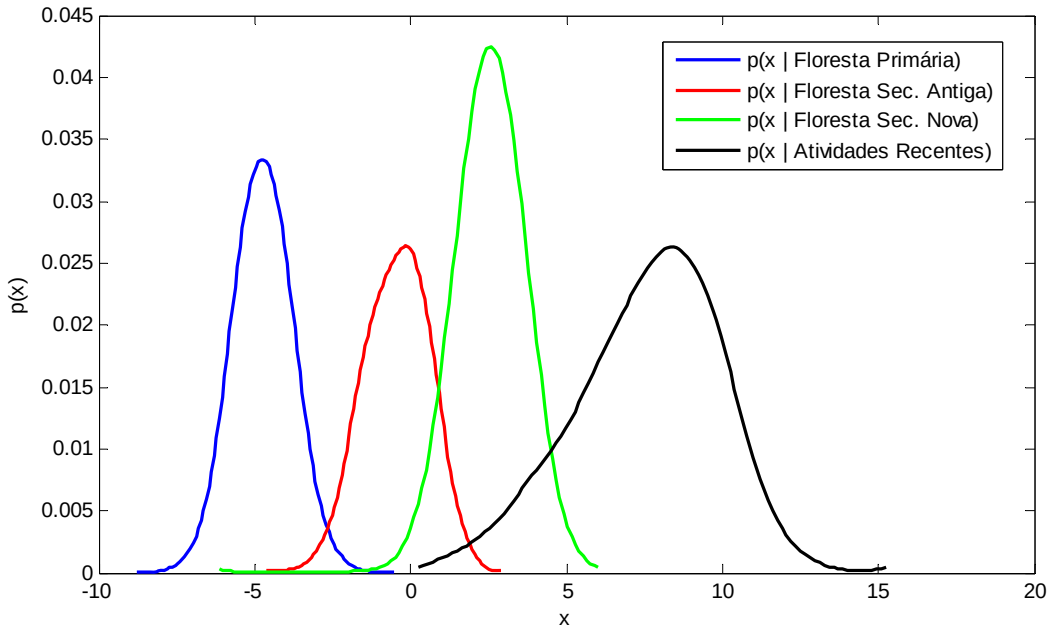


Figura 21. *Resultado da classificação da imagem de Tapajós utilizando esquema hierárquico supervisionado com PCA, ICA e LDA durante a extração de atributos para o caso 1-D.*



a)



b)

Figura 22*. Densidades condicionais estimadas para extração de atributos (Caso 1-D).*
a*. PCA (Não supervisionado).*
b*. Fusão hierárquica supervisionada.*

6.3.2 Caso II: Dimensionalidade moderada

Para o segundo experimento, os dados multivariados foram extraídos da imagem *Flc1.lan*, apresentada na Figura14b. Basicamente, foi adotada a mesma metodologia utilizada anteriormente: classificação *maxver* sob hipótese gaussiana, método *hodldout* (conjunto de treinamento com 14414 amostras e conjunto de teste com 70588 amostras são disjuntos), estimação ICA por máxima verossimilhança e esquema de fusão hierárquica supervisionada e não supervisionada. A idéia foi repetir o experimento anterior para dados com dimensionalidade moderada e verificar o desempenho em cada caso.

Uma observação importante para esse experimento é que o tamanho ótimo do subconjunto de atributos PCA (dimensionalidade ótima do subespaço-PCA), N_{PCA}, é obtida de maneira empírica, difere de caso para caso e depende diretamente do número desejado de componentes independentes estimados, N_{ICA}. Análises experimentais com os dados extraídos das imagens de sensoriamento remoto indicam que os melhores resultados ocorrem quando N_{PCA} e N_{ICA} são similares. O que se deve ter em mente é que em cada caso deve-se tentar buscar um equilíbrio entre dois critérios: representatividade dos dados (favorece dimensionalidade alta) e relação entre os valores próprios (favorece dimensionalidade baixa), ou seja, o objetivo é reter o máximo de informação representativa possível, porém evitando que a relação entre os valores próprios máximo e mínimo seja muito elevada (pode degradar desempenho).

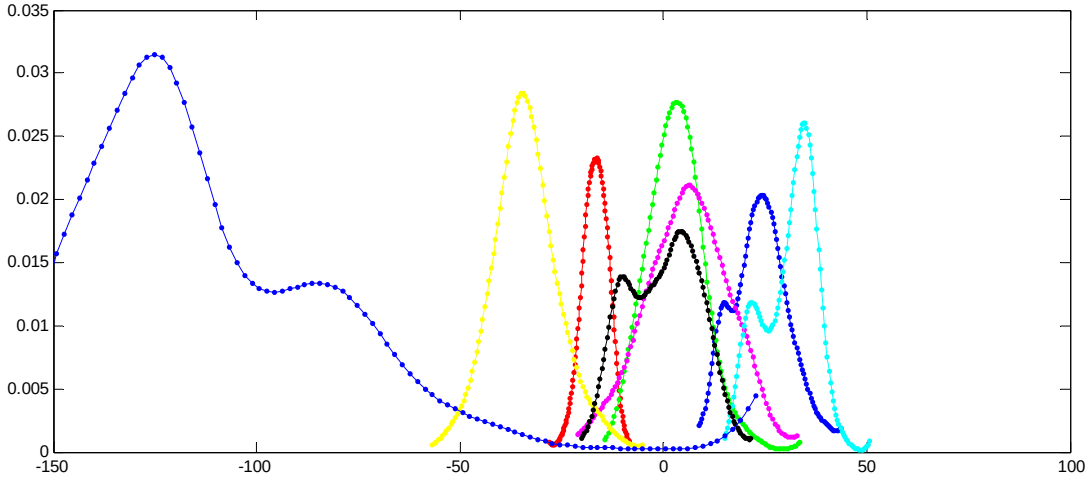
Os resultados apresentados na Tabela 5, mostram concordância com o caso anterior, ou seja, em alguns casos pode ser observada uma melhoria em relação aos casos supervisionados e não supervisionados padrão, PCA e LDA, através do esquema de fusão proposto (tanto supervisionada quanto não supervisionada). Os valores na Tabela 5 denotados por X correspondem aos casos em que o número de atributos é menor que $c-1$, onde c denota o número de classes e não é possível a aplicação do método LDA. Em particular, os coeficientes Kappa obtidos na fusão não supervisionada para subconjuntos de atributos de tamanho maior ou igual a 6, são estatisticamente equivalentes, tendo em vista os valores obtidos para as variâncias dos estimadores, $\sigma_{k_i}^2$ para $i = 6,\ldots,n$. Isso é um sinal de que após certo número de atributos (dimensionalidade intrínseca), o desempenho da classificação tende a se estabilizar, pois os dados se encontram fortemente concentrados num subespaço do espaço de atributos.

Em geral, o uso de ICA e combinação geram melhores atributos do ponto de vista de separabilidade entre classes. Gráficos ilustrando os espaços de atributos unidimensionais em cada caso foram obtidos pelo mesmo processo descrito anteriormente e são apresentados nas Figuras 23 e 24. Imagens temáticas referentes a cada um dos métodos de extração aplicados podem ser visualizadas na Figura 25, que inclui resultados para os casos com dois e três atributos. É possível notar que, embora sempre seja utilizado o mesmo classificador, o efeito da etapa de extração de atributos afeta diretamente o desempenho da classificação.

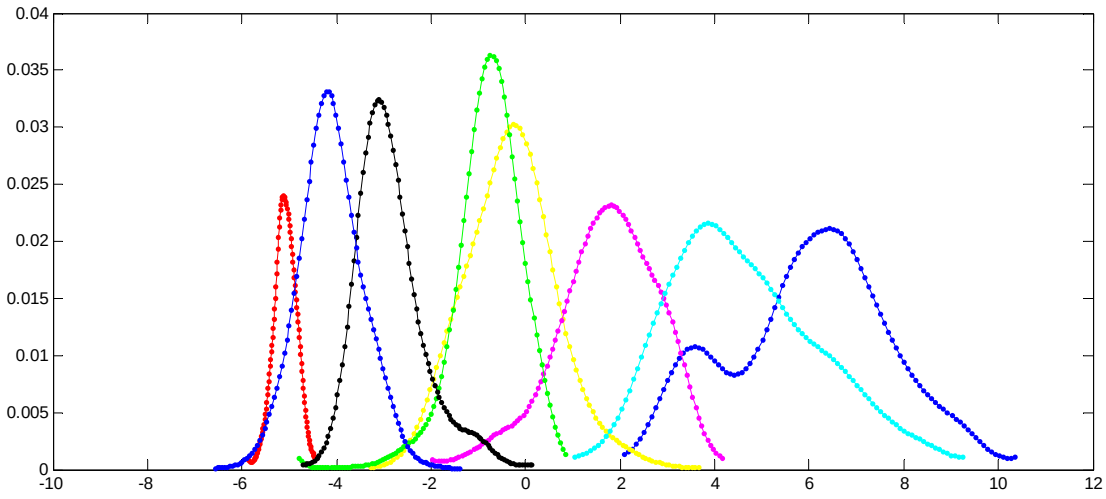
Nesse caso, também é interessante notar a influência dos efeitos negativos do mal-condicionamento presente nos diversos algoritmos ICA. Uma medida comum para esse tipo de comportamento é o cálculo do número de condição $C = \lambda_{Max} / \lambda_{Min}$ da matriz de covariância. Quanto maior a dimensionalidade do subespaço-PCA, ou seja, o número de atributos obtidos durante a etapa PCA, maior é o efeito negativo causado. Para os casos 4-D, 7-D e 12-D os valores de C_4, C_7 e C_{12} são respectivamente 78.3543, 264.1790 e 1467.7272, como pode ser visualizado na Figura 26. Pode ser visto que o desempenho da classificação é degradado na presença de mal-condicionamento, ou seja, o coeficiente Kappa para um dado subconjunto de d atributos, para d fixo, varia conforme se obtém mais ou menos atributos na etapa PCA. Para ilustrar, considere o caso em que $d = 4$. No primeiro caso o coeficiente Kappa é aproximadamente $0,9$ enquanto no segundo é próximo de $0,8$ e no terceiro chega a $0,65$. Ou seja, é preciso utilizar métodos de estimação ICA de maneira adequada, caso contrário o desempenho é bastante prejudicado.

***Tabela 6**. Desempenho da classificação para diferentes métodos de extração de atributos.*

Dimensionalidade	*PCA*		*Fusão não supervisionada*		*LDA*		*Fusão supervisionada*	
	Erro	*Kappa*	*Erro*	*Kappa*	*Erro*	*Kappa*	*Erro*	*Kappa*
1	0,5666	0,3097	0,3701	0,5371	0,3909	0,5119	0,3562	0,5587
2	0,3666	0,5603	0,1826	0,7834	0,1532	0,8240	0,1395	0,8265
3	0,0871	0,8925	0,0801	0,8991	0,1239	0,8456	0,1004	0,8712
4	0,0821	0,8966	0,0769	0,9031	0,1113	0,8611	0,0789	0,9006
5	0,0788	0,9008	0,0739	0,9067	0,1003	0,8747	0,0711	0,9103
6	0,0778	0,9019	0,0713	0,9101	0,0986	0,8768	0,0702	0,9116
7	0,0764	0,9038	0,0705	0,9110	0,0672	0,9146	0,0681	0,9148
8	0,0711	0,9105	0,0687	0,9134	X	X	X	X
9	0,0682	0,9139	0,0666	0,9159	X	X	X	X
10	0,0663	0,9168	0,0656	0,9173	X	X	X	X

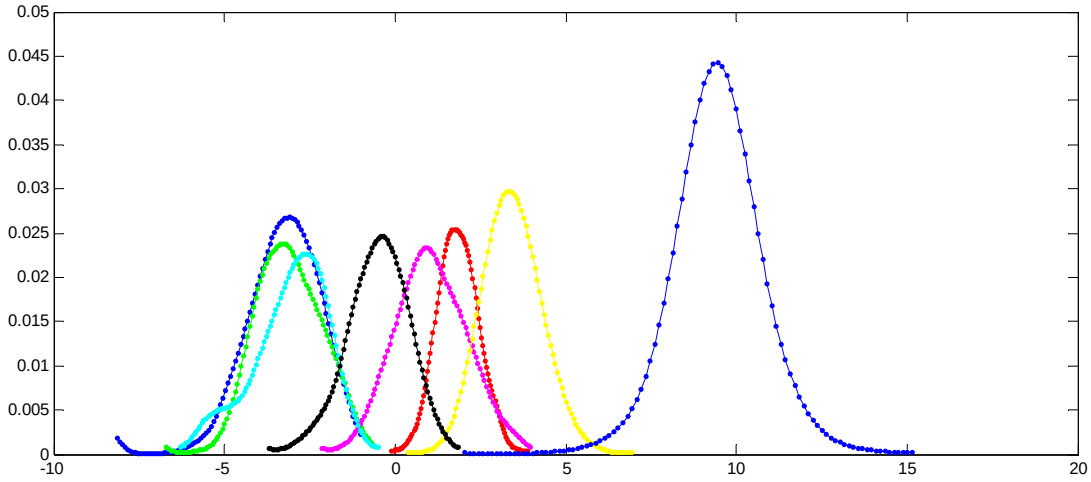


a)

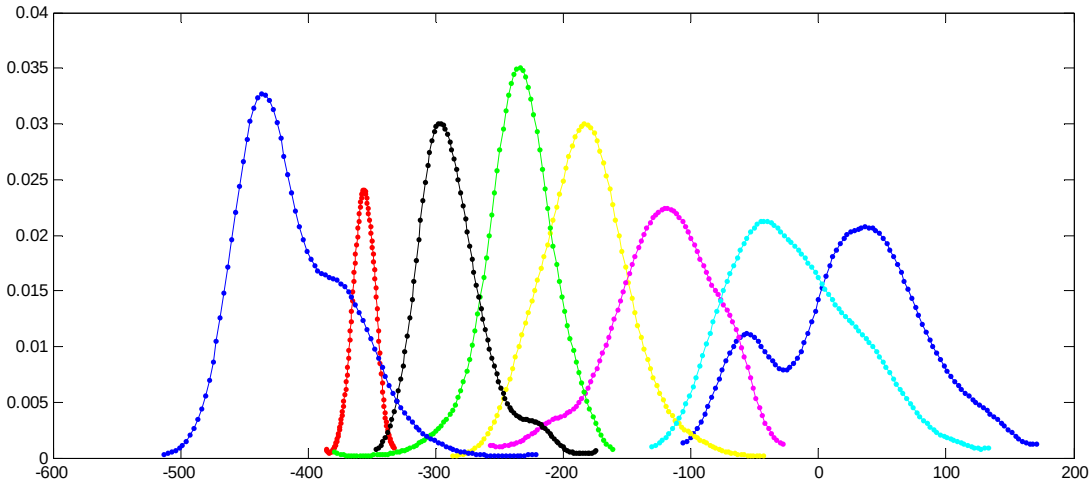


b)

Figura 23. *Densidades condicionais estimadas para diferentes métodos extração de atributos.*
a. *PCA.*
b. *Fusão Hierárquica Não Supervisionada.*

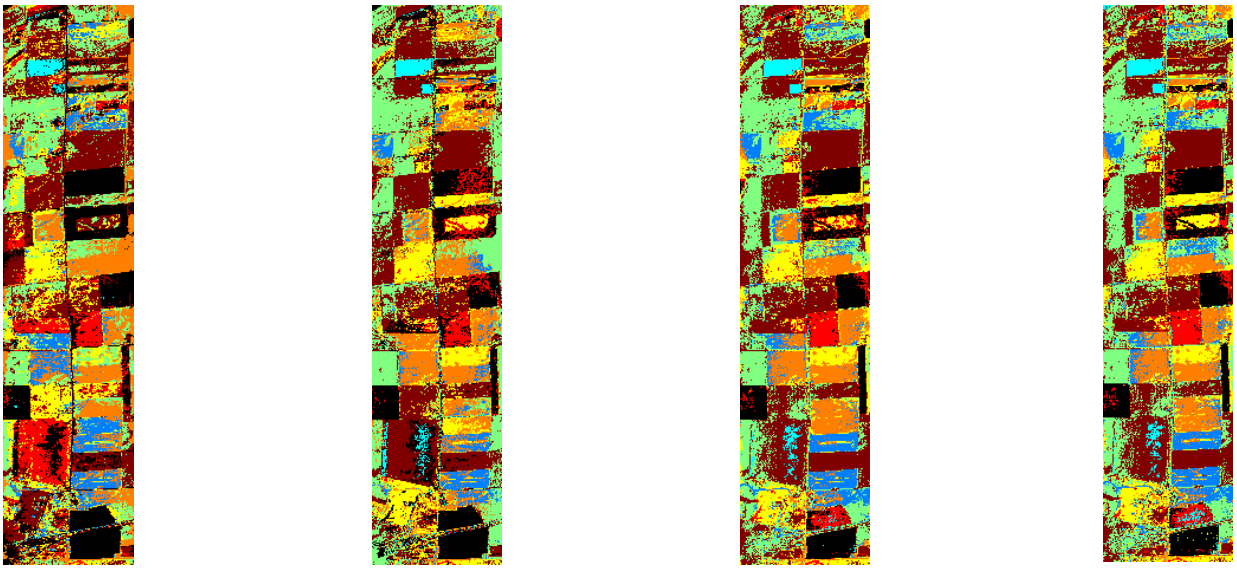


a)

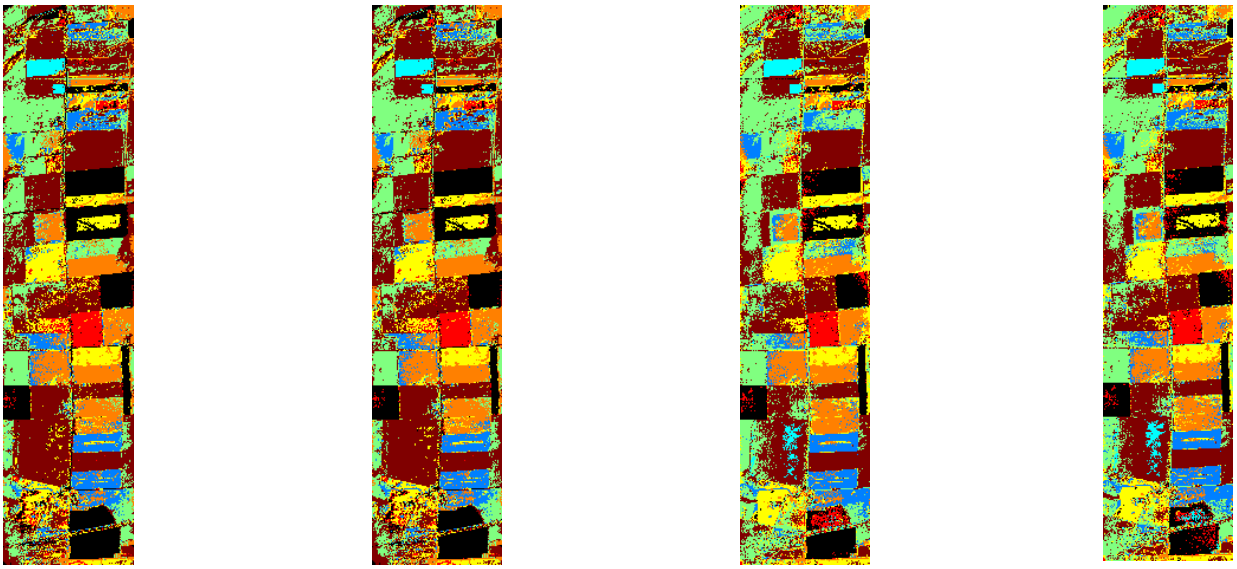


b)

Figura 24. *Densidades condicionais estimadas para diferentes métodos extração de atributos.*
a. *LDA.*
b. *Fusão Hierárquica Supervisionada.*



a) ***b)*** ***c)*** ***d)***



e) ***f)*** ***g)*** ***h)***

Figura 25*. Resultados da classificação para métodos de extração de atributos.*
a*. PCA 2-D.*
b*. Fusão Hierárquica Não Supervisionada 2-D.*
c*. LDA 2-D.*
d*. Fusão Hierárquica Supervisionada 2-D.*
e*. PCA 3-D.*
f*. Fusão Hierárquica Não Supervisionada 3-D.*
g*. LDA-3D.*
h*. Fusão Hierárquica Supervisionada 3-D.*

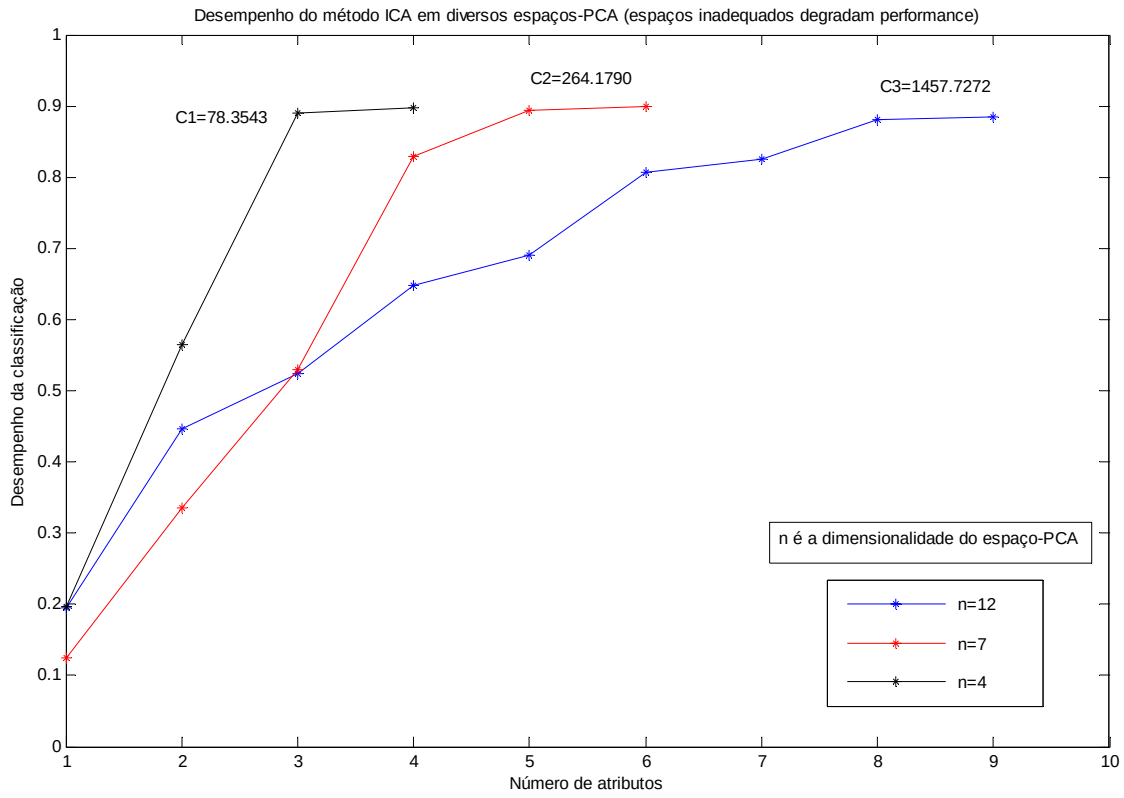


***Figura 26**. Efeito negativo do mal- condicionamento presente na matriz de branqueamento no desempenho da classificação.*

6.3.3 Caso III: Dimensionalidade alta

Para as análises e resultados seguintes, os dados multivariados utilizados nos experimentos foram extraídos da imagem hiperespectral (Figura 14c). Uma das características particulares dessa imagem é o elevado número de bandas (220), o que dificulta bastante a classificação (número de amostras torna-se insuficiente), aumentando ainda mais a importância de um processo extração de atributos. Além disso, verifica-se a presença de ruído em diversas bandas da imagem, de forma que os dados ruidosos são incorporados em todas as análises realizadas, ou seja, tanto na extração de atributos quanto na classificação. A Figura 27 mostra algumas bandas muito ruidosas da imagem.

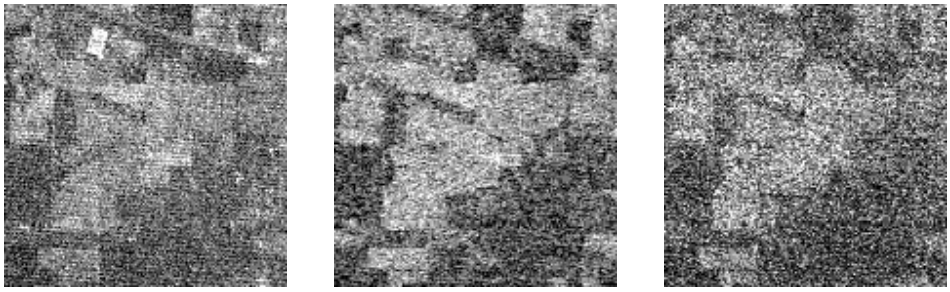


Figura 27. *Bandas ruidosas da imagem hiperespectral (Banda 1, Banda 109 e Banda 163).*

Atualmente, métodos ICA têm sido utilizados com sucesso em separação de sinais, ou *Blind Signal Separation* (BSS). Muitas aplicações utilizam algoritmos ICA para separar sinal e ruído, dado que, em geral, as propriedades estatísticas do sinal são diferentes das propriedades estatísticas do ruído, frequentemente modelado como gaussiano. Assim, a idéia é que com a presença de métodos ICA durante a extração de atributos exista a possibilidade de se reter menos ruído no conjunto de atributos resultante, ou seja, alguns atributos correspondentes a ruído sejam desconsiderados.

Basicamente, a metodologia adotada no primeiro experimento realizado consiste em aplicar o esquema de fusão hierárquica não-supervisionada, com PCA e ICA, e o esquema supervisionado, com PCA, ICA e uma posterior etapa de seleção de atributos utilizando o critério baseado em medidas de distância entre classes através de matrizes de espalhamento definido na equação (4.19). A não utilização de LDA se justifica pela séria limitação quanto ao número de atributos. No segundo experimento, a idéia é justamente tentar superar essa dificuldade através de um esquema de fusão concatenada, utilizando ICA e LDA. Maiores detalhes sobre cada experimento são fornecidos juntamente com os resultados obtidos, a seguir.

Para todos os experimentos que utilizaram os dados da imagem hiperespectral em questão, os erros de classificação foram estimados pelo método de validação cruzada (*leave-one-out cross-validation*) devido ao pequeno número de amostras de treinamento disponíveis para a maioria das classes, com o intuito para obter uma estimativa não tão otimista quanto no método *resubstitution* (mesmo conjunto para treinamento e teste).

A Figura 28 mostra os erros estimados de classificação para o primeiro experimento (fusão hierárquica não supervisionada). Para as curvas de PCA e LDA, o processo é trivial. Já o procedimento adotado para o esquema proposto foi, a partir do conjunto

original com 220 atributos, obter um subconjunto com 65 atributos utilizando-se o método PCA. Em seguida, a partir desse subconjunto-PCA foi aplicado ICA para se obter um novo subconjunto com 55 atributos. Assim, o erro de classificação foi estimado para todos os métodos, em cada dimensionalidade d (número de atributos) variando de 1 a 50.

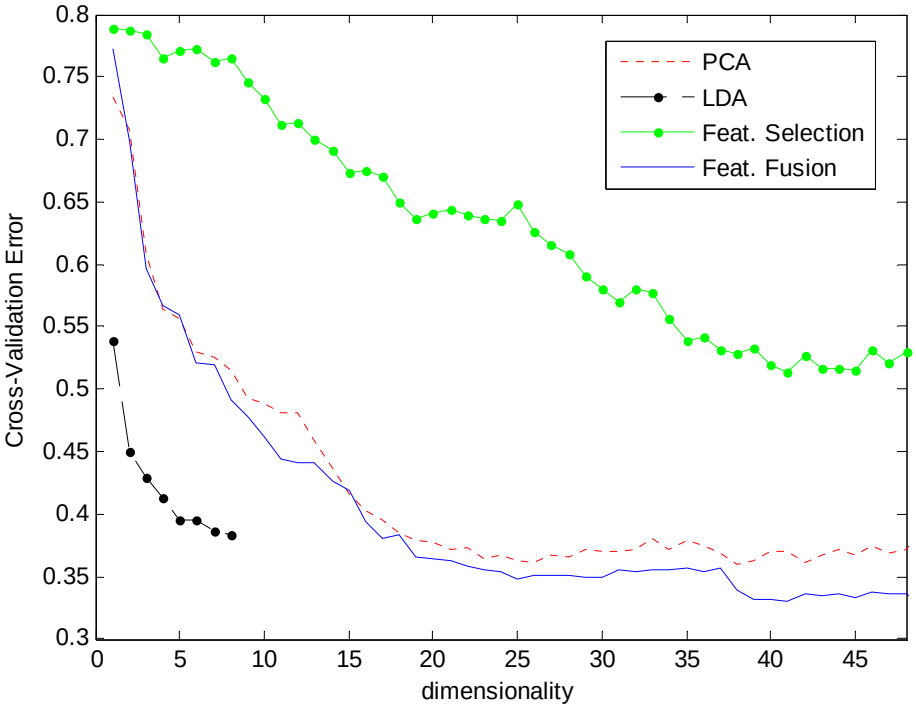


Figura 28. *Comparação entre erros de classificação para métodos de extração de atributos PCA, LDA, seleção de atributos e Fusão hierárquica não supervisionada.*

Para o esquema supervisionado, a idéia consiste em, a partir do subconjunto de 55 atributos ICA, aplicar um método de seleção de atributos com critério baseado em matrizes de espalhamento (vide equação (4.19)) e busca baseada no critério *Best Individual Features*, que não garante o subconjunto ótimo de atributos em cada dimensionalidade. O resultado obtido, ilustrado na Figura 29, mostra uma leve melhora no coeficiente Kappa para alguns subconjuntos de atributos (intervalo de 30 a 48 atributos). Nessa escala, pode-se notar que a curva de desempenho da classificação do PCA oscila mais que os outros casos, sendo que uma possível interpretação para esse fato pode ser a maior presença de ruído nos atributos, o que causa confusão na classificação dos dados.

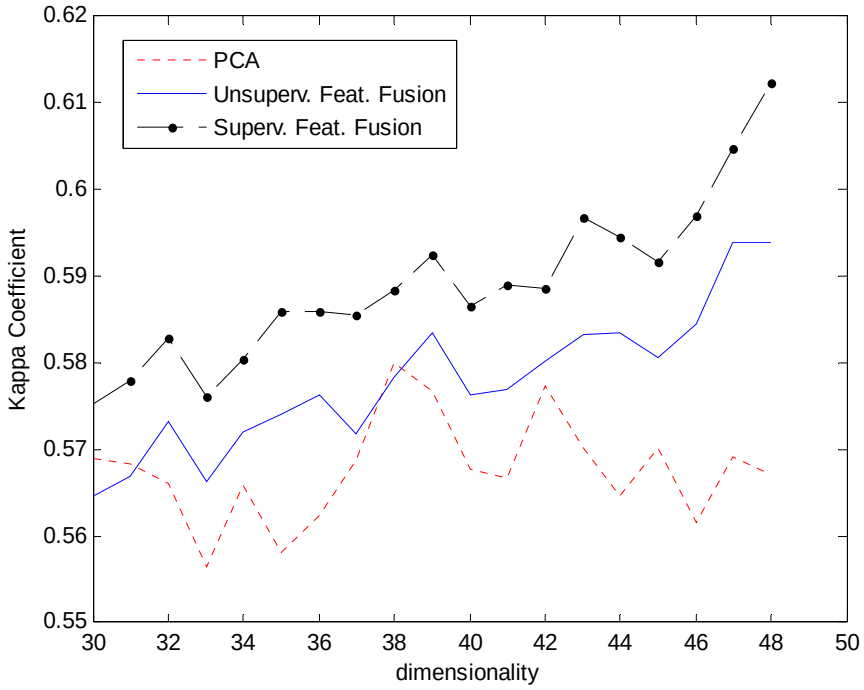


Figura 29. *Comparação do desempenho da classificação para métodos de extração de atributos PCA e Fusão hierárquica supervisionada e não supervisionada*

No segundo experimento, foi testado o esquema de fusão concatenada utilizando-se ICA e LDA. Nesse caso, o desempenho da classificação para a extração com ICA é significantemente degradado, pois os atributos foram obtidos diretamente do conjunto original (sem PCA). A idéia foi, a partir do conjunto original de atributos, gerar um subconjunto A de 40 atributos aplicando-se ICA no conjunto original de atributos, e um outro subconjunto B de 9 atributos obtido pela aplicação de LDA no subconjunto original de atributos. A seguir, os dois conjuntos de atributos, A e B, são concatenados para formar um único vetor de atributos com dimensionalidade 49. Um método de seleção de atributos idêntico ao citado anteriormente é utilizado para escolher os melhores atributos. O resultado obtido é mostrado a seguir na Figura 30. Nesse caso, combinando-se as curvas ICA e LDA foi possível obter a linha tracejada, indicando uma certa melhora.

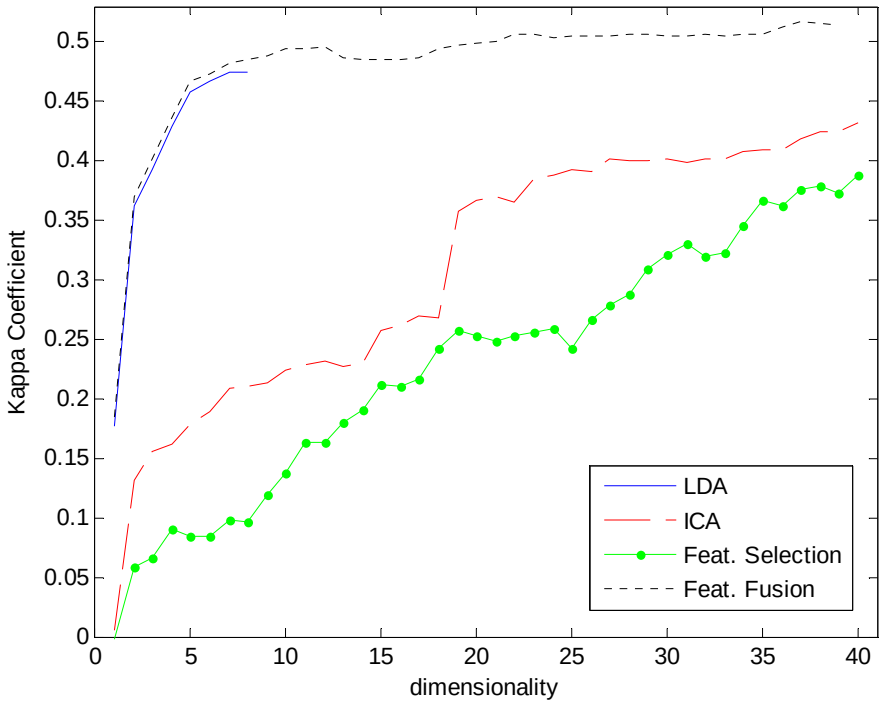


Figura 30. *Comparação entre o desempenho da classificação para métodos de extração de atributos PCA, LDA, seleção de atributos e Fusão concatenada.*

Para verificar como o tamanho do subconjunto de atributos PCA influencia na estimação dos atributos independentes, ou seja, como a dimensionalidade do subespaço-PCA atua na estimação ICA, foi realizada uma comparação. A Figura 31 ilustra o efeito negativo de subespaços PCA inadequados na obtenção dos componentes independentes.

Em resumo, a idéia é verificar para um dado intervalo de dimensionalidades ($d = 50$) como o erro de classificação varia de acordo com diversos subespaços-PCA. Em todos os casos, o ponto de partida é o conjunto original de 220 atributos. Na primeira curva, obtém-se um subconjunto de 80 atributos via PCA e posteriormente, a partir desses atributos, um novo subconjunto de 50 atributos via ICA. O erro estimado de classificação nesse caso, para $d = 50$, é aproximadamente 0,43. Num segundo caso, a partir do conjunto original de atributos, foram obtidos 120 atributos PCA e em seguida, foram estimados, da mesma forma que anteriormente, 50 atributos ICA. O erro de classificação passa a 0,46. Finalmente, no terceiro caso, a partir do conjunto original de atributos, são obtidos 200 atributos PCA e novamente um subconjunto resultante de 50 atributos ICA. Nessa situação, o erro de classificação passa para 0,52 e corresponde a um aumento de aproximadamente 21% em relação ao primeiro caso.

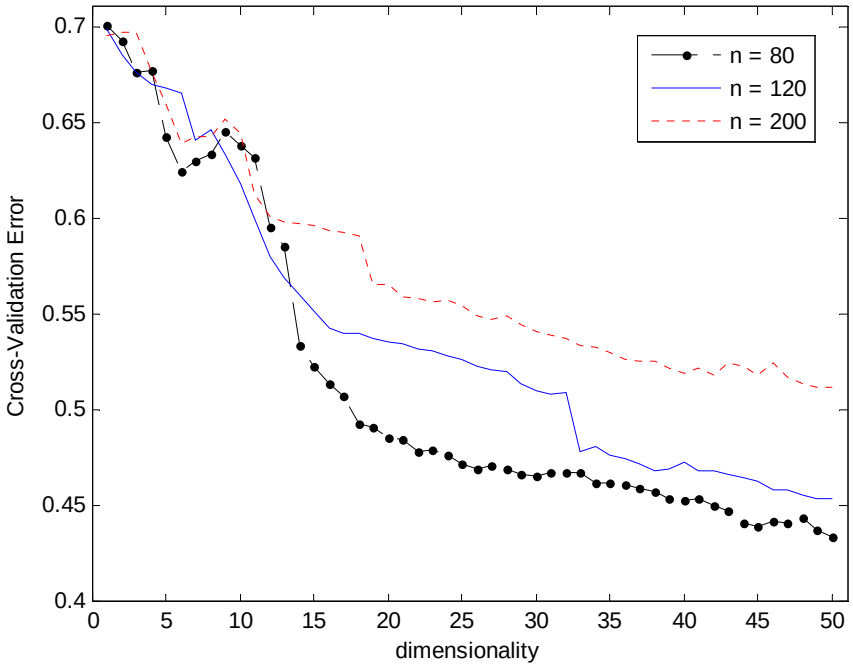


***Figura 301**. Efeito de diferentes subespaços-PCA no desempenho da classificação.*

A análise do problema hiperespectral possui algumas peculiaridades, pois trata-se de um caso bastante problemático. Em primeiro lugar, a medida de mal-condicionamento para a matriz de covariância no caso de considerar-se o espaço definido por todos os atributos (220) é da ordem de, $C_{220} = \lambda_{Max} / \lambda_{Min} \approx 1,5 \times 10^6$, ou seja, trata-se de um valor extremamente alto. Isso ocorre pois, nesse caso, a matriz de covariância $C_{220 \times 220}$ pode ser considerada esparsa, uma vez que possui vários elementos com valores muito próximos de zero, caracterizando ausência de correlação (0.00005) e outros elementos com valores muito próximos de um (0.995), o que caracteriza correlação total. Esses são alguns dos vários problemas causados pelo aumento drástico da dimensionalidade na análise de dados hiperespectrais. Como conseqüência, pode-se verificar o fraco desempenho obtido na classificação dos dados sem nenhuma etapa de extração de atributos, onde o erro estimado de classificação foi $\varepsilon_{220} = 0.8793$ e o coeficiente Kappa obtido foi $k = 0.0357$, caracterizando o desempenho como péssimo. Tal resultado indica a necessidade de se realizar uma boa etapa de extração de atributos. Nesses casos a etapa de extração de atributos é tão importante, senão mais, que a classificação propriamente dita.

O próximo experimento compara o desempenho de métodos estatísticos de ordens superiores (ICA) com métodos de segunda ordem (PCA) e seleção de atributos em espaços de dimensionalidade bastante elevada (hiperespaços). Devido a problemas com a convergência do algoritmo ICA baseado na abordagem de máxima verossimilhança, foi implementada e testada uma versão do método de ponto fixo iterativo baseado na maximização do comportamento não gaussiano.

Os resultados obtidos, apresentados na Figura 32, indicam que a presença de estatística de ordens superiores pode ser útil no processo de extração de atributos, especialmente quando o número de atributos aumenta. Além disso, segundo (NASCIMENTO; DIAS 2005), resultados mostram que esse tipo de informação presente em algoritmos ICA pode ser usado para separar misturas de materiais existentes em imagens hiperespectrais (*hyperespectral unmixing*), onde é realizada a decomposição do espectro dos pixels em uma coleção de espectros conhecidos (assinaturas espectrais) e seus correspondentes percentuais que indicam a proporção de cada material presente.

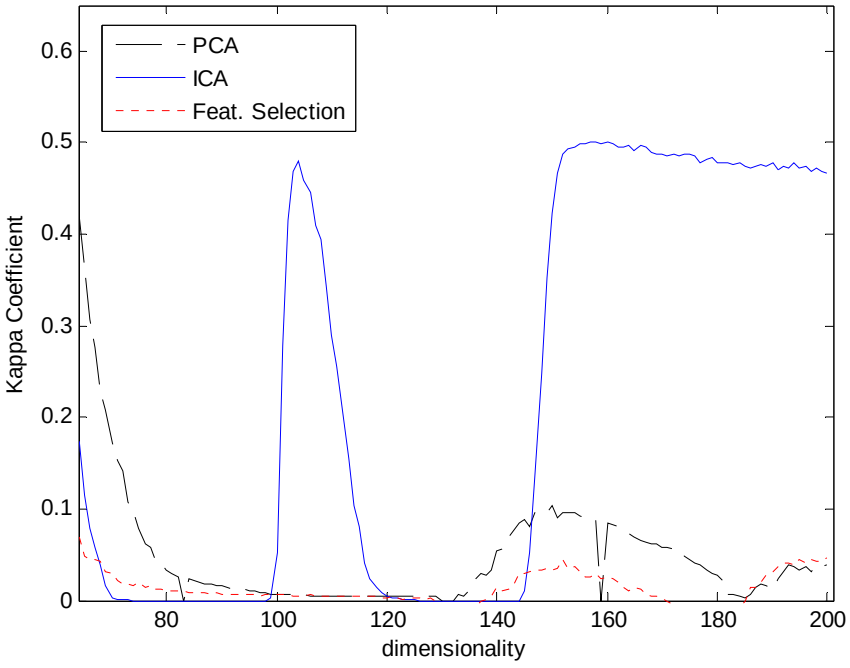


***Figura 312**. Desempenho da classificação para métodos de extração de atributos PCA, ICA e seleção de atributos em hiperespaços (dimensionalidade > 100).*

A curva que indica a performance da classificação utilizando ICA possui uma forma bastante incomum. Isso ocorre provavelmente devido as propriedades geométricas não intuitivas comuns aos dados hiperespectrais, definidas em (LANDGREBE, 2000). Outra possível explicação para essa diferença é a eventual presença de ruído independente em algumas bandas da imagem, retidas durante o processo aplicado na Transformação de Karhunen-Loève (PCA), ou ainda, advém do fato que algoritmos ICA são freqüentemente utilizados em *BSS* (*Blind Signal Separation*) para extrair um sinal a partir de uma mistura.

Além disso, a Transformação de Karhunen-Loève tenta maximizar a variação contida nos componentes transformados de baixa ordem (componentes principais são restritos a segunda ordem), relegando variações menos significantes presentes em componentes de ordens superiores (componentes independentes). Também, de acordo com (LANDGREBE, 1998), para altas dimensionalidades, o formato e a orientação dos dados (i.e, matriz de covariância) são mais significativos para a separação do que a informação sobre a localização propriamente dita (i.e, média), o que significa que estatísticas de segunda ordem são muito mais efetivas do que estatísticas de primeira ordem. Portanto, uma possível conclusão que se pode obter é que para melhorar ainda mais o desempenho da classificação, deve-se considerar a utilização de estatísticas de ordens superiores além de estatísticas de segunda ordem apenas em algoritmos de extração de atributos.

.

6.4 Conclusões

Nesse trabalho foi proposto um modelo de fusão de atributos utilizando métodos baseados em critérios estatísticos para tentar melhorar o desempenho da classificação de imagens de sensoriamento remoto, através da idéia de combinar métodos supervisionados e não supervisionados utilizando as abordagens hierárquica e concatenada. A adição de estatísticas de ordens superiores, através de algoritmos ICA aumentou o desempenho da classificação da abordagem não supervisionada (PCA), e juntamente com os esquemas de fusão de atributos fornece uma alternativa às sérias limitações dos métodos tradicionais. O preço para tal melhoria é o aumento no custo computacional dos algoritmos, visto que os métodos tradicionais como PCA e LDA são derivados de problemas de valores próprios relativamente simples e não necessitam de algoritmos iterativos para encontrar a solução desejada, obtida diretamente, ao contrário dos algoritmos ICA, que utilizam métodos de otimização como gradiente, gradiente conjugado, métodos quasi-Newton, entre outros para gerar algoritmos iterativos que nem sempre convergem em casos reais (dados obtidos através de imagens reais).

Também, foi verificado que, em geral, dados de sensoriamento remoto possuem uma estrutura dimensional inferior, ou seja, os dados possuem uma dimensionalidade d, mas encontram-se fortemente concentrados num subespaço com dimensionalidade m, com $m < d$. Logo, se os dados pertencem a um subespaço do espaço de atributos, é mais fácil restringir possíveis soluções (conjunto de atributos a serem obtidos) apenas ao subespaço em que os dados se encontram e não a todo o restante do espaço.

Um aspecto observado é o fato de que mesmo nos melhores casos, nota-se a presença de variações abruptas nos valores dos pixels classificados em todos os mapas de classes obtidos, mesmo em regiões suaves e homogêneas, resultando em visíveis esparsos erros de classificação. Tal comportamento é típico em classificadores pontuais, como é o caso do critério adotado na classificação *maxver*. Como solução a esse problema, pode-se adotar classificadores que incorporem informações contextuais, como por exemplo, os baseados em modelos de campos aleatórios markovianos.

Embora nos experimentos foram utilizadas somente imagens de sensoriamento remoto, outros tipos de dados multivariados podem ser explorados em variedade de aplicações em reconhecimento de padrões. Finalmente, os resultados obtidos indicaram que o esquema proposto pode obter bons atributos do ponto de vista discriminante, compondo um válido e interessante ferramental para análise de dados multivariados.

6.5 Trabalhos Futuros

Ao longo do período de desenvolvimento do trabalho, novas idéias contemplando diversas possibilidades de aplicações surgiam. Assim, pode ser interessante verificar a viabilidade de algumas possíveis idéias para trabalhos futuros.

Uma idéia consiste na possibilidade de pesquisar de maneira mais ampla os efeitos da extração/fusão de atributos sobre outras abordagens de classificadores (i.e, redes neurais, classificadores não paramétricos, classificadores contextuais) ou até mesmo em combinação de classificadores. Também pode-se aplicar esses métodos em outros tipos de dados, como imagens tomográficas de diversas bandas imagens médicas (ressonância magnética), entre outras.

Outra possibilidade é incorporar novos algoritmos, restrições ou etapas no esquema de combinação/fusão, ou seja, novos algoritmos ICA, métodos lineares generalizados para extração de atributos ou ainda algoritmos mais complexos de seleção de atributos.

Por fim, pode-se perceber que existe um grande potencial em possíveis problemas do tipo detecção, onde existem diversos atributos (imagens multiespectrais ou hiperespectrais), já que nessas ocasiões (detecção corresponde a existência de apenas duas classes, ou $c = 2$) os métodos tradicionais (i.e, em LDA, $d \leq c - 1$) são muito restritos ou sub-ótimos do ponto de vista de classificação.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AMARI, S.; CICHOCKI, A.; YANG, H.; ***A new learning algorithm for blind source separation***. Advances in Neural Information Processing Systems, v.8, p. 757–763, 1996.

BELL, A. J.; SEJNOWSKI, T. J.; ***An information maximization approach to blind separation and blind deconvolution***. Neural Computation, v. 7, p. 1129–1159, 1995.

BLASCHKE, T.; WISKOTT, L.; ***CuBICA: independent component analysis by simultaneous third and fourth-order cumulant diagonalization***. IEEE Transactions on Signal Processing, v. 52, n. 5, p. 1250-1256, 2004.

BORGNE, H. L.; GUÉRIN-DUGUÉ, A.; ANTONIADIS, *A.*; ***Representation of images for classification with independent features***. Pattern Recognition Letters, v. 25, n. 2, p. 141–154, 2004.

BORGNE, H. L. et al.; ***Classification of images: ICA filters vs. human perception***. In: Proceedings of the 7th International Symposium on Signal Processing and its Applications (ISSPA 2003). v. 2, p. 251–254, 2003.

CARDOSO, J. F.; ***Infomax and maximum likelihood for blind source separation***. IEEE Letters on Signal Processing, v. 4, n. 4, p. 112–114, 1997.

CARDOSO, J. F.; ***Higher-order contrasts for independent component analysis***. Neural Computation, v. 1, n. 1, p. 157–192, 1999.

CHEN, A.; BICKEL, P.J.; ***Consistent independent component analysis and prewhitening***. IEEE Transactions on Signal Processing, v. 53, n. 10(1), p. 3625-3632, 2005.

CICHOCKI, A.; UNBERHAUEN, R.; ***Robust neural networks with online learning for blind identification and blind separation of sources***. IEEE Transactions on Circuits and Systems, v. 43, p. 894–906, 1996.

COHEN, J. A.; ***Coefficient of agreement for nominal scales***. Educational and Measurement, v. 20, n. 1, p. 37–46, 1960.

COMON, P.; ***Independent component analysis - a new concept ?***. Signal Processing, v. 36, p. 287–314, 1994.

CONGALTON, R. G.; ***A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data***. Remote Sensing Environment, v. 37, p. 35-46, 1991.

COVER, T. M.; THOMAS, J. A. ***Elements of Information Theory***. Wiley, 1991.

DEVIJVER, P. A., KITTLER, J. ***Pattern Recognition: A Statistical Approach***. Prentice Hall, 1982.

DINH-TUAN PHAM. ***Fast algorithms for mutual information based independent component analysis***. IEEE Transactions on Signal Processing, v. 52, n. 10(1), p. 2690-2700, 2004.

DU, Y.C.; HU, W.C.; SHYU, L.Y.; ***The effect of data reduction by independent component analysis and principal component analysis in hand motion identification***. Proceedings of the 26th Annual International Conference of the IEEE EMBS, v. 1, p. 84-86, San Francisco, CA, USA, September, 2004.

DUDA, R. O.; HART, P. E.; STORK, D. G.; ***Pattern Classification***. Segunda edição, John Wiley & Sons, 2001.

DUIN, R. P. W.; ***PRTools - A Matlab Toolbox for Pattern Recognition***. 2003. Disponível em: <http://www.prtools.org/>.

ERIKSSON, J. KOIVUNEN, V.; ***Identifiability, separability, and uniqueness of linear ICA models***. IEEE Signal Processing Letters, v. 11, n. 7, p. 601-604, 2004.

FREITAS, C. C., SANT'ANNA, S. J. S., RENNÓ, C. D.; ***The use of JERS-1 and RADARSAT images for land use classification in the Amazon region***. In: International Geoscience and Remote Sensing Symposium. Proceedings. Hamburg v.3, 1649-1651, 1999.

FUKUNAGA, K.; ***Introduction to Statistical Pattern Recognition***. Segunda edição, Academic Press, 1990.

GOSE, E.; JOHNSONBAUGH, R.; JOST, S.; ***Pattern Recognition and Image Analysis***. Prentice Hall, 1996.

HANSEN, L. K.; LARSEN, J.; KOLENDA, T.; ***Blind detection of independent dynamic components***. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), v. 5, 2001.

HAYKIN, S.; ***Neural Networks - A Comprehensive Foundation***. Segunda edição, Prentice Hall, 1998.

HERRERO, G.G.; GOTCHEV, A.; CHRISTOV, I.; EGIAZARIAN, K.; ***Feature extraction for heartbeat classification using independent component analysis and matching pursuits***. IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP '05), v. 4, p. 725-728, 18-23 March 2005.

HUGHES, G. F.; ***On the mean accuracy of statistical pattern recognizers***. IEEE Transactions on Information Theory, v. 14, p. 55–63, 1968.

HWANG, J.; LAY, S.; LIPPMAN, A.; ***Nonparametric multivariate density estimation: A comparative study***. IEEE Transactions on Signal Processing, v. 42, n. 10, p. 2795–2810, 1994.

HYVÄRINEN, A.; ***Survey on independent component analysis***. Neural Computing Surveys, v. 2, p. 94–128, 1999.

HYVÄRINEN, A. et al.; ***FastICA for Matlab - Versão 2.3***. 2004. Disponível em: <http://www.cis.hut.fi/projects/ica/fastica>.

HYVÄRINEN, A.; KARHUNEN, J.; OJA, E.; ***Independent Component Analysis***. JohnWiley & Sons, 2001.

INKI, M.; ***Extensions of Independent Component Analysis for Natural Image Data***. Tese de Doutorado - Helsinki University of Technology, 2004.

JAIN, A. K.; DUIN, R. P.W.; MAO, J.; ***Statistical pattern recognition: A review***. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 22, n. 1, p. 4–37, 2000.

JIMENES, L. O.; LANDGREBE, D.; ***Supervised classification in high dimensional space: Geometrical, statistical and asymptotical properties of multivariate data***. IEEE Transactions on Systems, Man and Cybernetics, v. 28, n. 1, p. 39–54, 1998.

JUTTEN, C.; HERAULT, J.; ***Blind separation of sources: An adaptive algorithm based on neuromimetic architecture***. Signal Processing, v. 24, n. 1, p. 1–10, 1991.

KARHUNEN, J. et al.; ***A class of neural networks for independent component analysis***. IEEE Transactions on Neural Networks, v. 8, n. 3, p. 486–504, 1997.

LANDGREBE, D.; ***Information extraction principles and methods for multispectral and hyperspectral image data***. In: GHEN, C. H. (Ed.). Information Processing for Remote Sensing, World Scientific, 1999.

LEE, T. W.; GIROLAMI, M.; SEJNOWSKI, T. J.; ***Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources***. Neural Computation, v. 11, n. 2, p. 417–441, 1999.

MACKAY, D. J. C.; ***Maximum Likelihood and Covariant Algorithms for Independent Component Analysis***. 1999. Disponível em: http://wol.ra.phy.cam.ac.uk/mackay/abstracts/ica.html

MARTINEZ, W. L.; MARTINEZ, A. R.; ***Computational Statistics Handbook with Matlab***. Chapman & Hall, 2002.

NASCIMENTO, J.M.P.; DIAS, J.M.B.; ***Does independent component analysis play a role in unmixing hyperspectral data ?***. IEEE Transactions on Geoscience and Remote Sensing, v. 43, n. 1, p. 175-187, Jan. 2005.

NIKIAS, C.; PETROPULU, *A.*; ***Higher-Order Spectral Analysis: A Nonlinear Signal Processing Framework***. Prentice Hall, 1993.

OJA, E.; ***The nonlinear PCA learning rule in independent component analysis***. Neurocomputing, v. 17, n. 1, p. 25–46, 1997.

PAPOULIS, A.; ***Probability, Random Variables and Stochastic Processes***. Terceira edição, McGraw-Hill, 1991.

ROBILA, S. A.; VARSHNEY, P. K.; ***A fast source separation algorithm for hyperspectral image processing***. In: Proceedings of the IEEE Geoscience and Remote Sensing Symposium (IGARSS), v. 6, 2002.

ROBILA, S. A.; VARSHNEY, P. K.; ***Target detection in hyperspectral images based on independent component analysis***. In: Proceedings of the 7th SPIE Automatic Target Recognition, v. 4726, 2002.

ROSENBLATT, M.; ***Stationary Sequences and Random Fields***. Birkhauser, 1985.

SCOTT, D. W.; ***Multivariate Density Estimation***. John Wiley & Sons, 1992.

THEODORIDIS, S.; KOUTROUMBAS, K.; ***Pattern Recognition***. Segunda edição, Academic Press, 2003.

TRUNK, G. V.; ***A problem of dimensionality: A simple example***. IEEE Transactions on Signal Processing, v. 42, n. 10, p. 2795–2810, 1979.

XU, L.; ***Least mean square error reconstruction principle for self-organizing neural nets***. Neural Networks, v. 6, p. 627–648, 1993.

YANG, M.; AHUJA, N.; ***Face detection and gesture recognition for human-computer interaction***. Kluwer Academic Publishers, 2001.

YANG, J.; YE, H.; ZHANG, D.; ***A new LDA-KL combined method for feature extraction and its generalization***. Pattern Analysis and Applications, v. 7, p. 40-50, 2004.

YOUNG, T. Y.; CALVERT, T. W.; ***Classification, Estimation, and Pattern Recognition***. Elsevier, 1974.

ZHANG, X.; CHEN, C.; ***Independent component analysis by using joint cumulants and its application to remote sensing images***. Journal of VLSI Signal Processing, v. 37, n. 2-3, p. 293–303, 2004.