DISSERTAÇÃO DE MESTRADO

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM

CIÊNCIA DA COMPUTAÇÃO

“Extração de características e aprendizado não-supervisionados em imagens hiperespectrais ”

ALUNO: Eduardo Kazuo Nakao
ORIENTADOR: Alexandre L. M. Levada

São Carlos
Abril/2020

CAIXA POSTAL 676
FONE/FAX: (16) 3351-8233
13565-905 - SÃO CARLOS - SP
BRASIL

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EXTRAÇÃO DE CARACTERÍSTICAS E APRENDIZADO NÃO-SUPERVISIONADOS EM IMAGENS HIPERESPECTRAIS

EDUARDO KAZUO NAKAO

ORIENTADOR: PROF. DR. ALEXANDRE L. M. LEVADA

São Carlos – SP

Abril/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EXTRAÇÃO DE CARACTERÍSTICAS E APRENDIZADO NÃO-SUPERVISIONADOS EM IMAGENS HIPERESPECTRAIS

EDUARDO KAZUO NAKAO

Dissertação 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, área de concentração: Processamento de Imagens e Sinais
Orientador: Prof. Dr. Alexandre L. M. Levada

São Carlos – SP

Abril/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Eduardo Kazuo Nakao, realizada em 29/04/2020:

Alexandre Luis M. Levada

Prof. Dr. Alexandre Luis Magalhães Levada
UFSCar

Prof. Dr. Marcelo Andrade da Costa Vieira
EESC/USP

Prof. Dr. Cesar Henrique Comin
UFSCar

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Alexandre Luis Magalhães Levada Marcelo Andrade da Costa Vieira, Cesar Henrique Comin e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Alexandre Luis M Levada

Prof. Dr. Alexandre Luis Magalhães Levada

AGRADECIMENTOS

Alexandre Luis Magalhães Levada

Maria do Carmo Nicoletti

Cesar Henrique Comin

Marcelo Andrade da Costa Vieira

Denis Henrique Pinheiro Salvadeo

Grupo de Arquitetura e Processamento de Imagens e Sinais

Grupo de Inteligencia Computacional

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

RESUMO

Imagens hiperespectrais possuem centenas de bandas e maior capacidade de discriminação de diferenças sutis em comparação a imagens multiespectrais, o que beneficia aplicações de precisão. Entretanto, a alta resolução espectral e alta correlação de bandas inerentes dessas imagens, sugerem a possibilidade de ocorrência da maldição da dimensionalidade em processos de reconhecimento de padrões. Dessa forma, o estudo dos efeitos de métodos de redução de dimensionalidade é relevante para esse tipo de imagem. Adicionalmente, é relevante a comparação de comportamento de métodos de redução linear e não-linear. Nesse cenário, o objetivo do presente trabalho é analisar como a extração não-supervisionada de características e suas diferentes abordagens afetam uma tarefa de aprendizado não-supervisionado em imagens hiperespectrais. Para conduzir tal análise, os algoritmos *Principal Component Analysis*, *Isometric Feature Mapping* e *Locally Linear Embedding* foram executados em um conjunto de sete imagens. Sob cada execução, foram realizados agrupamentos pelos algoritmos *K-Means* e *Expectation Maximization*. Os desempenhos foram mensurados pelas medidas *Rand*, *Jaccard*, *Kappa*, *Entropy*, *Purity* e comparados pelos testes estatísticos de *Friedman* e *Nemenyi*. Os resultados dos testes de hipótese mostraram que, para 70% das imagens, a aplicação da extração de características aumentou significativamente o desempenho da tarefa de agrupamento e, em 60% desses casos, a extração não-linear gerou melhores resultados em comparação à linear.

Palavras-chave: sensoriamento remoto, imagens hiperespectrais, reconhecimento de padrões, extração de características, métodos não-supervisionados.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AVIRIS	Airborne Visible Infra-Red Imaging Spectrometer
DR	Dimensionality Reduction
EM	Expectation Maximization
EO-1	Hyperion sensor on Earth Observing-1
GMM	Gaussian Mixture Model
GT	Ground Truth
HI	Hyperspectral Image
ISOMAP	Isometric Feature Mapping
LLE	Locally Linear Embedding
MDS	Multidimensional Scaling
MSE	Mean Squared Error
NASA	National Aeronautics and Space Administration
NDR	No Dimensionality Reduction
PCA	Principal Component Analysis
RGB	Red Green Blue
ROSIS	Reflective Optics System Imaging Spectrometer
SVM	Support Vector Machine

LISTA DE SÍMBOLOS

D – dimensionalidade do espaço original

K – número de vizinhos de cada ponto

R^D *– espaço original de dimensão D*

R^d *– espaço transformado de dimensão d*

T – matriz de transformação linear

X – conjunto de pontos nas coordenadas da dimensão original

Y – conjunto de pontos nas coordenadas da dimensão reduzida

Σ *– matriz de covariância*

γ *– multiplicador de Lagrange*

λ *– autovalor da matriz de covariância de X ou da matriz B*

μ *– média*

ε *– erro*

d – dimensionalidade reduzida

f – função de transformação

k – quantidade de grupos

n – quantidade de pontos

x – ponto no espaço original

y – mapeamento do ponto x

SUMÁRIO

Capítulo 1

INTRODUÇÃO

A maior dimensionalidade das imagens hiperespectrais (*hyperspectral images* - HIs) em comparação a imagens multiespectrais (LOWE; POLCYN; SHAY, 1965) enriquece consideravelmente o conteúdo da informação - as primeiras possuem 100 ou mais bandas enquanto as segundas entre 10 e 20. Diferentemente de dados multiespectrais, a alta resolução hiperespectral aumenta a capacidade de discriminação de diferenças sutis em objetos, beneficiando aplicações de precisão (e.g., observação terrestre, agricultura de precisão, detecção de doenças). Assim, o entendimento comum é o de que, a dimensionalidade em sua totalidade deveria ser usada para definir fronteiras precisas no espaço de características.

Entretanto, a presença de centenas de bandas acarreta uma alta resolução espectro-espacial, isto é, transições suaves em ambos os domínios espectral e espacial, onde os valores em localizações e comprimentos de onda vizinhos são altamente correlacionados. Esse fato é observado pela presença de matrizes de covariância extremamente não diagonais e amplas funções de autocorrelação (CAMPS-VALLS et al., 2011). Portanto, em ambos os domínios, a variação intra-classe aumenta e a variação inter-classe diminui, o que leva a baixas acurácias de classificação. Se os vetores que representam os pixels estiverem imersos em um espaço de alta dimensão, o mal entendimento desse tipo de espaço leva a interpretações incorretas das imagens e dificulta a escolha adequada da técnica de processamento. Dessas observações, surge a suspeita da ocorrência da maldição da dimensionalidade em HIs.

Esses fatos justificam o desenvolvimento de estudos em redução de dimensionalidade (*dimensionality reduction* - DR) na área. Adicionalmente, a comum indisponibilidade de informação da classe para essas imagens justifica o uso de técnicas não-supervisionadas. Mostra-se que um espaço com altas dimensões é praticamente vazio e dados multivariados podem ser representados em dimensões menores, onde os efeitos da maldição da dimensionalidade são atenuados (WEGMAN, 1990; KESHAVA et al., 2000; BENEDIKTSSON; GHAMISI, 2015).

A DR em teoria não prejudica a separabilidade de classes (quando corretamente parametrizada) e existe uma tendência de formação de distribuições Normais nas projeções de baixa dimensão, o que favorece o uso de DR como etapa prévia a um procedimento de reconhecimento de padrões e mais especificamente ao aprendizado não-supervisionado por misturas de Gaussianas (*Gaussian Mixture Model* - GMM).

Landgrebe (1998) define que o papel de um algoritmo de análise espectral em imagens é o de particionar os dados em sub-regiões mutuamente excludentes, cada uma pertencendo à uma classe de interesse de cobertura da superfície. De maneira genérica (i.e. independente da implementação), a tarefa central dessa análise é localizar de forma ótima as fronteiras dessas sub-regiões.

Nessa mesma referência, é dito que, no caso de um número reduzido de bandas, o uso do modelo gaussiano multivariado para determinar as distribuições das classes é uma forma especialmente útil de proceder. O modelo é eficiente para se estimar as distribuições mesmo com a usual escassez de amostras em classes de substâncias sutis. O modelo também representa cada classe por uma gaussiana específica (i.e. diferentes gaussianas no conjunto de dados), o que também é eficiente mesmo que cada classe não seja uma gaussiana de fato.

Em um espaço Euclidiano de baixa dimensão, é plausível a execução do GMM diretamente nos vetores em sua dimensão original. Porém, HIs podem definir um espaço de alta dimensão em que a assunção Euclidiana não é válida. As propriedades de espaços de alta dimensão e suas consequências fazem com que a assunção de independência linear dos vetores da base de um espaço Euclidiano possa se tornar incorreta. Interpretando os vetores como variáveis aleatórias, a alta correlação entre essas variáveis sugere que uma possível premissa de independência estatística também esteja incorreta.

Assim, um método de agrupamento como o GMM, que utiliza assunções de distância Euclidiana e independência estatística dos dados, pode vir a ter seu comportamento comprometido. Surge então o interesse em transformar o espaço original em um segundo de comportamento garantidamente Euclidiano, para então, aplicar a técnica GMM. Portanto, um processo de DR prévio ao GMM pode vir a melhorar o desempenho do agrupamento.

Poderia-se tentar reduzir a dimensionalidade com o uso da técnica *Principal Component Analysis* (PCA) porém, existe a possibilidade de ocorrência de um erro de projeção por essa técnica em decorrência da relação entre a dispersão das classes e a geometria que estruturas assumem em espaços de alta dimensão. Devido a esse fato, técnicas lineares de extração de caracterśiticas como PCA podem produzir resultados limitados, fazendo assim com que técnicas não-lineares sejam exploradas.

Tendo em vista esse cenário, o objetivo deste trabalho de mestrado é analisar como a DR por extração de características não-supervisionada e suas diferentes abordagens afetam o aprendizado não-supervisionado em HIs. Ou seja, pela análise do desempenho do agrupamento por GMM, será verificado o benefício da extração de características e sob qual tipo de extração obtêm-se um melhor resultado.

Capítulo 2

TRABALHOS RELACIONADOS

Feature selection for classification of hyperspectral data by SVM **(PAL; FOODY, 2010)** A principal conclusão desse estudo é a de que, em HIs a acurácia de uma tarefa de classificação que utiliza exclusivamente a técnica *Support Vector Machine* (SVM), é influenciada pelo número de características usadas. Assim, mostra-se que o método SVM é de fato afetado pelo fenômeno de Hughes. Os resultados presentes nesse artigo mostram em um conjunto de dados do sensor AVIRIS, para todos os tamanhos de conjunto de treinamento avaliados, a adição de características levou a um declínio estatisticamente significativo na acurácia. Com o conjunto de dados DAIS, tal declínio também foi observado, tanto para conjuntos de treinamento pequenos (25 amostras por classe ou menos) quanto para uma amostragem de treinamento grande. Foi mostrado portanto que a seleção de características provê um conjunto de características reduzido com acurácia de classificação similar a de um conjunto de características maior. Adicionalmente, a análise de quatro diferentes métodos de seleção exibiu uma grande variação entre a quantidade de caraterísticas selecionadas por cada método, o que destaca a influência da escolha do método no processo.

Assim como em nosso estudo, essa referência mostrou a manifestação do fenômeno de Hughes em HIs e a influência do método de DR na estimação da dimensionalidade intrínseca. Entretanto, uma técnica de classificação foi usada ao invés de uma técnica de agrupamento, assim como técnicas de seleção de características foram usadas no lugar de extração.

Feature Mining for Hyperspectral Image Classification **(JIA; KUO; CRAWFORD, 2013)** Com o argumento de que o uso do menor número possível de características discriminantes como entrada para um classificador pode evitar o problema de Hughes, esse artigo fornece uma visão geral de métodos fundamentais e modernos de seleção e extração de características aplicáveis na classificação de dados hiperespectrais. A comum indisponibilidade da rotulação

de classes nesse tipo de dados é discutida. É enfatizado que em abordagens paramétricas supervisionadas, medidas de separabildiade de classes requerem grande quantidade de dados de treinamento, o que pode não ser acessível ou fisicamente viável na prática. Isso faz com que métodos não-paramétricos ou não-supervisionados sejam preferidos portanto. Informa-se que o foco das pesquisas atuais na área está em extração não-linear de características.

Esse artigo portanto também mostra que a DR contorna o problema de Hughes em HIs e complementa o nosso trabalho com a descrição de mais métodos tanto de seleção quanto de extração. É fornecido ainda mais um argumento sobre a relevância das categorias de métodos não-paramétricos, não-supervisionados e não-lineares, que foram as escolhidas e utilizadas em nosso trabalho. Por fim destaca-se a relevância atual de pesquisa em extração não-linear, também investigada por nós.

***Subspace Feature Analysis of Local Manifold Learning for Hyperspectral Remote Sensing Images Classification* (DING; TANG; LI, 2014)** Essa referência consegue mostrar mais especificamente que, em HIs, o uso das características obtidas por métodos de extração não-linear fornece maior acurácia de classificação do que o uso das características originais. Esse tipo de conclusão também reforça as evidências mostradas em nosso trabalho, não só do benefício da extração de características em HIs como também da possibilidade de benefício de métodos não-lineares em particular.

***Unsupervised Clustering and Active Learning of Hyperspectral Images with Nonlinear Diffusion* (MURPHY; MAGGIONI, 2018)** Nesse trabalho também é mostrado que no contexto de HIs, métodos lineares de DR geram melhores resultados em comparação com o caso de ausência de DR no processo de classificação. Adicionalmente, mostra-se também que a DR não-linear tem se mostrado superior à linear nesse contexto. Isso indica que HIs são mapeadas em estruturas cuja dimensionalidade intrínseca é menor do que a dimensionalidade original e também que tais estruturas são particularmente não-lineares. O estudo mostra também que, em HIs os métodos de agrupamento que utilizam distância Euclidiana tem desempenho inferior quando comparados a métodos que utilizam outras estratégias.

Novamente encontramos portanto mais uma referência que mostra o benefício da DR em HIs e particularmente da DR não-linear. Adicionalmente esse trabalho também traz a informação da falha de estratégias Euclidianas nesse tipo de imagem, reforçando a validade das suspeitas nessa mesma linha que se comprovaram no nosso caso.

***Evaluation of Nonlinear dimensionality reduction Techniques for Classification of Hyperspectral Images* (MYASNIKOV, 2018)** Aqui também se discute que, em muitos casos, a combinação da técnica PCA com o classificador SVM por exemplo fornece bons resultados para a classificação de HIs. Entretanto, cenas hiperepectrais podem trazer uma complexidade advinda de efeitos não-lineares, que fazem com que a escolha do PCA possa ser inapropriada. Afirma-se que para tais cenas portanto, é necessário escolher com cuidado a dimensionalidade de saída e considerar o uso de técnicas não-lineares de DR. É informado também que a principal desvantagem dos métodos não-lineares se caracteriza pela alta complexidade computacional em comparação ao PCA.

Além do mesmo argumento não-linear de outras referências exibidas, esse trabalho adicionalmente trata de problemas como dimensionalidade intrínseca e complexidade computacional, os quais também foram comentados por nós nas subseções 3.3.4 e 3.3.5.

***Unsupervised exploration of hyperspectral and multispectral images* (MARINI; AMIGO, 2020)** Alguns dos métodos de exploração não-supervisionada mais usados (principalmente técnicas de DR e agrupamento) são revisados e discutidos, com foco em aplicações em HIs e imagens multiespectrais. São providas instruções de utilização e destacados benefícios e desvantagens na tentativa de desmistificar equívocos comuns. Argumenta-se que, quando o foco final da análise é algum tipo de predição, o papel da análise de dados exploratória é subestimado com frequencia. Afirma-se que, na verdade, a análise exploratória provê uma riqueza de informações por si e fornece um primeiro insight dos dados que pode ser muito útil, mesmo que a exploração não seja o objetivo final. Conclui-se que métodos não-supervisionados fornecem informações extremamanete valiosas e podem ser aplicados em uma primeira abordagem para a análise de qualquer amostra multi ou hiperespectral, haja vista que tais métodos são independentes de hipóteses.

Capítulo 3

FUNDAMENTAÇÃO TEÓRICA

3.1 Sensoriamento remoto hiperespectral

A área denominada sensoriamento remoto envolve a obtenção de informação sem contato físico de um objeto ou cena. Essa obtenção é possível porque diferentes objetos refletem, absorvem e emitem radiação eletromagnética de forma única, baseados na sua textura e composição molecular (na Figura 3.1 é possível recordar a estrutura do espectro eletromagnético). Se a radiação que chega em um sensor é medida em um intervalo detalhado de comprimento de onda, a assinatura espectral consequente pode ser usada para identificar um dado objeto de interesse. Para esse fim, a intenção da tecnologia de HIs é a de capturar centenas de canais espectrais que podem caracterizar precisamente a composição química e discriminar materiais espectralmente similares.[1]

De acordo com Plaza et al. (2009), a espectroscopia de imagem (GOETZ et al., 1985), (também conhecida como sensoriamento remoto hiperespectral), se preocupa com a medição, análise e interpretação do espectro adquirido de uma dada cena ou objeto. Tal aquisição é feita por um veículo aéreo ou sensor de satélite a uma curta, média ou longa distância. O conceito se originou em 1980 quando A. F. H. Goetz e seus colegas do Laboratório de Jato-Propulsão da NASA desenvolveram novos instrumentos como o *Airborne Imaging Spectrometer*, que resultou no atual *Airborne Visible Infra-Red Imaging Spectrometer* (AVIRIS) (GREEN et al., 1998).

A espectroscopia de imagem em menos de trinta anos se transformou em um produto amplamente disponível à comunidade. Sensores hiperespectrais amostram principalmente a porção reflexiva do espectro eletromagnético: indo da região visível (0,4 - 0,7 μm) até a região Infra-Vermelho de Ondas Curtas - em torno de 2,4 μm. Essa amostragem é decomposta em centenas

[1] essa seção foi baseada em (GHAMISI et al., 2017) com adaptações e transcrições parciais de trechos

de canais espectrais estreitos e contíguos (e.g. 0,01 μm de largura cada), constituindo uma fina resolução espectral. O AVIRIS por exemplo possui 224 bandas espectrais, cobrindo uma região de comprimento de onda no intervalo de 0,4 a 2,5 μm em uma resolução espectral nominal de 0,01 μm. Existem entretanto outros tipos de sensores hiperespectrais que coletam dados no intervalo de ondas infravermelhas médias e longas.

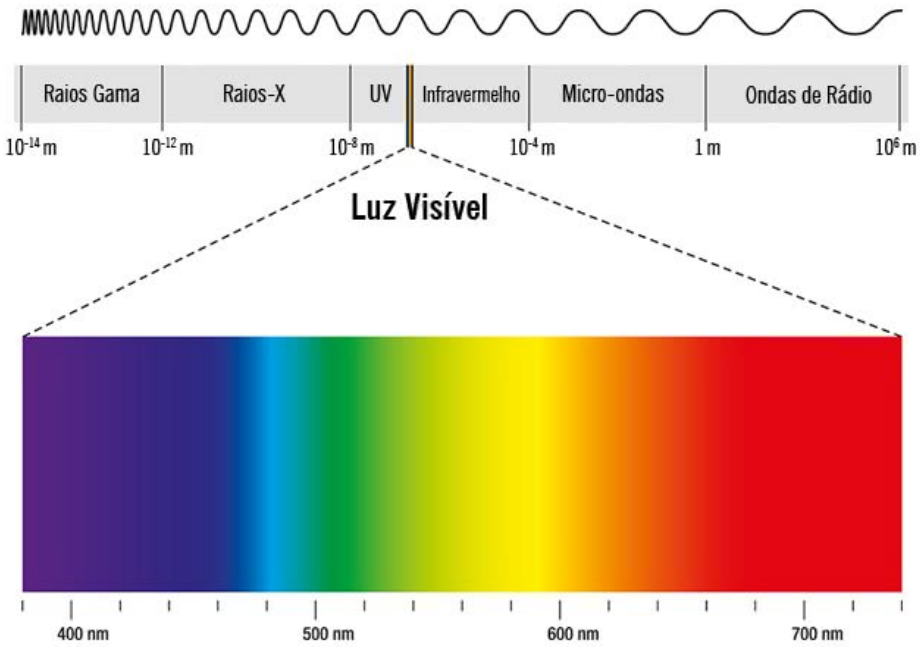


Figura 3.1: Espectro eletromagnético. Extraída de (ALVARENGA, 2019)

HIs geralmente são vistas como cubos hiperespectrais, onde as imagens de banda simples estão empilhadas de modo que a terceira dimensão do cubo é incrementada pelos comprimentos de onda amostrados. Uma HI que é coletada em D bandas espectrais pode ser pensada como uma nuvem de pontos (dados pelos pixels na imagem) em um espaço D-dimensional R^D. As D bandas espectrais na imagem formam D eixos de coordenadas no hiperespaço. Cada pixel é representado como um vetor D-dimensional tal que o valor da i-ésima coordenada é o valor estimado da refletância terrestre (ou da radiância do alcance do sensor) medida no i-ésimo comprimento de onda. Esse vetor é a representação geométrica do pixel no espaço espectral.

Um melhor entendimento das HIs pode ser obtido da Figura 3.2. Um cubo de dados hiperespectral tridimensional consiste em $n_1 \times n_2 \times D$ pixels, onde $n_1 \times n_2$ é o número de pixels em cada canal espectral e D representa o número de canais espectrais. Uma HI pode ser caracterizada por uma das seguintes definições:

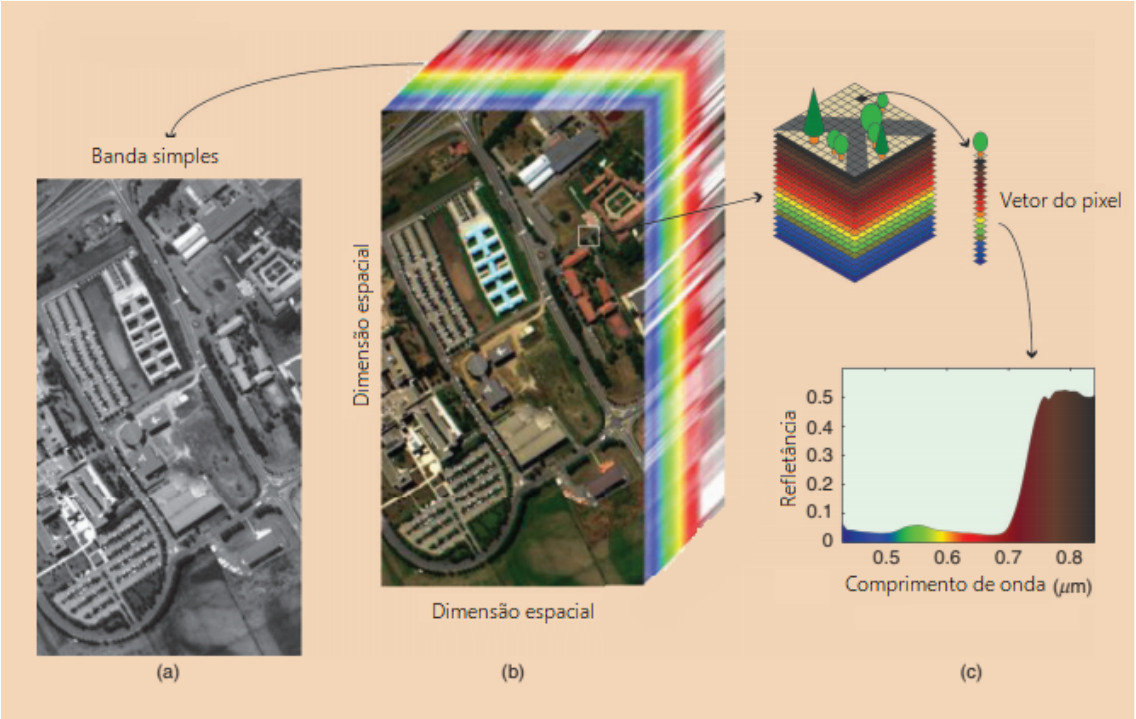


Figura 3.2: Exemplo de cubo de dados hiperespectral: (a) uma imagem em tons de cinza (b) um cubo hiperespectral (c) vetor de um pixel e sua correspondente assinatura espectral. Extraída de (GHAMISI et al., 2017)

1. Perspectiva espectral (ou dimensão espectral): sob essa perspectiva, o cubo é composto por $n_1 \times n_2$ pixels, onde cada pixel é um vetor de D valores. Cada pixel corresponde a radiação refletida em uma dada região da imagem e possui um valor de refletância para cada banda espectral. Figura 3.2(c) mostra o perfil espectral de um pixel.

2. Perspectiva espacial (ou dimensão espacial): nesse contexto, o cubo consiste em D imagens de tamanho $n_1 \times n_2$. Os valores de todos os pixels em uma banda espectral definem uma imagem de banda simples em duas dimensões, como mostrado na Figura 3.2(a), onde ambas dimensões são espaciais.

Ainda na Figura 3.2(b), note que a cena está representada por uma falsa composição RGB, que pode ser obtida da HI pela escolha subjetiva e empírica de três bandas que estariam fazendo o papel das três componentes R, G e B. A composição é então obtida pela soma dos valores de refletância dessas três bandas que resulta em uma quarta que será exibida de fato. Outra forma usual de se obter as três bandas é pelo uso do PCA com parâmetro igual a 3. Esse tipo de representação é útil para visualização humana para se ter uma ideia de como a cena seria enxergada a olho nu.

Outra representação dos dados é o Espaço de Características. O sensor amostra uma função contínua de energia vs comprimento de onda e a converte para um conjunto de medidas asso-

ciado a um pixel (que constitui um vetor no espaço D-dimensional). Essa conversão de uma função contínua para um vetor discreto é muito conveniente para uma análise algorítimica. As representações em Epaços de Características de duas e três dimensões podem ser usadas para visualizar um método de extração de características (Seção 3.3). Observe a Figura 3.3.

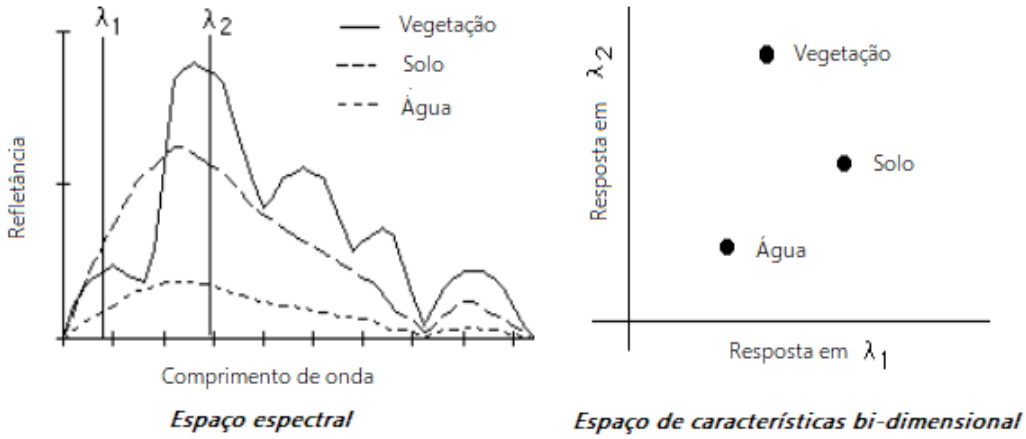


Figura 3.3: Representação de respostas espectrais em um espaço de características bidimensional. Extraída de (LANDGREBE, 1998)

Os avanços recentes na tecnologia de sensores e nas técnicas de processamento contribuem fortemente para o uso de dados hiperespectrais. Em baixas altitudes, aeronaves e veículos aéreos não-tripulados podem obter dados de altíssima resolução espectral e, de modo geral, os sensores possuem alta resolução espacial. Missões de observação e iniciativas de livre acesso aumentam a disponibilidade desses dados e permitem a caracterização, identificação e classificação das áreas de cobertura com maior precisão e robustez. A amostragem espectral detalhada é uma valiosa fonte de informação para uma grande variedade de aplicações.

Dentre as missões de observação existentes podemos citar: *Hyperion sensor on Earth Observing-1* (EO-1), AVIRIS, HyspIRI da *National Aeronautics and Space Administration* (NASA); PROBA, *Hyperspectral Imager Suite*, *Environmental Mapping and Analysis Program* (EnMAP) da *European Space Agency*; Gaofen da China; missões da *Japan Aerospace Exploration Agency*. Exemplos de iniciativas de acesso livre são a L.Biehl e o *Army Geospatial Center*.

Dentre as aplicações destacam-se: mapeamento urbano, gerenciamento e monitoramento ambiental, análise de plantio e agricultura de precisão (monitoramento do desenvolvimento e saúde das plantações), detecção mineral (mineralogia), detecção de estruturas camufladas em redondezas naturais, indústria de alimentos (caracterização da qualidade dos produtos), aplicações baseadas em segurança, imagens químicas, astronomia, ciências ecológicas, área médica, estudo da atmosfera, imagens de curto alcance.

3.2 Maldição da dimensionalidade

O termo maldição da dimensionalidade se refere a todas as dificuldades que ocorrem ao lidar com dados em alta dimensão. Sua primeira aparição ocorre em (BELLMAN; COLLECTION, 1961) onde se discute que, para estimar uma função com muitos atributos (dado um certo grau de acurácia), o número de amostras precisa aumentar junto com a quantidade de atributos.

Foi mostrado em (FUKUNAGA, 1990) que para a indução de classificadores lineares (e.g. *Nearest Mean Classifier*) o número de amostras necessárias é uma função linear do número de atributos. Para classificadores quadráticos é uma função quadrática (e.g. classificador Bayesiano sob hipótese Gaussiana com diferentes matrizes de covariância para cada classe). No caso de classificadores não-paramétricos, o número de amostras deve aumentar exponencialmente com o aumento dos atributos para se ter uma estimativa efetiva das densidades multivariadas (SCOTT, 1992; HWANG; LAY; LIPPMAN, 1994).

Na teoria de decisão Bayesiana, onde lidamos com um número ilimitado de amostras (já que as distribuições de probabilidade são completamente conhecidas a priori), o erro de classificação decresce monotonicamente (i.e. a probabilidade de erro se aproxima de zero) conforme o número de atributos aumenta (FUKUNAGA, 1990). Por outro lado, em um cenário real onde os parâmetros do modelo são estimados dos dados, o erro ε se mantém acima de um limite conforme o número de atributos aumenta (TRUNK, 1979).

O fenômeno de Hughes prova que para um número finito n de amostras, quando a quantidade de atributos cresce a partir de certo ponto, o desempenho da classificação é degradado (HUGHES, 1968). Um gráfico relacionado a esse fenômeno é o da Figura 3.4. Na ocorrência de muitos atributos é comum uma alta correlação entre as amostras, o que faz com que o erro quadrático ótimo na estimação de densidades seja grande mesmo que o número de amostras seja arbitrariamente grande (SCOTT, 1992).

Resultados relacionados a esse problema podem ser encontradas em (FUKUNAGA, 1990; SCOTT, 1992; JIMENEZ; LANDGREBE, 1998). Mais explicações e ilustrações se encontram em (JAIN; DUIN, 2000), (MURPHY, 2013) subseção 1.4.3, (DUDA; HART; STORK, 2001) seção 3.7, (BISHOP, 2006) seção 1.4, (LANDGREBE, 2019) e (LANDGREBE, 1998).

Esses fatos estão sob a óptica de uma tarefa de classificação no contexto de aprendizado de máquina. Porém há uma relação entre eles e a caracterização geométrica desse tipo de espaço. O termo maldição da dimensionalidade também costuma ser referenciado como fenômeno do espaço vazio, pois a quantidade de amostras disponíveis geralmente é restrita, fazendo com que espaços de alta dimensão sejam inerentemente esparsos.

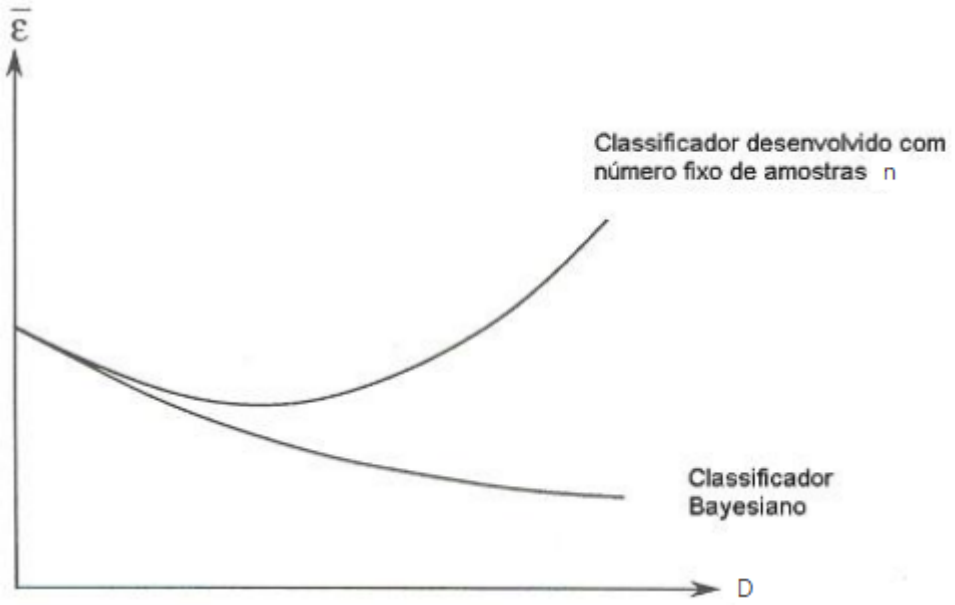


Figura 3.4: Ilustração do fenômeno de Hughes. Extraída de (FUKUNAGA, 1990)

Em contraste com o espaço Euclidiano usual, espaços de alta dimensão não são intuitivos e do ponto de vista geométrico e estatístico pode-se provar que[2] (JIMENEZ; LANDGREBE, 1998; LANDGREBE, 1998):

- vetores diagonais x_{diag} de um espaço de dimensão baixa se tornam ortogonais aos vetores da base em um espaço de dimensão alta, com isso, a projeção de um grupo de vetores quaisquer em um desses vetores x_{diag} pode arruinar a informação contida nos dados (WEGMAN, 1990; SCOTT, 1992).

- fenômeno da concentração (BEYER et al., 1999; FRANÇOIS, 2007): de acordo com Lee e Verleysen (2007), conforme a dimensionalidade cresce, o contraste provido por métricas usuais decresce (a distribuição das normas em um dado conjunto de pontos tende a se concentrar). Ou seja, métricas passam a ter baixo poder de discriminação. Por exemplo, a norma Euclidiana de vetores de muitas coordenadas que são independentes e identicamente distribuídos, se comporta de forma inesperada. A explicação e a prova podem ser encontradas em (DEMARTINES, 1994) mas, em suma, seja d a dimensão e x um vetor aleatório nessa dimensão, a norma de vetores aleatórios cresce proporcionalmente a $\sqrt{d}$, mas a variânca dos valores das normas se mantem mais ou menos constante para um d suficientemente grande. Isso significa também que o vetor x parece ser normalizado em alta dimensão. Mais precisamente, sendo μ a média, a probabilidade de que a norma de x esteja fora de um intervalo fixo centrado em $\mu_{\|x\|}$ se torna aproximadamente constante

[2] justificativas do item 1 e do caso da hiperesfera do item 2 (a) são exibidas no Apêndice A

quando d cresce. Conforme $\mu_{\|x\|}$ também cresce, o erro relativo de se tomar $\mu_{\|x\|}$ ao invés de $\|x\|$ se torna desprezível. As consequencias do fenômeno da concentração são:

- vetores originalmente aleatórios independentes e identicamente distribuídos de alta dimensão parecem estar localizados próximos à superfície de uma hiperesfera de raio $\mu_{\|x\|}$. O volume de uma hiperesfera ou hiperelipsóide se concentra em sua borda externa (WEGMAN, 1990; KENDALL, 2004) e o volume de um hipercubo se concentra em seus cantos (SCOTT, 1992; KENDALL, 2004). Isso significa que, não somente os vetores passam a possuir quase que o mesmo valor de norma, mas também que a distância Euclidiana entre dois vetores é aproximadamente constante. A distância Euclidiana é de fato a norma Euclidiana da diferença entre dois vetores aleatórios, e essa diferença também é um vetor aleatório. Outros resultados de normas e distâncias são dadas em (AGGARWAL; HINNEBURG; KEIM, 2001; FRANÇOIS, 2007).

- se os dados estão concentrados na casca, então estão distantes da média, fazendo com que dados Normalmente distribuídos estejam concentrados nas caudas da distribuição em vez de estarem ao redor da média. A diferença entre autovetores cresce, fazendo com que uma direção retenha muita variância e outra não. Seja Σ a matriz de covariância do conjunto de variáveis aleatórias (vetores) no espaço considerado e λs os autovalores de Σ. O número de condição de Σ é dado por $\lambda_{max}/\lambda_{min}$. Observa-se que a diferença entre λ_{min} e λ_{max} é alta, fazendo com que a matriz de covariância fique mal condicionada. Isso torna a estimação de densidade mais difícil. Veja a Figura 3.5: note a forma alongada indicando alta correlação dos dados e como a direção (autovetor) vertical retém pouca variância enquanto que a horizontal retém muita.

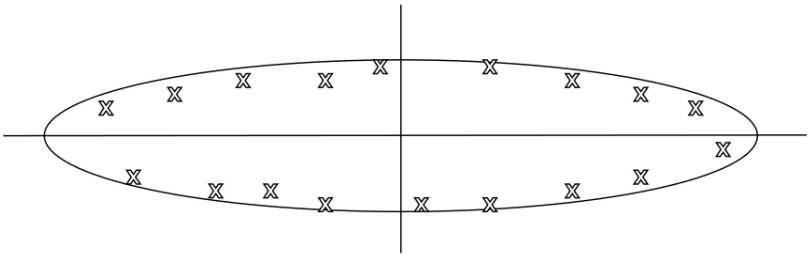


Figura 3.5: Ilustração de um hiperelipsóide com concentração de dados em borda

- vizinhanças locais são em sua maioria vazias, ou seja o problema da procura de vizinhos mais próximos se torna difícil de resolver (BEYER et al., 1999; BORODIN; OSTROVSKY; RABANI, 1999) o que causa perda de detalhes na estimação da densidade. Aggarwal, Hinneburg e Keim (2001) comentam que sob certas assunções razoáveis na

distribuição dos dados, a razão entre as distâncias do vizinho mais próximo e do mais distante é quase 1 para uma variedade de distribuições e funções de distância, conforme argumentado em (BEYER et al., 1999). Isso também contribui para que o problema de vizinhos mais próximos se torne mal condicionado dado que o contraste entre as distâncias de diferentes pontos não existe.

- fenômeno do espaço vazio que faz com que os dados possam ser projetados em um subespaço de menor dimensão sem perda significativa de informação de separabilidade entre classes (SCOTT; THOMPSON, 1983; CARREIRA-PERPINAN, 1997) .
- vetores projetados em dimensões menores tendem a formar uma distribuição Normal (HALL; LI, 1993; DIACONIS; FREEDMAN, 1984).

A Figura 3.6 ilustra como o aumento dimensional, ou seja, a representação de um memsmo conjunto de pontos em diferentes dimensões, acarreta no fenômeno do espaço vazio.

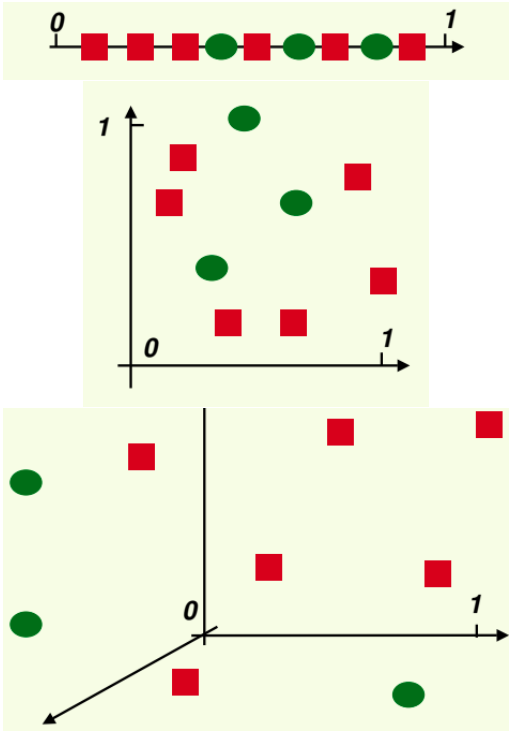


Figura 3.6: Extraída de http://www.cs.haifa.ac.il/~rita/ml_course/lectures/PCA_Fisher.pdf

Conhecida como rolo suíço, a Figura 3.7 apresenta uma possível configuração de um conjunto de dados de natureza originalmente bidimensional imerso em um espaço tridimensional (i.e. de alta dimensão nesse exemplo). Entende-se portanto que a dimensionalidade intrínseca desse estrutura seja igual a 2. A ideia de dimensionalidade intrínseca será formalizada e discutida com mais profundidade na seção a seguir. De forma complementar à Figura 3.5, a Figura

3.7 ilustra as propriedades citadas de espaço majoritariamente vazio com dados concentrados nas bordas, acompanhado do problema de detecção de vizinhos originalmente próximos e, de forma relacionada, da falha da métrica Euclidiana. Note no item a) e c) que a detecção de vizinhos pela métrica Euclidiana (azul tracejado) não reflete a vizinhança original (azul contínuo). Note também que os dados podem ser projetados para o espaço bidimensional com a manutenção da variância no conjunto. Esse exemplo de dispersão em particular será retomado em várias ilustrações na próxima seção para exemplificar, justificar e explicar os métodos de redução de dimensionalidade escolhidos.

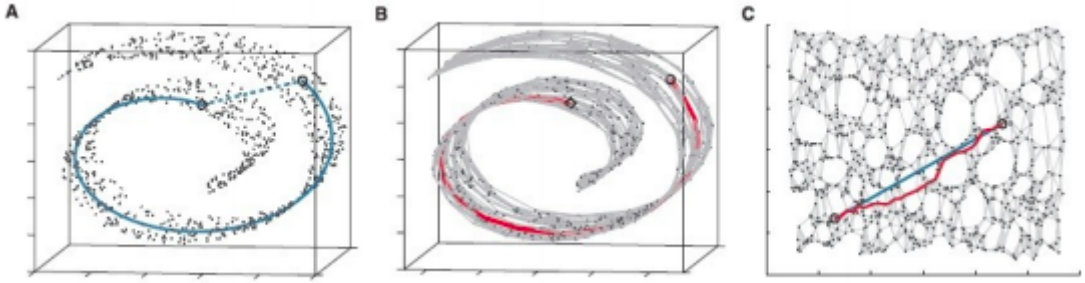


Figura 3.7: Aproximação da distância pela soma das arestas de um grafo no Rolo-Suíço. Extraída de (TENENBAUM; SILVA; LANGFORD, 2000)

3.3 Redução de dimensionalidade

A redução de dimensionalidade consiste na seleção e extração de características. A seleção de características constrói um subconjunto de d características $S = \{S_1, \ldots, S_d\}$ do conjunto original $F = \{F_1, \ldots, F_D\}$, onde $d \leq D$ e $S \subseteq F$, baseado em algum critério de desempenho em uma aplicação específica (e.g. classificação, detecção de objeto). Já o objetivo da extração de características é encontrar a função $f : R^D \rightarrow R^d$ que transforma o ponto $\left\{x_i \in R^D\right\}_{i=1}^{n}$ em $y_i = f(x_i)$, onde $\left\{y_i \in R^d\right\}_{i=1}^{n}$ e $d \leq D$, tal que a maior parte da informação dos dados é mantida. O termo f pode ser uma transformação linear ou não-linear. Uma ilustração de DR para o caso de HIs pode ser obtida na Figura 3.8. Ambas as categorias mantêm as relações de proximidade originais entre os pontos, são flexíveis à escolha do classificador posterior e têm se mostrado bem sucedidas na área de classificação (mesmo com a inevitável perda relativa de informação dos dados originais). Vale notar que a redução não pode ser excessiva pois o classificador pode perder a habilidade de discriminação (LI et al., 2011). Qualquer método de DR também é categorizado de acordo com o uso do rótulo da classe: não-supervisionado, supervisionado ou semi-supervisionado. Nesta subseção apresentaremos os três algoritmos de extração não-supervisionada escolhidos para o trabalho.

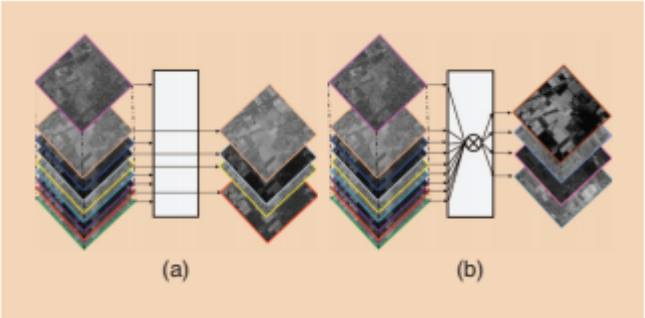


Figura 3.8: DR hiperespectral: (a) Seleção de Características (b) extração de características. Extraída de (GHAMISI et al., 2017)

3.3.1 *Principal Component Analysis*

De acordo com Cunningham e Ghahramani (2015), métodos lineares de DR foram desenvolvidos por muitas áreas da ciência (e.g. estatística, otimização, aprendizado de máquina) por mais de um século e se tornaram ferramentas matemáticas poderosas na análise de dados ruidosos e de alta dimensão. Parte de seu sucesso se deve a interpretações geométricas simples e propriedades computacionais atrativas. Em métodos lineares, procura-se uma matriz de projeção T que mapeie as amostras do espaço de caracterśticas original R^D, para um subespaço linear em R^d com $d < D$. Essa tarefa pode ser definida da seguinte forma:

Definição 1 (Redução de Dimensionalidade linear)**.** *Dados n pontos D-dimensionais*

$X = [x_1, ..., x_n] \in R^D$ *e uma escolha de dimensionalidade* $d < D$*, otimizar uma função objetivo* $f_X(.)$ *para produzir uma transformação linear* $T \in R^{d \times D}$ *responsável pelo mapeamento* $Y = TX \in R^{d \times n}$.

PCA é uma família de técnicas para tratar dados de alta dimensionalidade, que utilizam as dependências entre as variáveis para representá-los de uma forma mais compacta sem perda de informação relevante. É um dos métodos mais simples e robusto de DR e uma das técnicas mais antigas, tendo sido redescoberta muitas vezes em diversas áreas da ciência. PCA implementa a Transformação de Karhunen-Loève (ou Transformação de Hotteling) como é conhecida na literatura de reconhecimento de padrões (Jolliffe (2002) exibe na seção 1.2 um breve histórico com todas as referências relevantes). Além de ser o principal método linear existente, é também ainda o mais adotado na extração de características no geral. Se trata de um método clássico de análise estatística de dados multivariados que realiza a expansão de um vetor x em termos dos d autovetores da matriz de covariância associados aos d maiores autovalores. Por essa razão, é limitado à estatística de segunda ordem onde nenhuma hipótese sobre as densidades de probabilidade é necessária, uma vez que toda a informação pode ser estimada diretamente das

amostras. Dado um conjunto de dados multivariados, o objetivo é reduzir a dimensionalidade e a redundância existente para encontrar a melhor representação que minimize o critério de Erro Quadrático Médio (*Mean Squared Error* - MSE). A redundância é medida pela correlação entre os dados. Como os atributos mais significativos são utilizados, a representação no subespaço linear de fato minimiza esse critério.

O requisito básico em PCA é a existência de um vetor aleatório X com n elementos. Considera-se X como um vetor coluna. Devem estar disponíveis amostras $x_1, ..., x_n$ desse vetor. Nenhum modelo generativo é assumido para o vetor X, mas é necessário que os seus elementos sejam mutuamente correlacionados para que a compressão seja possível. Na transformação PCA, primeiramente os dados são centralizados subtraindo-lhes a média, que na prática é estimada através das amostras disponíveis. Em seguida, x é transformado linearmente em um outro vetor y contendo d elementos, com $d < D$, 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 x no novo sistema sejam não-correlacionados. Simultaneamente, as variâncias das projeções de x nos novos eixos são maximizadas, sendo que o primeiro eixo corresponde à maior variância, o segundo à maior na direção ortogonal ao primeiro e assim por diante. Pode ser mostrado em (FUKUNAGA, 1990) que, se λ_j e u_j são respectivamente o j-ésimo autovalor e autovetor da matriz de covariância de X , então para $j \neq k$

$$\lambda \geq 0$$
$$u_j \cdot u_k = 0$$

ou seja, todos os autovalores são positivos e os autovetores são mutuamente ortogonais entre si. Como consequência, para uma matriz de rank D, tem-se D autovetores ortonormais (assumindo que $\| u_j \| = 1$ para $j = 1, ..., D$) associados aos autovalores $\lambda_1, ..., \lambda_D$. Matematicamente, pode-se expressar a rotação do sistema de coordenadas definida pela Transformação Karhunen-Loève como uma matriz ortonormal $Z = \left[T^T, S^T\right]$, de dimensões $D \times D$, com

$T^T = [w_1, ..., w_d]_{D \times d}$ representando os eixos do novo sistema de coordenadas e

$S^T = [w_{d+1}, ..., w_D]_{D \times (D-d)}$ representando os eixos referentes às componentes eliminadas. A condição de ortonormalidade implica que $w_j \cdot w_k = 0$ para $j \neq k$, e $w_j \cdot w_k = 1$ para $j = k$. Pode-se escrever o vetor D-dimensional x através de sua expansão nos vetores da base:

$$x = \sum_{j=1}^{D} (x^T w_j) w_j = \sum_{j=1}^{D} c_j w_j \tag{3.1}$$

onde c_j é o produto interno entre x e w_j. Então, o novo vetor d-dimensional y é obtido por:

$$y^T = x^T T^T = \sum_{j=1}^{D} c_j w_j^T [w_1, ..., w_d] = [c_1, ..., c_d] \tag{3.2}$$

Desta forma, busca-se uma transformação T que maximize a variância dos dados, ou seja, otimize o critério PCA a seguir, com Σ_X sendo a matriz de covariância do vetor centralizado X:

$$J_1^{PCA}(T) = E\left[\| y \|^2\right] = E\left[y^T y\right] = \sum_{j=1}^{d} E\left[c_j^2\right] \tag{3.3}$$

porém, sabe-se que $c_j = x^T w_j$, e portanto:

$$J_1^{PCA}(w_j) = \sum_{j=1}^{d} E\left[w_j^T x x^T w_j\right] = \sum_{j=1}^{d} w_j^T E\left[xx^T\right] w_j = \sum_{j=1}^{d} w_j^T \Sigma_X w_j \tag{3.4}$$

sujeito a restrição $\| w_j \| = 1$. Trata-se de um problema de otimização com restrição de igualdade. É conhecido que a solução é encontrada através de multiplicadores de Lagrange:

$$J_1^{PCA}(w_j, \gamma_j) = \sum_{j=1}^{d} w_j^T \Sigma_X w_j - \sum_{j=1}^{d} \gamma_j (w_j^T w_j - 1) \tag{3.5}$$

Derivando a expressão acima em relação a cada componente de w_j e igualando a zero, chega-se ao seguinte resultado, encontrado em (YOUNG; CALVERT, 1974):

$$\Sigma_X w_j = \lambda_j w_j \tag{3.6}$$

Portanto, tem-se um problema de autovetores, ou seja, os vetores w_j da nova base que maximizam a variância dos dados transformados são os autovetores da matriz de covariância Σ_X. Porém, informações a respeito de como os d autovetores devem ser selecionados tornam-se mais claras na abordagem apresentada a seguir. É importante notar que, após essa transformação, os dados se encontram decorrelacionados. Ou seja, a matriz de covariância Σ_Y, onde Y é o resultado da aplicação da transformação T em X, é diagonal pela decomposição em autovalores da matriz Σ_X, onde $diag(\lambda_1, ..., \lambda_n)$ é a matriz diagonal dos valores próprios de Σ_X:

$$\Sigma_Y = T^T \Sigma_X T = T^T T diag(\lambda_1, ..., \lambda_n) T^T T = diag(\lambda_1, ..., \lambda_n) \tag{3.7}$$

Uma outra abordagem para o PCA é a da minimização do MSE. Nessa abordagem, busca-

se um conjunto de d vetores ortonormais de base que gerem um subespaço d-dimensional tal que, o MSE entre o vetor original x e sua projeção nesse subespaço, seja mínima. Denotando os vetores da base por $w_1, ..., w_m$, pela condição de ortonormalidade tem-se $w_i^T w_j = \delta_{ij}$ onde

$$\delta_{ij} = 1 \text{ se } i = j$$

$$\delta_{ij} = 0 \text{ se } i \neq j$$

A projeção de x no subespaço gerado pelos vetores w_j, com $j = 1, ..., d$, é dada pela equação 3.1 e portanto o critério MSE a ser minimizado torna-se:

$$J_{MSE}^{PCA}\left(w_j\right) = E\left[\parallel x - \sum_{j=1}^{d}\left(x^T w_j\right) w_j \parallel^2\right] \tag{3.8}$$

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

$$J_{MSE}^{PCA}\left(w_j\right) = E\left[\parallel x \parallel^2\right] - E\left[\sum_{j=1}^{d}\left(x^T w_j\right)^2\right] =$$

$$E\left[\parallel x \parallel^2\right] - \sum_{j=1}^{d} E\left[w_j^T x x^T w_j\right] = E\left[\parallel x \parallel^2\right] - \sum_{j=1}^{d} w_j^T \Sigma_X w_j \tag{3.9}$$

como o primeiro termo não depende de w_j, para minimizar o critério MSE basta maximizar

$$\sum_{j=1}^{d} w_j^T \Sigma_X w_j \tag{3.10}$$

Porém, da equação 3.4, esse mesmo problema de otimização foi resolvido através de multiplicadores de Lagrange, e o resultado obtido é que os vetores w_j devem ser os autovetores de Σ_X. Então, substituindo-se a equação 3.6 em 3.9, tem-se:

$$J_{MSE}^{PCA}\left(w_j\right) = E\left[\parallel x \parallel^2\right] - \sum_{j=1}^{d} \gamma_j \tag{3.11}$$

Esse resultado mostra que para minimizar o MSE, deve-se escolher os d autovetores associados aos d maiores autovalores da matriz de covariância. Foi mostrado em (FUKUNAGA, 1990) que o valor do mínimo MSE é:

$$J_{MSE}^{PCA}(w_j) = \sum_{j=d+1}^{D} \gamma_j \tag{3.12}$$

Para concluir, sumarizando então o algoritmo PCA:

Algorithm 1 Principal Component Analysis

1: **function** PCA(X)
2: Calcule a média e a matriz de covariância das amostras
$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$
$\Sigma_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)^T$
3: Calcule os autovalores e os autovetores de Σ_x
4: Defina a matriz de transformação $T = [w_1, w_2, ..., w_d]$ com os d autovetores associados aos d maiores autovalores
5: Projete os dados X no subespaço PCA
$y_i = Tx_i \quad \text{for} \quad i = 1, 2, ..., n$
6: **return** Y
7: **end function**

3.3.2 *Isometric Feature Mapping*

O *Isometric Feature Mapping* (ISOMAP) é um dos algoritmos pioneiros em DR não-linear. Seus autores propõem uma abordagem que combina os principais recursos algorítmicos do PCA e do *Multidimensional Scaling* (MDS) (COX; COX, 2001; BORG; GROENEN, 2005) - eficiência computacional, otimização global e garantias de convergência assintótica - com a flexibilidade de aprender uma ampla classe de estruturas não-lineares (TENENBAUM; SILVA; LANGFORD, 2000). A ideia é construir um grafo unindo os vizinhos mais próximos, computar os menores caminhos entre cada par de vértices e, conhecendo as distâncias entre os pontos, encontrar um mapeamento para um espaço de menor dimensão que preserve essas distâncias. Assume-se que os caminhos mínimos no grafo aproximam as distâncias reais (BERNSTEIN et al., 2000) conforme Figura 3.7. A Figura 3.9 fornece uma ideia visual do procedimento.

O algoritmo pode ser dividido em três passos:

Passo 1: induzir um grafo a partir do conjunto de dados $X = \{x_i\}$ para $i = 1, ..., n$ onde x_i denota o vetor de características que representa a i-ésima amostra. Existem basicamente duas formas de se criar um grafo G a partir dos dados. Em ambas, cada amostra representará um vértice do grafo e, se duas amostras estiverem ligadas em G, o peso da aresta será a distância entre os respectivos vetores.

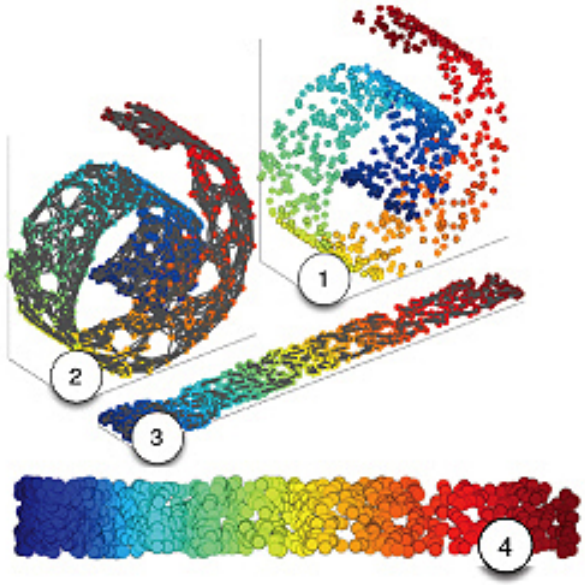


Figura 3.9: Desdobramento do Rolo-Suíço. Extraída de (WEINBERGER et al., 2006)

1. Grafo KNN:
   Parâmetro de entrada K: número de vizinhos de cada vértice
   Para cada amostra x_i do conjunto:
   Calcular a distância Euclidiana de x_i a cada outro x_j:
   $L_2(x_i,x_j) = \sqrt{(x_{i1}-x_{j1})^2+...+(x_{iD}-x_{jD})^2}$
   Selecionar as K amostras mais próximas de x_i
   Criar em G as arestas entre x_i e as K amostras selecionadas

2. Grafo R-vizinhança (LUXBURG, 2007):
   Parâmetro de entrada ε: raio que define a região de influência de cada amostra
   Para cada amostra x_i do conjunto:
   Calcular a distância Euclidiana de x_i a cada outro x_j
   Se $L_2(x_i,x_j) \leq \varepsilon$:
   Criar aresta entre x_i e x_j em G

Passo 2: Montar a matriz D de distâncias ponto a ponto: para cada amostra x_i do conjunto, aplicar o algoritmo de Dijkstra para obter os caminhos mínimos de x_i aos demais. Fazer $D_{ij} =$ tamanho do menor caminho entre x_i e x_j.

Passo 3: De posse da matriz D, encontrar um conjunto de pontos em um espaço de menor dimensão tal que as distâncias sejam preservadas. Esse problema é solucionado pelo algoritmo MDS cujo objetivo é, dada uma matriz de distâncias par a par, recuperar as coordenadas dos pontos $y_i \in R^d$, $i = 1,..,n$ onde d é definido pelo usuário.

A distância entre os vetores y_i e y_j é

$$L_{2ij}^2 = \|y_i - y_j\|^2 = (y_i - y_j)^T (y_i - y_j) \tag{3.13}$$

Para aliviar a notação vamos chamar L_{2ij}^2 de L_{ij}^2. A matriz de distâncias é dada por $D = \{L_{ij}^2\}$, $i, j = 1, ..., n$ (i é linha, j é coluna). Seja B a matriz dos produtos internos. $B = \{b_{ij}\}$, onde $b_{ij} = y_i^T y_j$. O método MDS baseia-se na resolução de dois sub-problemas: i) Encontrar a matriz B a partir de D; ii) Recuperar as coordenadas dos pontos a partir de B.

Sub-problema 1: Encontrar B a partir de D

Hipótese: a média dos dados é nula (pontos estão ao redor do vetor nulo).

$$\sum_{i=1}^{n} y_i = 0 \tag{3.14}$$

caso contrário há infinitas possibilidades, basta transladar os pontos. De L_{ij}^2 a partir da distributiva, temos:

$$L_{ij}^2 = y_i^T y_i + y_j^T y_j - 2 y_i^T y_j \tag{3.15}$$

assim, a partir da matriz D, podemos obter a média de uma coluna j como:

$$\frac{1}{n}\sum_{i=1}^{n} L_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^T y_i + \frac{1}{n}\sum_{i=1}^{n} y_j^T y_j - \frac{2}{n}\sum_{i=1}^{n} y_i^T y_j = \frac{1}{n}\sum_{i=1}^{n} y_i^T y_i + y_j^T y_j \tag{3.16}$$

analogamente, podemos computar a média de uma linha i como:

$$\frac{1}{n}\sum_{j=1}^{n} L_{ij}^2 = \frac{1}{n}\sum_{j=1}^{n} y_i^T y_i + \frac{1}{n}\sum_{j=1}^{n} y_j^T y_j - \frac{2}{n}\sum_{j=1}^{n} y_i^T y_j = y_i^T y_i + \frac{1}{n}\sum_{j=1}^{n} y_j^T y_j \tag{3.17}$$

e finalmente, podemos computar a média dos elementos de D como:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i^T y_i + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_j^T y_j - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i^T y_j =$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i^T y_i + \frac{1}{n}\sum_{j=1}^{n} y_j^T y_j = \frac{2}{n}\sum_{i=1}^{n} y_i^T y_i \tag{3.18}$$

note que de 3.15 é possível definir b_{ij} como:

$$b_{ij} = y_i^T y_j = -\frac{1}{2}(L_{ij}^2 - y_i^T y_i - y_j^T y_j) \tag{3.19}$$

mas de 3.17 podemos isolar o termo $-y_i^T y_i$:

$$-y_i^T y_i = -\frac{1}{n}\sum_{j=1}^{n} L_{ij}^2 + \frac{1}{n}\sum_{j=1}^{n} y_j^T y_j \tag{3.20}$$

e de 3.16 podemos isolar o termo $-y_j^T y_j$:

$$-y_j^T y_j = -\frac{1}{n}\sum_{i=1}^{n} L_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n} y_i^T y_i \tag{3.21}$$

então, fazendo 3.20 - 3.21 temos:

$$-y_i^T y_i - y_j^T y_j = -\frac{1}{n}\sum_{i=1}^{n} L_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} L_{ij}^2 + \frac{2}{n}\sum_{i=1}^{n} y_i^T y_i \tag{3.22}$$

de 3.18 podemos escrever:

$$\frac{2}{n}\sum_{i=1}^{n} y_i^T y_i = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}^2 \tag{3.23}$$

de modo que temos uma expressão completa para b_{ij} em função dos elementos de D:

$$b_{ij} = -\frac{1}{2}(L_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} L_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} L_{ij}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}^2) \tag{3.24}$$

chamando de $a_{ij} = -\frac{1}{2}L_{ij}$ podemos escrever as médias na linha i, na coluna j e em D respectivamente como:

$$a_{i.} = \frac{1}{n}\sum_{j=1}^{n} a_{ij} \tag{3.25}$$

$$a_{.j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij} \tag{3.26}$$

$$a_{..} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \tag{3.27}$$

expressando b_{ij}:

$$b_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..} \tag{3.28}$$

definindo a matriz $A = \{a_{ij}\}$, $i, j = 1, ..., n$ como $A = -\frac{1}{2}D$ pode-se mostrar que a relação entre B e A é ainda mais simplificada, sendo dada por $B = HAH$ onde a matriz H é definida por

$$H = I - \frac{1}{n}\vec{1}\vec{1}^T \tag{3.29}$$

sendo $\vec{1}^T = [1, ..., 1]$ (vetor de 1s com n dimensões). Dessa forma tem-se que

$$\vec{1}\vec{1}^T = U = \begin{bmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 1 \end{bmatrix} \tag{3.30}$$

note que $B = HAH$ nada mais é que a forma matricial da equação 3.28 uma vez que

$$B = HAH = \left(I - \frac{1}{n}U\right)A\left(I - \frac{1}{n}U\right) = \left(A - \frac{1}{n}UA\right)\left(I - \frac{1}{n}U\right) = A - A\frac{1}{n}U - \frac{1}{n}UA + \frac{1}{n^2}UAU \tag{3.31}$$

portanto, temos a matriz B.

Sub-problema 2: Recuperar as coordenadas $y_i \in R^d$ a partir de B.

Note que a matriz B dos produtos internos pode ser expressa por:

$$B_{n\times n} = Y_{n\times d}Y^T_{d\times n} \tag{3.32}$$

onde n denota o número de amostras e d denota o número de dimensões. A matriz B possui três propriedades importantes:

a) simétrica

b) o *rank* de B é d (número de linhas/colunas linearmente independentes)

c) positiva e semi-definida: $\forall y \in R^n$, $y^T By \geq 0$

Isso implica em dizer que a matriz B possui d autovalores não negativos e $n-d$ autovalores nulos. Assim, pela decomposição espectral de B pode-se escrever:

$$B = V\Lambda V^T \tag{3.33}$$

onde $\Lambda = diag(\lambda_1, ..., \lambda_n)$ é a matriz diagonal dos autovalores de B e

$$V = \begin{bmatrix} | & | & ... & | \\ | & | & ... & | \\ v_1 & v_2 & ... & v_n \\ | & | & ... & | \\ | & | & ... & | \end{bmatrix}_{n \times n} \tag{3.34}$$

é a matriz dos autovetores de B. Sem perda de generalidade iremos considerar $\lambda_1 \geq ... \geq \lambda_n$. Devido aos $n-d$ autovalores nulos, B pode ser escrita como:

$$B = V'\Lambda'V'^T \tag{3.35}$$

onde $\Lambda' = diag(\lambda_1, ..., \lambda_d)$ é a matriz diagonal dos autovalores de B e

$$V' = \begin{bmatrix} | & | & ... & | \\ | & | & ... & | \\ v_1 & v_2 & ... & v_d \\ | & | & ... & | \\ | & | & ... & | \end{bmatrix}_{n \times d} \tag{3.36}$$

mas como $B_{n\times n} = Y_{n\times d}Y^T_{d\times n} = V'\Lambda'V'^T = V'\Lambda'^{1/2}\Lambda'^{1/2}V'^T$ temos finalmente que

$$Y_{n\times d} = V'_{n\times d}\Lambda'^{1/2} \tag{3.37}$$

onde $\Lambda'^{1/2} = diag(\sqrt{\lambda_1}, ..., \sqrt{\lambda_d})$. Cada linha de $Y_{n\times d}$ terá a coordenada de um vetor $y_i \in R^d$, onde d é um parâmetro que controla o número de dimensões do espaço de saída. A seguir, os algoritmos MDS e ISOMAP são sumarizados:

Multidimensional Scaling:

1. Entrada: $D = \{L^2_{ij}\}$ (obtida criando grafo e executando Dijkstra)
2. Faça $A = -\frac{1}{2}D$
3. Faça $H = I - \frac{1}{n}\vec{1}\vec{1}^T$
4. Calcule $B = HAH$
5. Encontre os d autovetores associados aos d maiores autovalores de B e construa $V'_{n\times d}$ e $\Lambda' = diag(\lambda_1, ..., \lambda_d)$
6. Calcule $Y_{n\times d} = V'_{n\times d}\Lambda'^{1/2}_{d\times d}$

Algorithm 2 Isometric Feature Mapping

function ISOMAP(X)
  Dos dados de entrada $X_{D\times n}$ construa um grafo KNN.
  Calcule a matriz de distâncias ponto a ponto $D_{n\times n}$.
  Calcule $A = -\frac{1}{2}D$.
  Calcule $H = I - \frac{1}{n}U$, onde U é uma matriz $n \times n$ de 1's.
  Calcule $B = HAH$.
  Encontre os autovetores e autovalores da matriz B.
  Selecione os $d < D$ maiores autovetores e autovalores de B e defina:

$$V = \begin{bmatrix} | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \\ v_1 & v_2 & \dots & \dots & v_d \\ | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \end{bmatrix}_{n\times d} \tag{3.38}$$

$$\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_d) \tag{3.39}$$

  Calcule $Y = \Lambda^{1/2}V^T$
  return Y
end function

3.3.3 *Locally Linear Embedding*

O algoritmo ISOMAP é um método global no sentido de que, para encontrar as coordenadas de um dado vetor $x \in R^D$, ele usa a informação sobre todas as amostras através da matriz B. Em contrapartida, o *Locally Linear Embedding* (LLE), como o nome sugere, é um método local, ou seja, as novas coordenadas de qualquer $x \in R^D$ dependem apenas da vizinhança desse ponto. A hipótese principal por trás do LLE é a de que, para uma densidade suficientimente grande de amostras, é esperado que o vetor x e seus vizinhos formem um *patch* linear, ou seja, pertençam todos a um sub-espaço Euclidiano (ROWEIS; SAUL, 2000), conforme Figura 3.10.

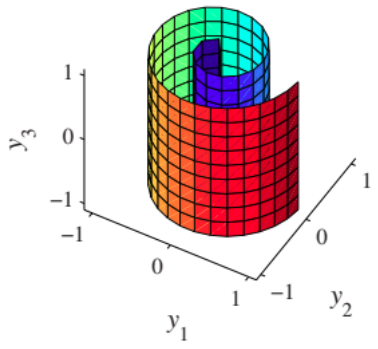


Figura 3.10: Rolo-Suíço interpretado como conexões de planos locais. Extraída de (LEE; VERLEYSEN, 2007)

Dessa maneira, é possível caracterizar a geometria local por coeficientes lineares:

$$x_i \approx \sum_j w_{ij} x_j \tag{3.40}$$

para $x_j \in U(x_i)$ isto é, pode-se reconstruir um vetor como combinação linear de seus vizinhos. Basicamente, o algoritmo LLE requer como entrada uma matriz $X_{n\times D}$ com linhas x_i, o número de dimensões desejado $d < D$ e um inteiro $K > d+1$ para encontrar as vizinhanças locais. A saída é uma matriz $Y_{n\times d}$ com linhas y_i. O algoritmo LLE é dividido em três etapas (ROWEIS; SAUL, 2000; SAUL; ROWEIS, 2003), conforme Figura 3.11:

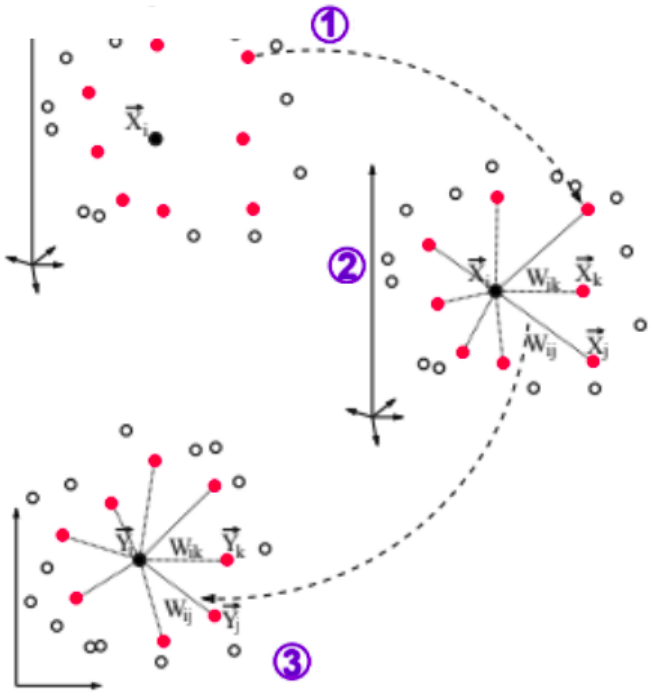


Figura 3.11: Etapas do LLE. Extraída de (SAUL; ROWEIS, 2003)

1. Para cada $x_i \in R^D$ encontre os K vizinhos mais próximos;

2. Encontre a matriz de pesos W que minimize o erro de reconstrução para cada $x_i \in R^D$

$$E(W) = \sum_{i=1}^{n} \| x_i - \sum_j w_{ij} x_j \|^2 \tag{3.41}$$

onde $w_{ij} = 0$ a menos que x_j seja um dos K vizinhos mais próximos de x_i e para cada i, $\sum_j w_{ij} = 1$

3. Encontre as coordenadas Y que minimizem o erro de reconstrução usando pesos ótimos

$$\Phi(Y) = \sum_{i=1}^{n} \| y_i - \sum_j w_{ij} y_j \|^2 \tag{3.42}$$

sujeito às restrições $\sum_i Y_{ij} = 0$ para cada j, e $Y^T Y = I$

A seguir será descrito como obter a solução de cada passo do LLE.

Encontrando as vizinhanças lineares locais No algoritmo usual do LLE, um número fixo de vizinhos mais próximos é definido para cada exemplar através da distância Euclidiana. Outros critérios também podem ser usados para a escolha de vizinhos, como escolher todos os pontos pertencentes à bola de raio fixo. Além disso, o número de vizinhos não precisa ser o mesmo para todos os pontos. Regras alternativas também são possíveis, por exemplo, selecionar todos os pontos dentro de um certo raio até uma quantidade máxima, ou selecionar um certo número de vizinhos onde nenhum esteja fora de um raio máximo (SAUL; ROWEIS, 2003).

Muitos critérios devem ser levados em consideração para escolher K. Primeiramente, o algoritmo vai recuperar somente as imersões cuja dimensionalidade d seja estritamente menor do que K. Em segundo lugar, LLE é baseado na assunção de que um ponto e seus vizinhos sejam localmente lineares. Para conjuntos de dados não-lineares, a escolha de um K muito grande vai em geral violar essa assunção. Finalmente, no caso não-usual onde $K > D$, cada ponto pode ser perfeitamente reconstruído de seus vizinhos, e os pesos de reconstrução local deixam de ser unicamente definidos. Nesse caso, regularização adicional tem de ser adicionada para quebrar essa degeneração (SAUL; ROWEIS, 2003).

O algoritmo LLE também precisa checar se o grafo KNN é conexo. Se o grafo é disconexo (ou com baixa conectividade) então o LLE deve ser aplicado separadamente para cada um dos componentes conexos do grafo; caso contrário, a regra de seleção de vizinhança tem de ser refinada para gerar um grafo com maior conectividade (SAUL; ROWEIS, 2003).

Estimação dos pesos por mínimos quadrados A segunda etapa do LLE é reconstruir cada ponto de seus vizinhos mais próximos. Os pesos de reconstrução ótima podem ser calculados de forma fechada. Sem perda de generalidade, pode-se expressar o erro de reconstrução local no ponto x_i como:

$$E(w) = \| \sum_j w_j (x_i - x_j) \|^2 = \sum_j \sum_K w_j w_K (x_i - x_j)(x_i - x_K)^T \tag{3.43}$$

definindo a matriz C como:

$$C_{jk} = (x_i - x_j)(x_i - x_K)^T \tag{3.44}$$

temos a seguinte expressão para o erro de reconstrução local:

$$E(w) = \sum_j \sum_K w_j C_{jK} w_K = w^T C w \tag{3.45}$$

A restrição $\sum_j w_{ij} = 1$ pode ser entendida de forma geométrica ou probabilística. Do ponto de vista geométrico, ela provê invariância à translação, ou seja, a adição de qualquer vetor constante c a x_i e a todos os seus vizinhos não altera o erro de reconstrução. Seja $x_i = x_i + c$ e $x_j = x_j + c$. Então o novo erro de reconstrução local é dado por:

$$\begin{aligned}
\widetilde{E}(w) =& \| x_i - \sum_j w_j x_j \|^2 \\
=& \| x_i + c - \sum_j w_j (x_j + c) \|^2 \\
=& \| x_i + c - \sum_j w_j x_j - \sum_j w_j c \|^2 \\
=& \| x_i + c - \sum_j w_j x_j - c \|^2 \\
=& E(w)
\end{aligned} \tag{3.46}$$

Em termos de probabilidade, impor que os pesos somem 1 faz de W uma matriz de transição estocástica (SAUL; ROWEIS, 2003). Vamos mostrar que na minimização do erro quadrático, a solução é encontrada através de um problema de autovalores. Na verdade, a estimação da matriz W se reduz a n problemas de autovalores: como não há restrições entre as linhas de W, podemos encontrar os pesos ótimos para cada amostra x_i separadamente, o que simplifica os cálculos. Assim, temos n problemas independentes de otimização com restrição dados por:

$$\arg\min_{w_i} w_i^T C_i w_i \tag{3.47}$$

sujeito a $\vec{1}^T w_i = 1$ para $i = 1, ..., n$

Usando multiplicadores de Lagrange, escreve-se a função Lagrangiana como:

$$L(w_i, \gamma) = w_i^T C_i w_i - \gamma(\vec{1}^T w_i - 1) \tag{3.48}$$

Derivando em relação a w_i:

$$\frac{\partial}{\partial w_i} L(w_i, \gamma) = 2 C_i w_i - \gamma \vec{1} = 0 \tag{3.49}$$

o que leva em

$$C_i w_i = \frac{\gamma}{2} \vec{1} \tag{3.50}$$

Se a matriz C_i é invertível, temos a solução em forma fechada:

$$w_i = \frac{\gamma}{2} C_i^{-1} \vec{1} \tag{3.51}$$

onde λ pode ser ajustado para assegurar que $\sum_j w_i(j) = 1$. Na verdade, há uma expressão fechada para $w_i(j)$ (SAUL; ROWEIS, 2000):

$$w_i(j) = \frac{\sum_K C_i^{-1}(j,K)}{\sum_K \sum_l C_i^{-1}(K,l)} \tag{3.52}$$

Para acelerar o algoritmo, em vez de se calcular a inversa da matriz C, resolve-se o seguinte sistema linear:

$$C_i w_i = \vec{1} \tag{3.53}$$

e então normaliza-se a solução para garantir que $\sum_j w_i(j) = 1$ dividindo-se cada coeficiente do vetor w_i pela soma de todos os coeficientes:

$$w_i(j) = \frac{w_i(j)}{\sum_j w_i(j)} \tag{3.54}$$

para $j = 1, ..., D$

Se $K > D$, então no geral os K vetores distintos geram todo o espaço. Isso significa que x_i pode ser escrito exatamente como combinação linear dos seus k vizinhos mais próximos. De fato, se $K > D$, há infinitas soluções para $x_i = \sum_j w_j x_j$ pois há mais variáveis desonhecidas K do que equações D. Nesse caso, o problema de otimização é mal-condicionado e uma regularização é necessária. Uma técnica de regularização comum é a regularização de Tikonov que, em vez de minimizar diretamente

$$\| x_i - \sum_j w_j x_j \|^2 \tag{3.55}$$

adiciona um termo de penalização para o problema de mínimos quadrados

$$\| x_i - \sum_j w_j x_j \|^2 + \alpha \sum_j w_j^2 \tag{3.56}$$

onde α controla o grau de regularização. Em outras palavras, seleciona-se os pesos que minimizem a combinação do erro de reconstrução e a soma dos quadrados dos pesos. Quando $\alpha \to 0$ tem-se o problema de mínimos quadrados. No limite oposto, $\alpha \to \infty$, o termo de erro quadrático se torna desprezível e busca-se minimizar a norma Euclidiana do vetor de peso w. Geralmente, α é setado para ser um valor pequeno porém diferente de zero. Nesse caso, os n

problemas independentes de otimização com restrição são:

$$\arg\min_{w_i} w_i^T C_i w_i + \alpha w_i^T w_i \tag{3.57}$$

sujeito a $\vec{1}^T w_i = 1$ para $i = 1, ..., n$

A função Lagrangiana é definida por:

$$L(w_i, \gamma) = w_i^T C_i w_i + \alpha w_i^T w_i - \gamma(\vec{1}^T w_i - 1) \tag{3.58}$$

Tomando a derivada em relação a w_i e igualando a zero:

$$2C_i w_i + 2\alpha w_i = \gamma\vec{1} \tag{3.59}$$

$$(C_i + \alpha I)w_i = \frac{\gamma}{2}\vec{1} \tag{3.60}$$

$$w_i = \frac{\gamma}{2}(C_i + \alpha I)^{-1}\vec{1} \tag{3.61}$$

onde λ é escolhido para normalizar w_i.

Encontrando as coordenadas A ideia principal por trás do terceiro passo do LLE é usar os pesos de reconstrução ótimos estimados por mínimos-quadrados para encontrar as novas coordenadas. Desse modo, fixando a matriz de pesos W, o objetivo é resolver outro problema de minimização quadrática:

$$\Phi(Y) = \sum_{i=1}^{n} \| y_i - \sum_j w_{ij} y_j \|^2 \tag{3.62}$$

Em outras palavras, tem de se responder à questão: quais são as coordenadas $y_i \in R^d$que são reconstruídas pelos pesos de W?

Para não degenerar o problema, duas restrições são impostas:

1. a média dos dados no espaço transformado é zero, caso contrário teríamos um infinito número de soluções

2. a matriz de covariância dos dados transformados é igual à Identidade, isto é, não há correlação entre os componentes de $y \in R^d$

Contudo, diferentemente da estimação dos pesos W, encontrar as coordenadas não se simplifica em n problemas independentes, porque cada linha de Y aparece em Φ diversas vezes, uma vez como o vetor central y_i e novamente como um dos vizinhos de outros vetores.

Primeiro, vamos re-escrever a equação 3.62 de maneira mais significativa usando matrizes. Note que:

$$\Phi(Y) = \sum_{i=1}^{n} \left[\left(y_i - \sum_j w_{ij} y_j \right)^T \left(y_i - \sum_j w_{ij} y_j \right) \right] \tag{3.63}$$

Aplicando a distributiva:

$$\Phi(Y) = \sum_{i=1}^{n} \left[y_i^T y_i - y_i^T \left(\sum_j w_{ij} y_j \right) - \left(\sum_j w_{ij} y_j \right)^T y_i + \left(\sum_j w_{ij} y_j \right)^T \left(\sum_j w_{ij} y_j \right) \right] \tag{3.64}$$

Expandindo o somatório:

$$\Phi(Y) = \sum_{i=1}^{n} y_i^T y_i - \sum_{i=1}^{n} \sum_j y_i^T w_{ij} y_j - \sum_{i=1}^{n} \sum_j y_j^T w_{ji} y_i + \sum_{i=1}^{n} \sum_j \sum_K y_j^T w_{ji} w_{ik} y_K \tag{3.65}$$

Denotando por Y a matriz $d \times n$ onde cada coluna y_i para $i = 1, ..., n$ armazena as coordenadas do i-ésimo ponto e sabendo que $w_i(j) = 0$ a menos que y_j seja um dos vizinhos de y_i, podemos escrever $\Phi(Y)$ como:

$$\begin{aligned}
\Phi(Y) &= Tr(Y^T Y) - Tr(Y^T W Y) - Tr(Y^T W^T Y) + Tr(Y^T W^T W Y) \\
&= Tr(Y^T Y) - Tr(Y^T (WY)) - Tr((WY^T)Y) + Tr((WY)^T (WY)) \\
&= Tr(Y^T (Y - WY) - (WY)^T (Y - WY)) \\
&= Tr((Y - WY)^T (Y - WY)) \\
&= Tr(((I - W)Y)^T ((I - W)Y)) \\
&= Tr(Y^T (I - W)^T (I - W) Y)
\end{aligned} \tag{3.66}$$

Definindo a matriz $M_{n \times n}$ como:

$$M = (I - W)^T (I - W) \tag{3.67}$$

tem-se o seguinte problema de otimização:

$$\arg\min_Y Tr(Y^T M Y) \tag{3.68}$$

sujeito a $\frac{1}{n} Y^T Y = I$

Assim, a função Lagrangiana é dada por:

$$L(Y, \gamma) = Tr(Y^T M Y) - \gamma \left(\frac{1}{n} Y^T Y - I \right) \tag{3.69}$$

Derivando e igualando a zero:

$$2MY - 2\frac{\gamma}{n}Y = 0 \tag{3.70}$$

$$MY = \beta Y \tag{3.71}$$

onde $\beta = \frac{\gamma}{n}$, evidenciando que Y é composto pelos autovetores da matriz M. Já que temos um problema de minimização, queremos que Y seja composto pelos d autovetores associados aos d menores autovalores. Note que sendo M uma matriz $n \times n$, ela possui n autovalores e n autovetores ortogonais. Apesar dos autovalores serem reais e não-negativos, o menor deles é sempre zero com autovetor constante $\vec{1}$. Esse autovetor corresponde à média de Y e deve ser descartado para garantir a restrição $\sum_i^n y_i = 0$ (RIDDER; DUIN, 2002). Note que cada linha de W deve somar um, e portanto:

$$W\vec{1} = \vec{1} \tag{3.72}$$

$$\vec{1} - W\vec{1} = 0 \tag{3.73}$$

$$(I - W)\vec{1} = 0 \tag{3.74}$$

$$(I - W)^T(I - W)\vec{1} = 0 \tag{3.75}$$

$$M\vec{1} = 0 \tag{3.76}$$

Portanto, para obter $y_i \in R^d$, onde $d < D$, temos de selecionar os $d+1$ menores autovetores e descartar o autovetor constante com autovalor zero. Em outras palavras, temos de selecionar os d autovetores associados aos menores autovalores diferentes de zero.

O algoritmo a seguir mostra um resumo do método LLE. A entrada é uma matriz $X_{D\times n}$ cujas colunas são as amostras e a saída é a matriz $Y_{n\times d}$ cujas linhas representam as novas coordenadas dos pontos.

3.3.4 Considerações adicionais

Em um espaço de alta dimensão relativa, as amostras estão correlacionadas, garantindo assim o pré-requisito de aplicação do PCA. Os três métodos apresentados nessa seção, ao final de suas execuções obtêm uma base ortonormal em um espaço onde os vetores de padrões podem ser expressos como combinação linear dos componentes da base. No PCA isso é feito de forma explícita, ou seja, temos acesso direto a cada vetor da base. Nos métodos não-lineares isso é feito de maneira implícita, ou seja, ao final de cada algoritmo não temos acesso aos vetores da base em si, apenas às coordenadas locais no espaço Euclidiano de saída (pois não existe uma

Algorithm 3 Locally Linear Embedding

function LLE(X, K, d)
  Dos dados de entrada $X_{D\times n}$ construa um grafo KNN.
  for $\vec{x}_i \in X^T$ **do**
    Calcule a matriz C_i $K \times K$ como:

$$C_i(j,K) = (\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_K)^T \tag{3.77}$$

    Resolva o sistema linear $C_i\vec{w}_i = \vec{1}$ para estimar os pesos $\vec{w}_i \in R^K$.
    Normalize os pesos em $\vec{w}_i$ tal que $\sum_j \vec{w}_i(j) = 1$.
  end for
  Construa a matriz W $n \times n$, cujas linhas são os $\vec{w}_i$ estimados.
  Calcule $M = (I-W)^T(I-W)$.
  Encontre os autovalores e os autovetores da matriz M.
  Selecione os d menores autovetores de M e defina:

$$Y = \begin{bmatrix} | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \\ v_1 & v_2 & \dots & \dots & v_d \\ | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \end{bmatrix}_{n\times d} \tag{3.78}$$

  return Y
end function

matriz de projeção construída explicitamente). Note que o PCA determina as coordenadas dos pontos y dadas na matriz Y da seguinte forma:

$$Y_{d\times n} = T_{d\times D} \cdot X_{D\times n}$$

ou seja, como uma transformação linear de X. Já os métodos não-lineares exibem as novas coordenadas em suas próprias matrizes, onde cada coluna é um autovetor:

$$\begin{bmatrix} -- & y_1 & -- \\ & . & \\ & . & \\ & . & \\ -- & y_n & -- \end{bmatrix}$$

Portanto, a matriz Y não é dada por uma transformação linear de X e assim, tais técnicas de redução são denominadas não-lineares. Além disso, como os vetores são vistos como variáveis aleatórias por alguns algoritmos subsequentes, é assegurado também que essas variáveis não estão correlacionadas, o que é uma consequência direta da propriedade anterior. Em resumo, em todos os três métodos, as coordenadas são obtidas a partir de autovetores de uma determinada

matriz, que por definição são ortogonais. Ao final do processo de qualquer um dos três métodos, temos a garantia que o espaço reduzido é linear.

Para justificar e ilustrar a aplicação de métodos não-lineares, um exemplo clássico é o "desdobramento" do Rolo Suíço. A Figura 3.12 mostra esse conjunto de dados e como ele não é bem representado pelo PCA em termos da separação das classes definidas nesse caso.

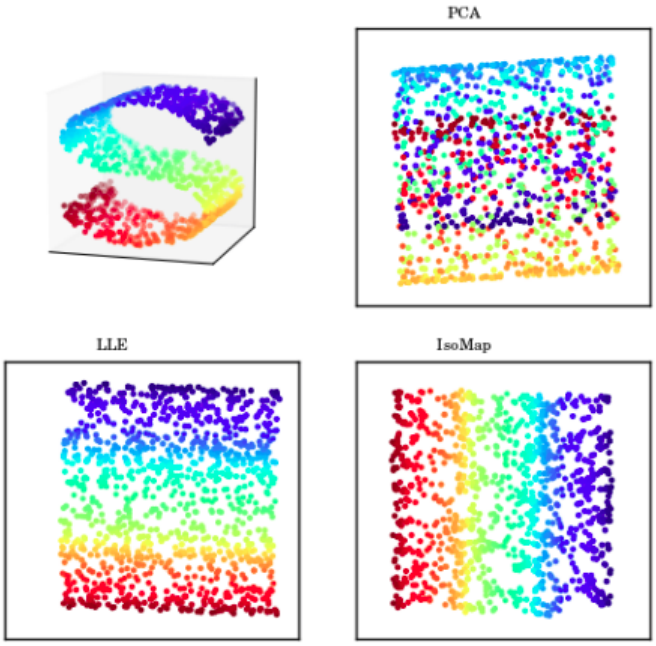


Figura 3.12: Projeções do Rolo-Suíço pelo PCA, ISOMAP e LLE. Extraída de (IVEZIĆ et al., 2014; PEDREGOSA et al., 2011)

O exemplo do Rolo-Suíço evidencia uma possível limitação de métodos lineares. Entretanto vale ressaltar também as limitações dos métodos não-lineares. Primeiramente, tendo em vista a evidência apresentada de que a DR não-linear é mais adequada quando há dúvida sobre a natureza estrutural dos dados, seria plausível então questionar por que sequer considerar usar o PCA em algum caso. Porém, note que, se para um dado problema os dados já residirem originalmente em um hiperplano, a redução linear é a melhor escolha devido à simplicidade, velocidade de processamento e escalabilidade. Mais precisamente:

- Complexidade computacional: o PCA é mais eficiente tanto em tempo de execução quanto em uso de memória do que métodos de DR não-linear. Esses fatores são cruciais quando se trata de conjunto de dados grandes (o que é comum nas diversas áreas de aplicação). Mesmo implementações simples do PCA funcionam em grandes conjuntos de dados.

- Pertinência da linearidade: em certas ocasiões, os dados de fato definem uma estrutura

quase linear em um espaço de menor dimensão. Nesses casos, o PCA tende a fornecer uma aproximação suficiente mesmo que a estrutura não seja perfeitamente linear.

- Facilidade de uso: PCA é de uso direto. Dada uma implementação particular, o único parâmetro a ser dado é a dimensionalidade-alvo. Técnicas não-lineares tipicamente requerem a seleção de mais de um parâmetro. O ajuste de parâmetros pode aumentar ainda mais o custo computacional desses métodos. É necessária também a seleção entre as diversas técnicas não-lineares disponíveis e essa escolha nem sempre é óbvia. Diferentes técnicas funcionam bem em diferentes circunstâncias e pode-se não conhecer a priori qual é mais apropriada.

- Mapeamento direto: o PCA fornece uma transformação que pode ser aplicada em dados que não faziam parte do conjunto de treinamento (no caso de uma tarefa de classificação por exemplo). Algumas técnicas não-lineares não possuem esse mapeamento direto.

- Pré-processamento: no caso do LLE por exemplo, o PCA se faz necessário em algumas situações como uma ferramenta de pré-processamento para viabilizar a execução do método não-linear. Isso acontece em outras situações também.

- *Overfitting*: no caso de uma tarefa de classificação, é sabido que as técnicas não-lineares possuem certa tendência ao overfitting devido ao aumento de complexidade que o modelo não-linear possui.

- Interpretabilidade: em alguns casos, pode-se desejar usar DR para se entender o processo de geração dos dados. Os pesos do PCA facilitam a conclusão a respeito da dimensão original, o que não é o caso de muitos métodos não-lineares.

- Popularidade: PCA é um método antigo, confiável e bem conhecido pela comunidade. Métodos não-lineares são novos e ainda estão em processo de estudo a respeito da categorização de adequação em função dos tipos de problemas a serem resolvidos. Além disso, implementações desses métodos ainda não estão amplamente disponíveis.

A respeito dos problemas específicos da extração não-linear, vale ressaltar que melhores maneiras de se construir um grafo de vizinhança de X por técnicas que não a do KNN e da ε-vizinhança, ainda estão sendo investigadas. Mesmo na utilização da técnica KNN, ainda não há uma definição precisa a respeito da escolha do parâmetro K para tal. A definição de K é uma oportunidade para o analista incorporar um conhecimento a priori na tarefa e depende das características dos dados, como densidade de amostragem e geometria da estrutura.

Como o ISOMAP e o LLE são baseados em intuições diferentes, os problemas que cada um pode trazer também são de naturezas distintas. O ISOMAP tenta preservar as propriedades geométricas globais do conjunto de dados (i.e. separação entre pontos longínquos) com o custo de distorcer geometrias locais. Foi mostrado que o ISOMAP é vulnerável a erros de curto-circuito se o K é muito grande em relação a estrutura dos dados, ou se a presença de ruído fizer com que alguns pontos estejam ligeiramente fora do padrão da estrutura (BALASUBRAMANIAN; SCHWARTZ, 2002). Mesmo um pequeno erro de curto-circuito pode alterar muitas entradas na matriz D, o que pode levar a imersões totalmente diferentes. Por outro lado, se K é muito pequeno, o grafo de vizinhanças pode se tornar muito esparso para aproximar os caminhos entre pontos corretamente.

Em contraste, o LLE tenta preservar as propriedades geométricas locais, que são caracterizadas pelos coeficientes lineares locais de reconstrução. Ainda que as vizinhanças sejam sobrepostas, o acoplamento entre pontos distantes pode ser severamente enfraquecido se os dados são ruidosos, esparsos ou com baixa conectividade. Portanto, o problema mais comum do LLE é mapear pontos de entrada distantes em pontos de saída próximos. O algoritmo LLE tem três parâmetros importantes: a dimensionalidade intrínseca d, o número de vizinhos mais próximos K, e em alguns casos o parâmetro de regularização α. A DR com LLE é muito sensível a variações nesses parâmetros.

Se K é muito pequeno, o mapeamento não vai refletir nenhuma propriedade global; se é muito grande, o mapeamento vai perder o caráter não-linear e vai se comportar como o PCA (já que todo o conjunto de dados será interpretado como uma única vizinhança). Afirma-se que dado um valor de K, o LLE só será capaz de recuperar uma imersão de dimensão $d < K$, e quando $d > K$, o conjunto de dados possui baixa dimensionalidade intrínseca e cada ponto pode ser perfeitamente reconstruído de seus vizinhos - fazendo com que os pesos de reconstrução local não sejam mais unicamente definidos. De modo geral, para um bom funcionamento, o LLE espera uma curvatura e densidade de amostragem tal que cada ponto tenha na ordem de $2d$ vizinhos que definam um *patch* aproximadamente linear. Saul e Roweis (2003) afirmam que os resultados do LLE são tipicamente estáveis para uma série de valores de K. Se α é incorreto, a análise espectral pode não convergir (RIDDER; DUIN, 2002). Se d é muito grande, o mapeamento vai realçar ruído; se é muito baixo, partes distintas do conjunto de dados podem ser mapeadas umas sobre as outras. A seguir vamos discutir mais a respeito da dimensionalidade intrínseca.

3.3.5 Dimensionalidade intrínseca

Diz-se que um conjunto de dados $X \subset R^D$ possui dimensionalidade intrínseca $d \leq D$, se X pode ser descrito em termos de d parâmetros livres. A interpretação geométrica é a de que o conjunto de dados em sua totalidade reside em uma hipersuperfície d-dimensional em R^D (THEODORIDIS; KOUTROUMBAS, 2008). A detecção da dimensionalidade intrínseca é um problema em investigação na área de DR. Não há uma solução única ou geral pois, há influência das características de cada conjunto de dados. Diferentes métodos foram propostos e a exibição de todos foge do escopo desse trabalho. Para o leitor interessado no assunto, recomenda-se a obtenção de maiores detalhes em (MINKA, 2000; CAMASTRA, 2003; FAN et al., 2010; SUN; MARCHAND-MAILLET, 2014; HE et al., 2014; CERUTI, 2014; CAMPADELLI et al., 2015).

Comentaremos o tema brevemente nessa subseção devido à sua importância em qualquer problema de DR. Se faz necessário definir uma dimensionalidade-alvo razoável que garanta suficientemente a retenção de boa parte da variabilidade dos dados, caso contrário, uma representação inadequada compromete os demais processos da análise. Por outro lado, uma dimensionalidade-alvo significativamente maior do que a intrínseca é igualmente indesejável. Se uma estrutura unidimensional por exemplo estiver imersa em um espaço tridimensional, uma redução de três para duas dimensões poderia representar a estrutura na forma de uma curva no plano, distorcendo assim sua característica originalmente linear. Portanto é importante que a dimensionalidade-alvo não seja muito inferior nem muito superior à intrínseca.

Algumas estratégias são empíricas. Zheng et al. (2016) testa exaustivamente a acurácia obtida em sua tarefa de classificação em função de uma série de valores de dimensão para observar o ponto de estabilização, conforme Figura 3.13 (que nesse caso se deu com dez dimensões).

A estratégia escolhida em nosso estudo também é empírica: a razão da soma dos maiores autovalores da respectiva matriz do método de DR sobre a soma de todos autovalores (Equação 3.79). Considera-se a matriz de covariância para o PCA, a matriz B para o ISOMAP e a matriz M para o LLE

$$\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i} \tag{3.79}$$

Procura-se os d primeiros maiores autovalores onde o valor da razão ultrapasse um certo limiar, que na literatura e na maior parte dos experimentos existentes varia entre $0,70$ e $0,95$. Esse valor da razão é entendido como a porcentagem retida da variância total dos dados. Portanto d é a quantidade de autovetores necessários para se reter essa porcentagem de variância, isto é, o número de dimensões mínimo suficiente para projeção (ou ainda, os autovetores associados aos maiores autovalores). Essa abordagem é simples e amplamente empregada em

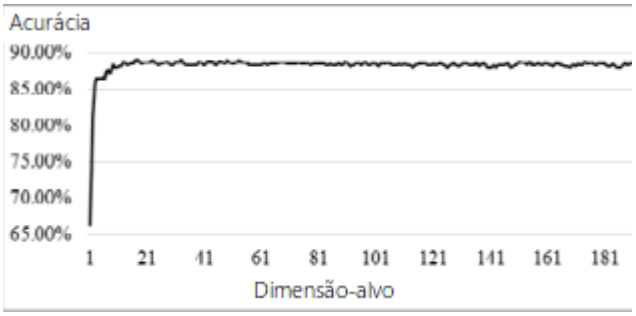


Figura 3.13: Acurácia em função da dimensionalidade. Extraída de (ZHENG et al., 2016)

aplicações de PCA e MDS, sendo inclusive sugerida em (COX; COX, 2001).

Esse mesmo critério é citado em (IVEZIĆ et al., 2014) subseção 7.3.2 , (WEBB, 1999) subseção 9.3.1, e (WANG; CHANG, 2006). É apresentada nessas referências a relação entre, o valor de corte da fração, e o ponto de inflexão do gráfico dos autovetores pelos autovalores. No caso do PCA e MDS, a plotagem da sequência de autovetores pela variância gera um gráfico decrescente e o ponto de inflexão é a dimensionalidade procurada. No ISOMAP, assim como no PCA e MDS, a verdadeira dimensionalidade pode ser estimada do decrescimento no erro conforme o aumento da dimensionalidade-alvo (TENENBAUM; SILVA; LANGFORD, 2000). A Figura 3.14 permite a visualização desse fato assim como a relação entre a curva de estimação e o conjunto de dados particular.

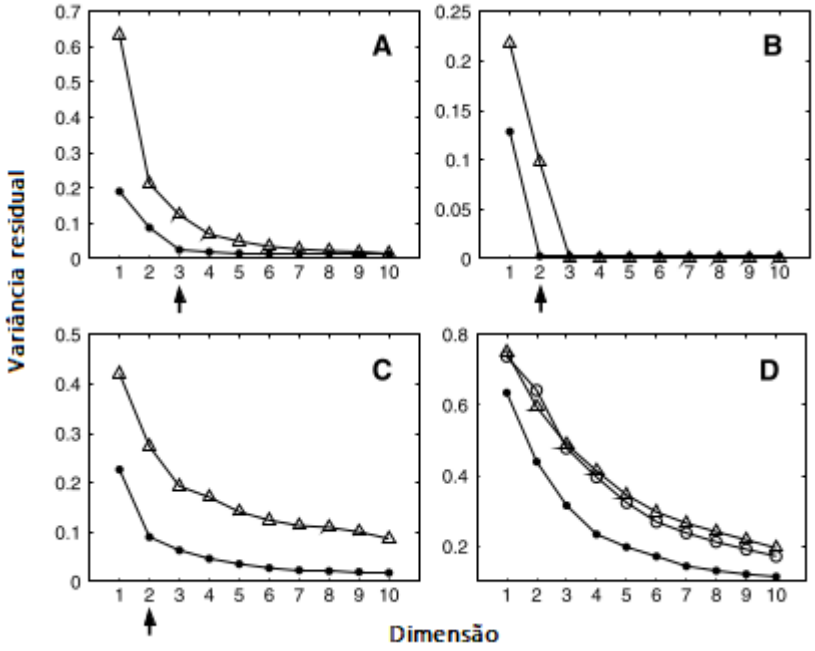


Figura 3.14: Variância residual em diferentes conjuntos de dados (A-D) do PCA (triângulos abertos), MDS (triângulos abertos de A-C; círculos abertos em D) e ISOMAP (círculos preenchidos). As setas indicam a dimensionalidade verdadeira quando conhecida. Extraída de (TENENBAUM; SILVA; LANGFORD, 2000)

Vale reforçar que essa fórmula foi escolhida devido à sua popularidade e simplicidade, porém é sabido que existem dificuldades na determinação do ponto de inflexão em alguns casos - conforme abordado nas referências citadas anteriormente e como indica a Figura 3.14 D - e também problemas em relação à sua efetividade - exibidos em (CHANG, 2003; CHANG; DU, 2004; RAMAKRISHNA et al., 2006)[3].

No LLE essa mesma análise do ponto de inflexão também pode ser feita, porém o gráfico é crescente (pois o cálculo das novas coordenadas envolve a minimização de uma função). Tal minimização é dada pelos menores autovalores da matriz M (i.e. os autovalores não-nulos mais próximos de 0). Contudo, anomalias podem existir (assim como para o PCA e o ISOMAP) no LLE por sensibilidade a ruído por exemplo (POLITO; PERONA, 2002; GOLDBERG; RITOV, 2008). Uma forma de amenizar a influência de ruído é se aplicar primeiro o PCA (reduzindo as componentes em 25% por exemplo), para depois aplicar LLE - conforme afirmado em (LEE; VERLEYSEN, 2007) subseções 7.2.3 e 7.2.4. Outro problema particular do LLE também é citado em (SAUL; ROWEIS, 2003) na seção 5.3, onde se afirma que tal estratégia só funciona para dados que residam em uma estrutura essencialmente linear ou que foram amostrados de maneira uniforme (o que não é garantido para todos os casos). Algumas alternativas para contornar o problema são apresentadas nessa referência. Como ilustração, Saul e Roweis (2003) apresentam na Figura 3.15 como o ponto de inflexão não é evidente em alguns casos onde é sabido que a dimensionalidade intrínseca é 2.

3.4 *Bootstrap aggregating*

Dado um conjunto de dados X com n elementos, a técnica *bootstrap aggregating* (BREIMAN, 1996) (também conhecida por *bagging*), tem como função gerar, por amostragem aleatória uniforme de X (com ou sem reposição), m novos conjuntos X_i, cada um com n' elementos. Uma mesma tarefa de aprendizado de máquina é aplicada em cada subconjunto e os resultados de cada tarefa são então combinados por um critério conveniente ao estudo particular (e.g. média, votação). Essa combinação é uma aproximação ou estimação do resultado que o procedimento teria se fosse aplicado sobre o conjunto X original. A técnica pode ser encarada como uma abordagem de divisão e conquista. A Figura 3.16 (adaptada de https://towardsdatascience.com) ilustra o procedimento.

[3] como o foco do trabalho não é a investigação dessas medidas, vamos apenas nesse momento citar a existência desses problemas (que poderão ser investigados futuramente)

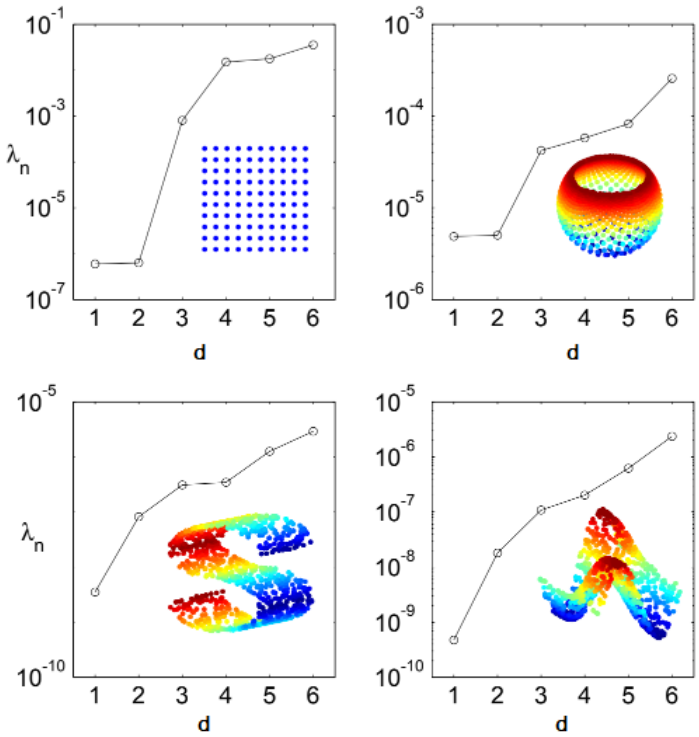


Figura 3.15: Note como o ponto de inflexão não é evidente especialmente para as duas formas inferiores. Extraída de (SAUL; ROWEIS, 2003)

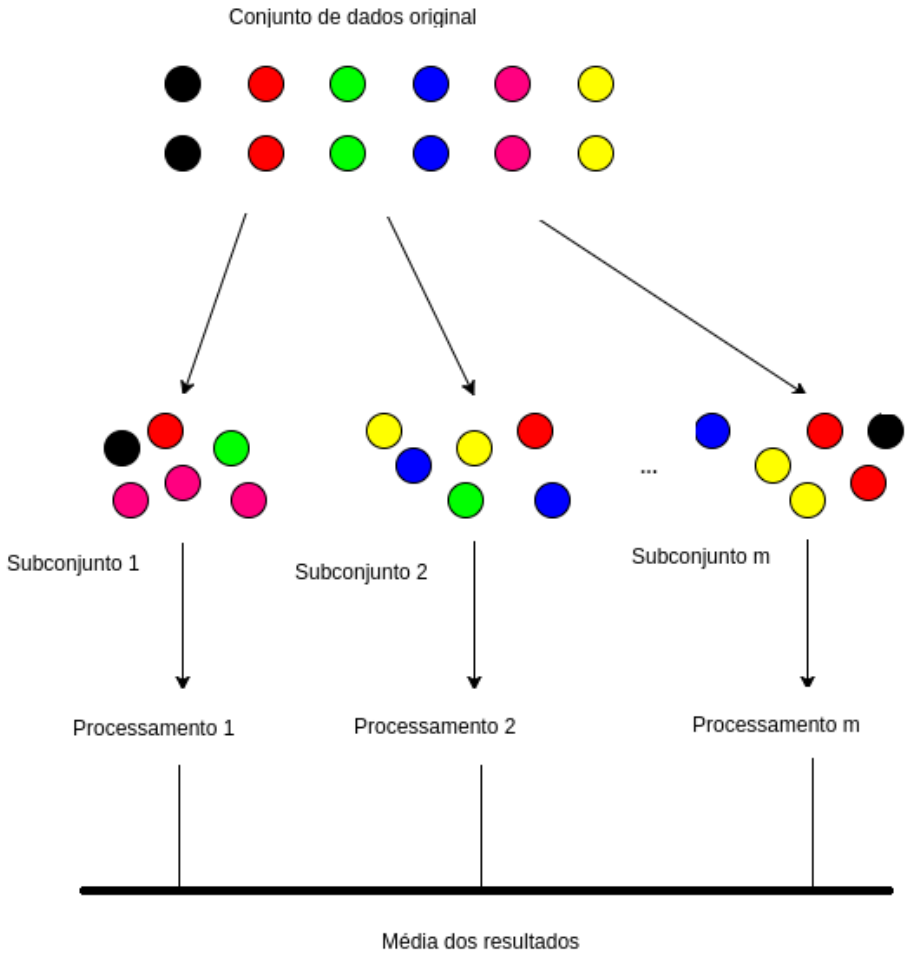


Figura 3.16: *Bootstrap aggregating* - ilustração do procedimento

3.5 Avaliação de agrupamento

Avaliar algoritmos de DR ainda é um problema. A metodologia usual consiste em executar o algoritmo em algum conjunto de dados artificial e analisar visualmente se os resultados são intuitivamente adequados (como ilustrado na Figura 3.12). Esse procedimento é subjetivo e medidas objetivas são necessárias para permitir uma análise quantitativa. Para realizar uma avaliação experimental robusta, se faz necessário conhecer a Verdade Terrestre[4] (*Ground Truth* - GT) e determinar se a informação está sendo preservada.

Nesse trabalho, tal determinação será feita pela comparação entre o resultado do procedimento de agrupamento por GMM e o GT. A seguir serão apresentadas as medidas de avaliação de agrupamento escolhidas que serão usadas para tal comparação. Todas possuem a propriedade de que quanto maior o valor retornado, melhor é o desempenho do resultado de agrupamento avaliado.

3.5.1 *Kappa*

O coeficiente *Kappa* foi originalmente proposto por (COHEN, 1960) como um método para determinar concordância entre especialistas. No contexto de uma tarefa de agrupamento, o coeficiente determina um grau de concordância *a posteriori*. Ou seja, dadas amostras previamente rotuladas em classes (GT), ele mede qual é a concordância entre a rotulação prévia e o resultado do agrupamento dado pela partição do conjunto de dados em grupos. A análise dessa estatística indica que, para valores negativos, não há concordância alguma e, quando o valor resultante é igual a 1, a concordância é total. Uma possível interpretação é a de que o coeficiente expressa a proporção de erros que o agrupamento evita cometer quando comparado a um resultado puramente aleatório. Assim, seja

L quantidade de classes

k quantidade de grupos no conjunto todo[5]

m_{ij} quantidade de elementos da classe i pertencentes ao grupo j[6]

[4] i.e. classes verdadeiras e seus pixels associados ou, rotulação feita pelo especialista de domínio

[5] note que nesse trabalho $k = L$

[6] há um requisito de correspondência semântica entre os rótulos do agrupamento e da classificação prévia. O apêndice B fornece a fundamentação teórica da solução que adotamos para garantir tal requisito

C matriz de Confusão definida como

$$C = \begin{bmatrix} m_{11} & \cdots & m_{1L} \\ \vdots & \ddots & \vdots \\ m_{L1} & \cdots & m_{LL} \end{bmatrix} \tag{3.80}$$

m_{i+} soma dos elementos da linha i

m_{+j} soma dos elementos da coluna j (i.e. quantidade de pontos no grupo j)

n total de observações (i.e. quantidade de pontos no conjunto todo)

então a expressão do cálculo do coeficiente é dada por (CONGALTON, 1991)

$$Kappa = \frac{n\sum_{i=1}^{L} m_{ii} - \sum_{i=1}^{L} m_{i+}m_{+j}}{n^2 - \sum_{i=1}^{L} m_{i+}m_{+j}} \tag{3.81}$$

3.5.2 *Entropy e Purity*

Adicionalmente às definições anteriores, seja

p_{ij} probabilidade de um ponto do grupo j pertencer à classe i

$$p_{ij} = \frac{m_{ij}}{m_{+j}} \tag{3.82}$$

e_j entropia de um grupo j

$$e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij} \tag{3.83}$$

$purity_j$ pureza de um grupo j

$$purity_j = \max p_{ij} \tag{3.84}$$

então, os coeficientes *Entropy* e *Purity* (SHANNON, 1948) são dados por

$$e = \sum_{j=1}^{k} \frac{m_{+j}}{n} e_j \tag{3.85}$$

$$purity = \sum_{j=1}^{k} \frac{m_{+j}}{n} purity_j \tag{3.86}$$

onde e é entropia total do conjunto de grupos e $purity$ é a pureza geral de um agrupamento. Como os nomes sugerem, essas medidas mensuram a variação e homogenidade intra-grupo (ou sobreposição inter-grupos) e consequente incerteza, aleatoriedade e desordem do agrupamento.

Ambas são baseadas em p_{ij}, que pode ser interpretada como um grau de pertinência do grupo j à classe i. Note que a entropia será igual a zero quando todas as suas parcelas forem iguais a zero, ou seja, quando $e_j = 0, \forall j$. Observando o termo e_j, notamos que o mesmo assume valor zero quando

- $p_{ij} = 0$ (i.e. o grupo não compartilha ponto algum com determinada classe)
- $\log_2 p_{ij} = 0 \Rightarrow p_{ij} = 1$ (i.e. o grupo compartilha todos os pontos com determinada classe)

A pureza resulta valores em $[0,1]$. Já a entropia resulta valores em $[-\log_2 L, 0]$ e assume valor máximo quando, $\forall$ grupo, todos os pontos do grupo pertencem à mesma classe. Informalmente: quanto maior o valor da medida, mais bem definida é a partição gerada pelo agrupamento; por outro lado, quanto menor o valor da medida, mais incerto é o agrupamento (i.e. há uma indicação de ausência de estrutura de agrupamento ou inabilidade do algoritmo de agrupamento em revelar tal estrutura).

3.5.3 *Rand* e *Jaccard*

Dado o resultado de um agrupamento e uma classificação prévia, um par de pontos pode ser unicamente rotulado como

SS se ambos os pontos pertencem ao mesmo grupo e à mesma classe

SD se pontos pertencem ao mesmo grupo porém à classes distintas

DS se pertencem a grupos distintos porém à mesma classe

DD se pertencem a grupos distintos e à classes distintas

Sejam a, b, c, d as quantidades de pares SS, SD, DS, DD, respectivamente. As medidas R (RAND, 1971) e J (JACCARD, 1912) são definidas como[7]

$$R = (a+d) \,/\, (a+b+c+d) \tag{3.87}$$

$$J = a/(a+b+c) \tag{3.88}$$

Portanto, essas medidas se baseiam na comparação entre discordância e concordância. A ideia é a de que, os grupos podem ser definidos tanto pelos pontos que contêm, como pelos que nao

[7] os termos a, b, c, d foram utilizados para facilitar o entendimento; a correspondência entre esses termos e os elementos da matriz de Confusão pode ser encontrada em (JAIN; DUBES, 1988) páginas [172,175]

contêm. Assim, se os pontos de um par estão atribuídos ao mesmo grupo e à mesma classe, isso representa concordância entre agrupamento e classificação prévia. Da mesma forma, se os pontos do par pertencem a grupos distintos e classes distintas, isso também representa concordância entre agrupamento e classificação prévia. A avaliação R é baseada na sobreposição entre as medidas de concordância SS e DD. A avaliação J é similar, porém considera somente a concordância SS (o termo d não está presente em seu cálculo).

Capítulo 4

EXPERIMENTOS

Os detalhes das imagens utilizadas são apresentados na seção 4.1; a metodologia do projeto será detalhada na seção 4.2; os resultados experimentais nas seções 4.3 e 4.4; o texto se encerra na seção 4.5 com as conclusões obtidas dos resultados e possibilidades de trabalhos futuros.

4.1 Conjunto de imagens

Sete HIs públicas tradicionais foram usadas para experimentação[1]: *Indian Pines, Salinas, SalinasA, Botswana, Kennedy Space Center, Pavia University, Pavia Centre*. Todas são imagens da superfície terrestre e foram obtidas por veículos aéreos ou satélites. Os equipamentos fonte são: AVIRIS, *Reflective Optics System Imaging Spectrometer* (ROSIS), *NASA EO-1 Satellite*.

As subseções seguintes trazem os detalhes de cada imagem respectivamente, com exceção da última subseção, que tem como papel somente sumarizar os aspectos principais de todas as imagens.

4.1.1 *Indian Pines*

Coletada em Junho de 1992 pelo sensor AVIRIS na região noroeste do estado americano de Indiana, a cena identificada como *Indian Pines* é composta de 145×145 pixels e 224 bandas de refletância com comprimento de onda no intervalo de 0,4 a 2,5 μm. Utilizamos a versão corrigida da imagem cujas bandas que cobrem regiões de absorção de água foram removidas (bandas [104-108], [150-163], 220). Uma visualização em tons de cinza é dada na Figura 4.1. Uma visualização em falsa composição *Red Green Blue* (RGB) é dada na Figura 4.2. Dois terços da área total correspondem a áreas de agricultura e um terço de floresta e outras vegetações

[1] http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral Remote Sensing Scenes

naturais. Existem duas rodovias principais de mão dupla e um trilho de trem, assim como pequenos conjuntos de moradias e outras construções e ruas. Haja vista que a captura se deu em Junho, algumas plantações como milho e soja se encontram no estágio inicial de crescimento, com menos de 5% de cobertura. O GT é dividido em dezesseis classes além da classe 0 e pode ser visualizado pelo mapa de rótulos da Figura 4.3. A Tabela 4.1 apresenta a descrição e a quantidade de pixels de cada classe. A função do mapa de rótulos é exibir na dimensão espacial de banda simples, não o valor de refletância de cada pixel em uma determinada banda, mas sim uma coloração ou tonalidade correspondente ao rótulo da classe associada a cada pixel. A correspondência coloração-numeração do rótulo da classe é dada pela legenda vertical à direita no mapa. Essa legenda substitui os intervalos do espectro comumente apresentado nas representações de refletância em dimensão espacial de banda simples (presente nas Figuras 4.20 e 4.22 por exemplo).

Figura 4.1: *Indian Pines* - imagem em tons de cinza. Extraída de (BAUMGARDNER; BIEHL; LANDGREBE, 2015)

Figura 4.2: *Indian Pines* - falsa composição RGB. Extraída de (SAQUI, 2018)

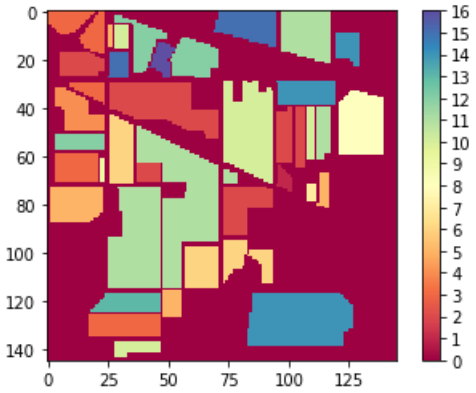


Figura 4.3: *Indian Pines* - mapa de rótulos do GT

Rótulo	Descrição	Quantidade de pixels
1	Alfafa	46
2	Milho - primeira fase	1428
3	Milho - segunda fase	830
4	Milho - terceira fase	237
5	Grama - pastagem	483
6	Grama - árvores	730
7	Grama - pastagem cortada	28
8	Feno	478
9	Aveia	20
10	Soja - primeira fase	972
11	Soja - segunda fase	2455
12	Soja - terceira fase	593
13	Trigo	205
14	Bosques	1265
15	Construções - Grama - Árvores - Ruas	386
16	Rochas - Estruturas Férreas - Edifícios	93

Tabela 4.1: *Indian Pines* - tabela de classes

4.1.2 *Salinas*

Essa cena foi coletada pelo sensor AVIRIS sobre o Vale Salinas na Califórnia e possui uma alta resolução espacial (3,7 metros por pixel). A área coberta inclui 512×217 pixels. Foram descartadas vinte bandas de absorção de água ([108-112], [154-167], 224) e portanto foram utilizadas 204 bandas. Uma visualização em tons de cinza é dada na Figura 4.4. Uma

visualização em falsa composição RGB é dada na Figura 4.5. A imagem inclui vegetais, terra nua e campos de vinhedos. O GT inclui 16 classes (Figura 4.6, Tabela 4.2).

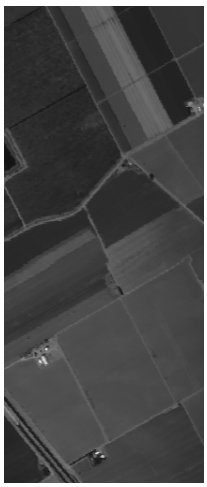


Figura 4.4: *Salinas* - imagem em tons de cinza. Extraída de (MYASNIKOV, 2018)

Figura 4.5: *Salinas* - falsa composição RGB. Extraída de (SAQUI, 2018)

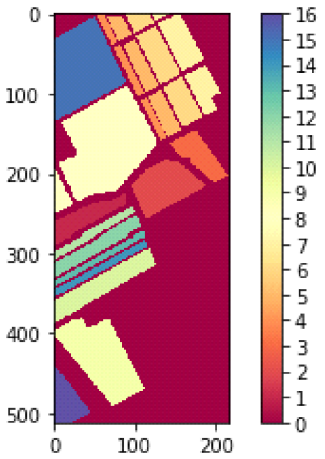


Figura 4.6: *Salinas* - mapa de rótulos do GT

Rótulo	Descrição	Quantidade de pixels
1	Brócolis verde erva 1	2009
2	Brócolis verde erva 2	3726
3	Pousio	1976
4	Arado de pousio	1394
5	Pousio liso	2678
6	Restolho	3959
7	Salsão	3579
8	Uva	11271
9	Solo para plantio de vinhedo	6203
10	Milho verde ervas velhas	3278
11	Alface quarta semana	1068
12	Alface quinta semana	1927
13	Alface sexta semana	916
14	Alface sétima semana	1070
15	Vinhedo em crescimento	7268
16	Vinhedo treliças verticais	1807

Tabela 4.2: *Salinas* - tabela de classes

4.1.3 *SalinasA*

*Salinas*A é a menor imagem dentre as utilizadas e é um recorte da imagem *Salinas*. Uma visualização em tons de cinza é dada na Figura 4.7. Essa imagem possui apenas 86×83 pixels e 6 classes, descritas na Tabela 4.3 e mapeadas na Figura 4.8.

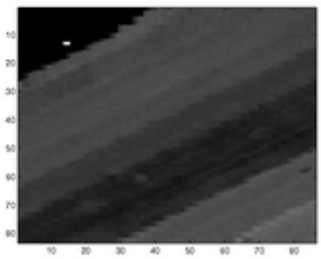


Figura 4.7: *SalinasA* - imagem em tons de cinza. Extraída de (GRANA; VEGANZONS; AYERDI, 2020)

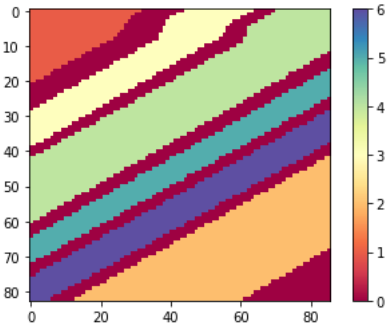


Figura 4.8: *SalinasA* - mapa de rótulos do GT

Rótulo	Descrição	Quantidade de pixels
1	Brócolis verde erva 1	391
2	Milho verde ervas velhas	1343
3	Alface quarta semana	616
4	Alface quinta semana	1525
5	Alface sexta semana	674
6	Alface sétima semana	799

Tabela 4.3: *SalinasA* - tabela de classes

4.1.4 *Pavia Centre*

Existem duas cenas adquiridas pelo sensor ROSIS durante uma campanha de voos sobre a cidade de Pavia no norte da Itália: *Pavia Centre* e *Pavia University.*

Pavia Centre possui 102 bandas e 1096×1096 pixels com resolução geométrica de 1,3 metros. Uma visualização em tons de cinza é dada na Figura 4.9. Alguns pixels não continham informação e tiveram de ser descartados antes da constituição da imagem final. Na Figura 4.10 observa-se um setor escuro na região desses pixels. Uma visualização em falsa composição RGB é dada na Figura 4.11. Nove classes estão presentes (Figura 4.12 e Tabela 4.4).

Figura 4.9: *Pavia Centre* - imagem em tons de cinza. Extraída de (PLAZA et al., 2005)

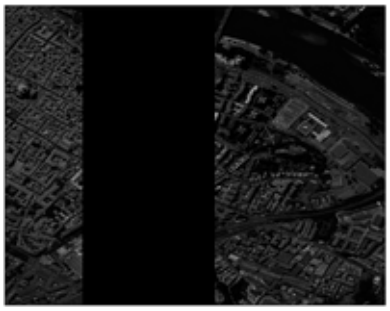


Figura 4.10: *Pavia Centre* - faixa de pixels descartados. Extraída de (GRANA; VEGANZONS; AYERDI, 2020)

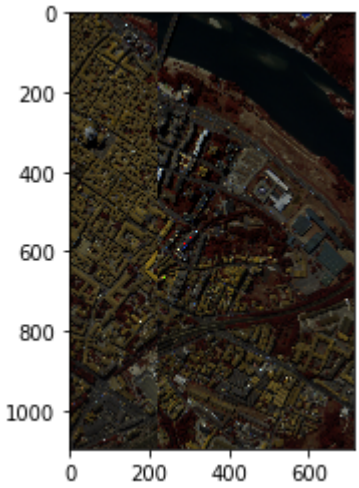


Figura 4.11: *Pavia Centre* - falsa composição RGB

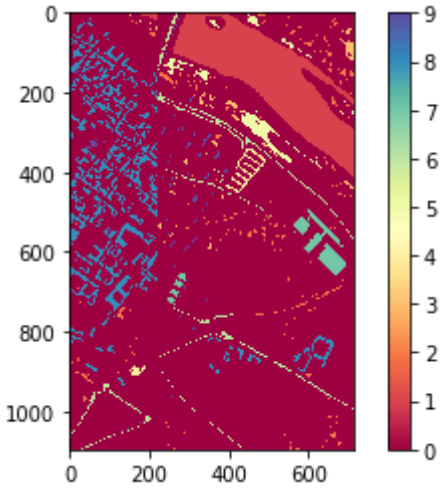


Figura 4.12: *Pavia Centre* - mapa de rótulos do GT

Rótulo	Descrição	Quantidade de pixels
1	Água	824
2	Árvores	820
3	Asfalto	816
4	Tijolos	808
5	Betume	808
6	Telhado	1260
7	Sombras	476
8	Prado	824
9	Solo nu	820

Tabela 4.4: *Pavia Centre* - tabela de classes

4.1.5 *Pavia University*

Esse conjunto de dados foi capturado sobre a Universidade de Pavia na Itália pelo instrumento aéreo ROSIS-03. O sensor possui 115 canais com cobertura espectral entre 0,43 e 0,86 µm. Doze canais foram removidos por causa da existência de ruído e os 103 canais espectrais restantes foram processados para compor a imagem. Os dados foram corrigidos atmosfericamente, mas não geometricamente. A resolução espacial é de 1,3m por pixel. A imagem possui 610×610 pixels, porém, assim como *Pavia Centre*, alguns pixels foram descartados (Figura 4.13). Uma visualização em tons de cinza é dada na Figura 4.14. Duas possíveis visualizações em falsa composição RGB são dadas na Figura 4.15. O conjunto de dados cobre a Escola

de Engenharia e consiste em 9 diferentes classes incluindo árvores, asfalto, betume, cascalho, trechos metálicos, sombra, tijolos, prado e solo (Figura 4.16 e Tabela 4.5).

Figura 4.13: *PaviaU* - faixa de pixels descartados. Extraída de (GRANA; VEGANZONS; AYERDI, 2020)

Figura 4.14: *PaviaU* - imagem em tons de cinza. Extraída de (GHAMISI et al., 2017)

Figura 4.15: *PaviaU* - falsa composição RGB. Extraída de (GHAMISI et al., 2017) e (BIOUCAS-DIAS et al., 2013)

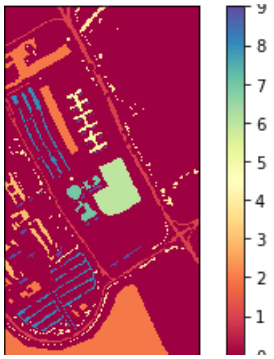


Figura 4.16: *PaviaU* - mapa de rótulos do GT

Rótulo	Descrição	Quantidade de pixels
1	Asfalto	6631
2	Prado	18649
3	Cascalho	2099
4	Árvores	3064
5	Forros metálicos	1345
6	Solo nu	5029
7	Betume	1330
8	Tijolos	3682
9	Sombras	947

Tabela 4.5: *PaviaU* - tabela de classes

4.1.6 *Kennedy Space Center*

Essa cena foi capturada pelo sensor AVIRIS sobre o Centro Espacial Kennedy na Flórida em 23 de Março de 1996 a uma altitude de aproximadamente 20 kilômetros com resolução espacial de 18 metros. A cena possui 512×614 pixels. Após a remoção de bandas correspondentes à absorção de água e bandas com baixa relação sinal-ruído , 176 bandas foram utilizadas para compor a imagem. Uma visualização em tons de cinza é dada na Figura 4.17. Uma visualização em falsa composição RGB é dada na Figura 4.18. A Figura 4.20 exibe três diferentes bandas dentre as 176. Os dados de treinamento foram selecionados usando mapas de cobertura terrestre derivados de fotografias coloridas infravermelho fornecidas pelas imagens do Centro e do *Landsat Thematic Mapper*. O esquema de classificação da vegetação foi desenvolvido pelo pessoal do Centro em um esforço para definir os tipos funcionais que são discerníveis na resolução espacial do Landsat e esses dados do AVIRIS. A discriminação da cobertura terrestre para esse ambiente é difícil devido à similaridade de assinaturas espectrais para certos tipos de vegetação. Treze classes foram definidas[2]. A Figura 4.19 mostra o mapa de rótulos do GT.

[2]para as imagens *Kennedy Space Center* e *Botswana* não conseguimos obter a tabela de correspondência de classes como as exibidas nas imagens anteriores. Porém o arquivo usual da correspondência de classes (i.e rótulos genéricos [1,2,...,k]) foi provido, o que já é suficiente para todo o nosso processo.

Figura 4.17: *Kennedy Space Center* - imagem em tons de cinza. Extraída de (MYASNIKOV, 2018)

Figura 4.18: *Kennedy Space Center* - falsa composição RGB. Extraída de (SAQUI, 2018)

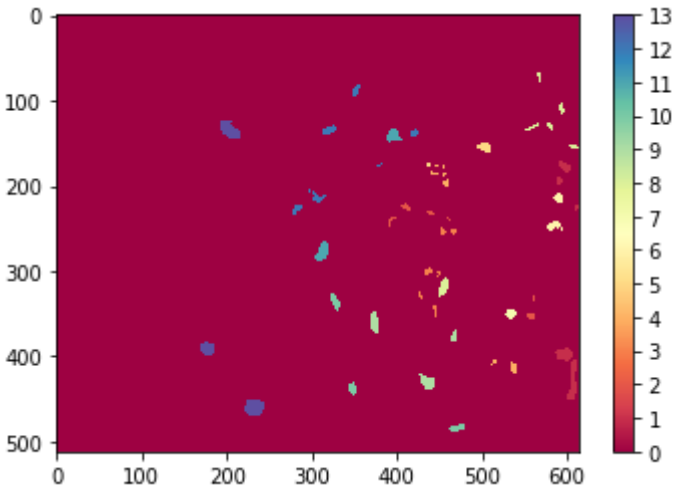


Figura 4.19: *Kennedy Space Center* - mapa de rótulos do GT

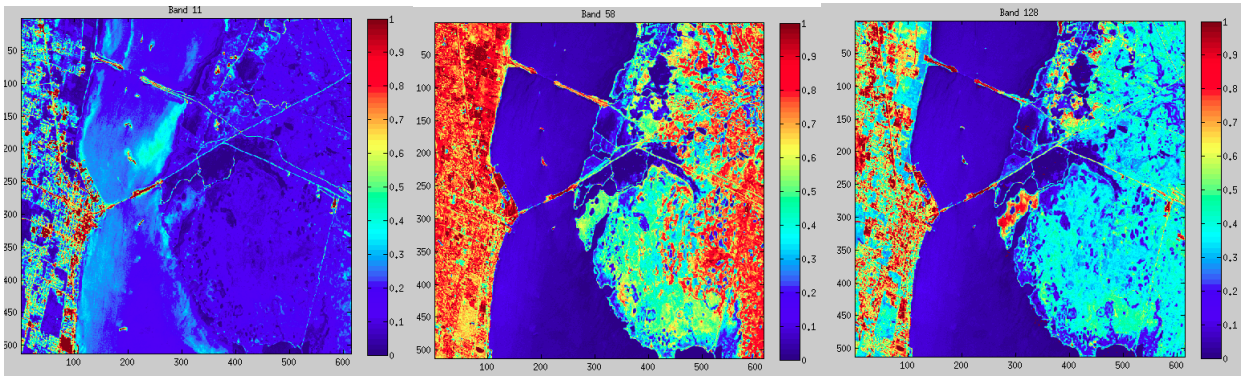


Figura 4.20: *Kennedy Space Center* - da esquerda para a direita: bandas 11, 58, 158. Disponível em http://www.ehu.eus/ccwintco/uploads/2/28/KSC.gif

4.1.7 *Botswana*

O satélite EO-1 da NASA adquiriu uma sequência de dados sobre o Delta do Okavango, Botswana entre 2001 e 2004. O sensor Hyperion EO-1 adquire dados com resolução de 30 m / pixel sobre uma faixa de 7,7 km em 242 bandas cobrindo a faixa de 400-2500 nm do espectro em janelas de 10 nm. O pré-processamento dos dados foi realizado pelo Centro de Pesquisas Espaciais da Universidade de Tenessee para mitigar os efeitos de detectores ruins, descalibração inter-detector e anomalias intermitentes. Bandas não calibradas e ruidosas correspondentes à absorção de água foram removidas e as 145 bandas remanescentes foram incluídas como candidatas: [10-55, 82-97, 102-119, 134-164, 187-220]. A cena possui 1476×256 pixels. Uma visualização em falsa composição RGB é dada na Figura 4.21. A Figura 4.22 exibe quatro bandas distintas. Quatorze classes representam os seguintes tipos de cobertura terrestre: pântanos sazonais, pântanos ocasionais e florestas mais secas localizadas na porção periférica do Delta. A Figura 4.23 mostra o mapa de rótulos do GT.

4.1.8 Resumo

Sumarizamos na Tabela 4.6 o tamanho em megabytes e as quantidades de pixels, classes (sem contabilizar a classe *background*) e bandas (após os respectivos descartes mencionados nas subseções anteriores) de cada imagem.

Figura 4.21: *Botswana* - falsa composição RGB. Extraída de (SAQUI, 2018)

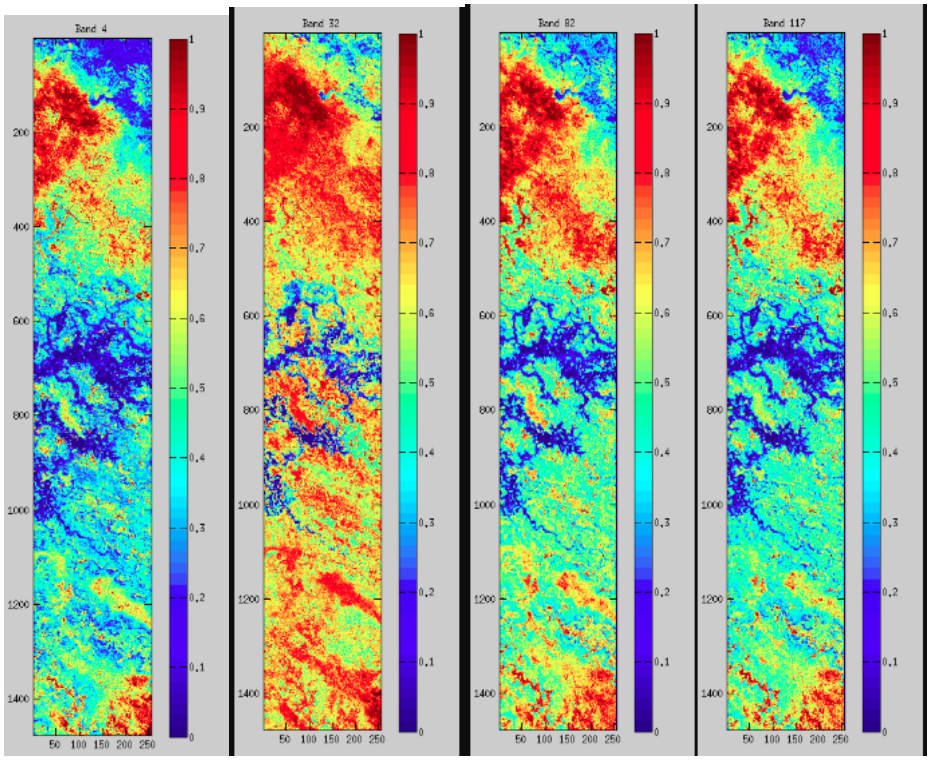


Figura 4.22: *Botswana* - bandas 4, 32, 82 e 117. Extraída de (GRANA; VEGANZONS; AYERDI, 2020)

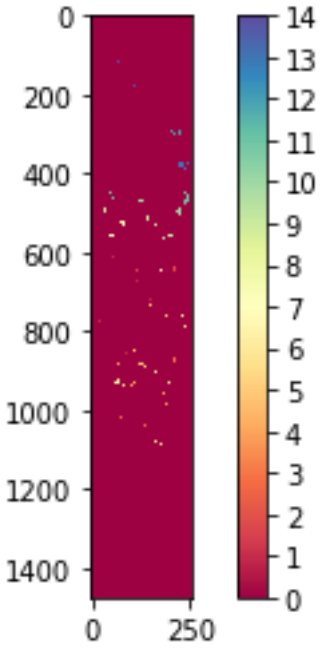


Figura 4.23: *Botswana* - mapa de rótulos do GT

Imagem	Pixels	Bandas	Tamanho MB	Classes
Indian Pines	145×145	204	5,7	16
Salinas	512×217	204	25,3	16
SalinasA	86×83	204	1,5	6
Pavia Centre	1096×1096	102	123,6	9
Pavia University	610×610	103	33,2	9
Kennedy Space Center	512×614	176	56,8	13
Botswana	1476×256	242	78,9	14

Tabela 4.6: Resumo das informações do conjunto de imagens

4.2 Metodologia

O objetivo deste trabalho portanto é investigar se a DR beneficia o agrupamento em HIs e, nesse caso, se a abordagem não-linear de DR é mais eficiente que a linear. Para conduzir tal investigação, foram comparados os desempenhos de agrupamentos feitos sem o uso de redução, com o uso de redução linear e com o uso de redução não-linear. Os algoritmos de extração de características PCA, ISOMAP e LLE foram executados em um conjunto de sete HIs. Sob cada uma dessas três distintas execuções e também sob a imagem original sem aplicação de DR (*No Dimensionality Reduction* - NDR), foram realizados agrupamentos pelo *K-Means* (FORGY,

1965; MACQUEEN, 1967; LLOYD, 1982) e o *Expectation Maximization* (EM) (DEMPSTER; LAIRD; RUBIN, 1977). Como o conjunto de imagens selecionado possui GT, os desempenhos dos agrupamentos foram mensurados pelas medidas externas *Rand*, *Jaccard*, *Kappa*, *Entropy* e *Purity*. Por fim, os desempenhos foram comparados pelos testes estatísticos de *Friedman* (FRIEDMAN, 1937) e *Nemenyi* (NEMENYI, 1963).

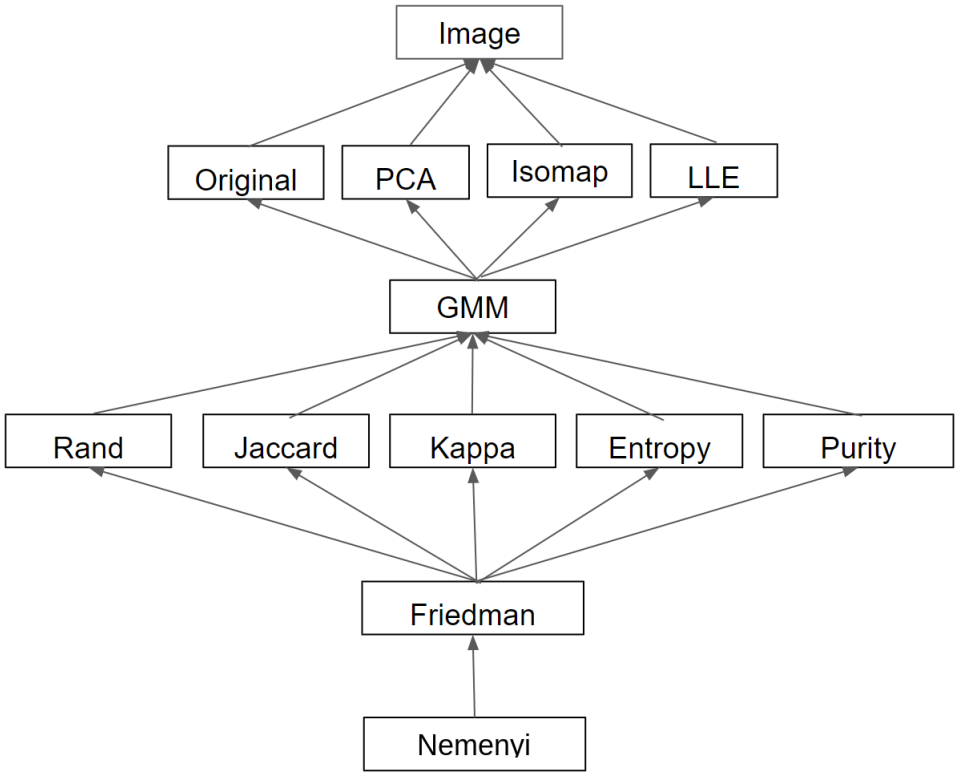


Figura 4.24: Etapas experimentais

Nos nossos experimentos, com exceção de *Indian Pines* e *SalinasA*, as imagens eram muito grandes para o hardware disponível[3]. A tentativa de execução de DR não-linear não era possível devido a falta de espaço de processamento. O ISOMAP por exemplo, em seu segundo passo requer, para todos os pontos, o cálculo e armazenamento da distância de um ponto aos demais, (o que é computacionalmente custoso). Para contornar esse problema, subconjuntos de pixels foram escolhidos com o uso da técnica *Bagging*. No nosso estudo, para cada imagem foi fixada uma porcentagem de pixels de cada classe e uma quantidade de amostragens. Os procedimentos seguintes do processo experimental levaram em consideração cada uma dessas amostragens para compor uma análise global. De modo geral, procurou-se para cada imagem selecionar aproximadamente vinte mil pixels (capacidade de processamento do hardware). Será mencionada nas subsubseções seguintes a quantidade de amostragens realizadas em cada imagem e a porcentagem do tamanho total da imagem que cada amostragem corresponde. As seguintes

[3]memória RAM 24GB

correspondências foram adotadas: 20% da imagem - 10 execuções; 10% - 20; 5% - 40; 2% - 100. O fato dos algoritmos de DR escolhidos se basearem somente no domínio espectral (e não espacial) quando aplicados em HIs favorece o uso dessa técnica de amostragem.

Sendo possível executar DR com essa redução amostral, a dimensionalidade-alvo de cada método para cada imagem foi estimada pela técnica abordada na Subsubseção 3.3.5. Como os métodos não-lineares necessitam do valor de dimensão-alvo como parâmetro, utilizamos para tal o valor de dimensão original de cada imagem. A estimação não requer múltiplas execuções dos algoritmos - é necessária somente a acumulação sequencial dos autovalores dessas matrizes (que já estão ordenados) sobre a soma total de autovalores até se atingir o limiar (95% para o PCA e ISOMAP e 1% para o LLE[4]). Seguindo a recomendação mencionada no último parágrafo da Subsubseção 3.3.5, adotamos para quantidade de vizinhos o valor de 2D (sendo D a dimensão original - que nesse caso também é a dimensão de parâmetro). Como o tema ainda está em estudo na área, experimentamos reduzir a quantidade de vizinhos para diminuir a conectividade do grafo e possivelmente obter melhores resultados. Essa diminuição é limitada (como comentado na mesma Subsubseção) e, de fato, valores muito baixos geraram erros de execução por falta de conectividade. As quantidades específicas de vizinhos utilizadas em cada imagem serão mencionadas nas próximas subseções.

A seguir o PCA, o ISOMAP e o LLE foram executados com as dimensões estimadas em cada imagem. Somente uma execução de cada método é necessária pois, fixados os parâmetros, os resultados de todos são determinísticos (invariantes). A quantidade de vizinhos utilizada foi a menor usada na estimação. A dimensão utilizada foi a correspondente a essa quantidade de vizinhos. Esses valores são mencionados para cada imagem nas próximas subseções.

Em todas as imagens existe uma classe *background* (que no GT é sempre representada pelo rótulo 0). Tal classe pode caracterizar regiões de transição que não puderam ser rotuladas pelo especialista ou que não foram relevantes para a rotulação. Como o *background* não está no interesse de discriminação da nossa tarefa, uma transformação foi aplicada definindo as coordenadas de todos os pixels dessa classe para (0,...,0). Essa transformação foi aplicada após a DR e antes do agrupamento, evitando assim distorções no agrupamento[5] e na geometria original para a DR.

Dada uma imagem portanto, (haja vista que para as demais o mesmo procedimento é repetido), sob cada uma das três distintas execuções (PCA, ISOMAP, LLE) e também sob a imagem

[4]como discutido na subsubseção 3.3.5, ainda não se dispõe de critérios objetivos teóricos para a definição precisa do limiar no caso do LLE. É comum esse valor ser determinado empiricamente e assim foi feito também nesse trabalho. Contudo, desejamos futuramente investigar possiveis melhorias e critérios para essa estimativa.

[5]foi verificado experimentalmente que, após o agrupamento, todos os pixels da classe 0 se mantiveram em um mesmo grupo

original sem aplicação de DR, foram realizados agrupamentos pelo *K-Means* e o EM. Esses algoritmos foram escolhidos pelas suas assunções de não-correlação de variáveis e dispersão em espaço Euclidiano, o que é garantido após a aplicação de todos os três métodos de DR utilizados (subseção 3.3.4). Tal escolha foi considerada a mais apropriada devido as discussões previamente exibidas no capítulo 1. Adicionalmente, vale lembrar que os algoritmos de agrupamento não são o foco da nossa investigação. Várias execuções foram realizadas na tentativa de minimizar o conhecido efeito de inicialização aleatória de centroides. A especificação de quantas execuções será dada individualmente para cada imagem nas próximas subsubseções pois, a mesma varia em função do número de amostragens *Bagging*.

Em relação a essa etapa de agrupamento, aqui vale uma breve ressalva teórica. Note que, formalmente, o *K-Means* é o caso mais simples do GMM, onde se assume matriz de covariância igual à identidade e portanto se estima somente o parâmetro "média" de cada gaussiana (grupo). Note também que o GMM emprega a técnica *Expectation Maximization* em seu processo de estimação dos parâmetros (consequentemente assim também o faz o *K-Means*). A comunidade de aprendizado de máquina muitas vezes referencia o GMM por EM quando da estimação das médias e variância de cada grupo. Portanto, por simplicidade e sem perda de rigor, há momentos que referenciamos a etapa de agrupamento do nosso processo somente pelo termo GMM.

O conjunto de imagens usado nesse trabalho felizmente dispõe da distribuição original das classes nos dados fornecida por especialistas (GT). A avaliação dos resultados das diferentes formas de extração de características adotadas, será feita pela aplicação de agrupamento nos mesmos e pela consequente comparação entre a estrutura fornecida pelo agrupamento e o GT. As medidas de avaliação de agrupamento adotadas foram aplicadas em todos as distintas execuções. O valor representativo escolhido de cada medida foi dado pelo máximo obtido (dentre todos os resultados dessa medida). Especificamente para o coeficiente *Kappa*, um emparalhemento entre os rótulos gerados pelo agrupamento e os rótulos do GT foi feito antes da aplicação do coeficiente. Portanto, ao final do processo experimental, dada uma imagem, obtiveram-se para cada método de DR (PCA, ISOMAP, LLE) e também para a não aplicação de DR, cinco valores de avaliação de agrupamento (cada uma dada pelo valor máximo de uma medida de avaliação específica).

O teste de *Friedman* foi aplicado para indicar se houve diferença estatística significativa entre algum par de métodos (levando em consideração cada possível par de métodos de DR distintos). Nos casos de detecção de diferença, o teste de *Nemenyi* foi aplicado para indicar os pares específicos que diferiram. Os resultados desses dois testes estatísticos permitem analisar se a DR foi benéfica para a tarefa de agrupamento e em quais imagens a DR não-linear superou

a linear. Os resultados dos testes são mostrados e analisados na seção 4.4.

A linguagem de programação utilizada foi Python[6]. Utilizamos também o SciPy[7], uma biblioteca de rotinas numéricas do Python que possui sub-pacotes para computação matemática e científica. Dentre esses pacotes, utilizamos o NumPy[8], que também é um pacote científico que facilita a manipulação de arrays e o uso de técnicas de álgebra linear. Outro pacote utilizado foi o Matplotlib[9], para plotagem de gráficos. O Scikit-learn[10] é uma ferramenta construída sobre os três pacotes supra-citados que possui uma infinidade de métodos de reconhecimento de padrões, inclusive os seguintes usados nesse projeto: PCA, ISOMAP, LLE, *K-Means*, EM, *Kappa*, *Jaccard*, *Rand*. A implementação dos testes de significância utilizada foi a da biblioteca STAC[11]. No pareamento de rótulos, a construção da matriz de custos foi implementada por nós e a implementação do algoritmo Húngaro utilizada foi a *scipy.optimize.linear_sum_assignment*. As técnicas *bootstrap aggregating*, *Entropy* e *Purity* também são de própria implementação. Todos os códigos e arquivos necessários para reproduzir os experimentos se encontram disponíveis em https://drive.google.com/open?id=1NJNw-kd-JNx4-NE5EKnD68oqoKQPfuda.

4.3 Resultados - *Bagging*, DR e agrupamento

Da mesma forma como fizemos na seção 4.1, apresentaremos nas próximas subseções, para cada imagem, os resultados referentes as etapas de *Bagging*, DR e agrupamento. Ao final de cada subseção é apresentada a respectiva tabela de valores máximos das medidas de validação de agrupamento por método de DR[12] (Tabelas 4.8 - 4.14), onde a última linha apresenta a média dos valores das medidas dado um método. Para se ter uma ideia comparativa também a respeito das dimensionalidades intrínsecas estimadas, sumarizamos na Tabela 4.7 tais valores para cada método de DR.

[6] versão 3.6.5 - https://www.python.org/downloads/release/python-365/
[7] versão 1.0.0 - https://www.scipy.org/scipylib/download.html
[8] versão 1.15.0 - https://www.numpy.org/
[9] versão 3.0.0 - https://matplotlib.org/
[10] versão 0.20 - https://scikit-learn.org/stable/
[11] http://tec.citius.usc.es/stac/doc/index.html
[12] truncamos todos os valores em três casas decimais ou na primeira casa decimal possível

Imagem	NDR	PCA	ISOMAP	LLE
Indian Pines	204	5	55	36
Salinas	204	2	10	47
SalinasA	204	2	4	42
Pavia Centre	102	2	20	22
Pavia University	103	3	17	18
Kennedy Space Center	176	47	110	65
Botswana	242	2	30	30

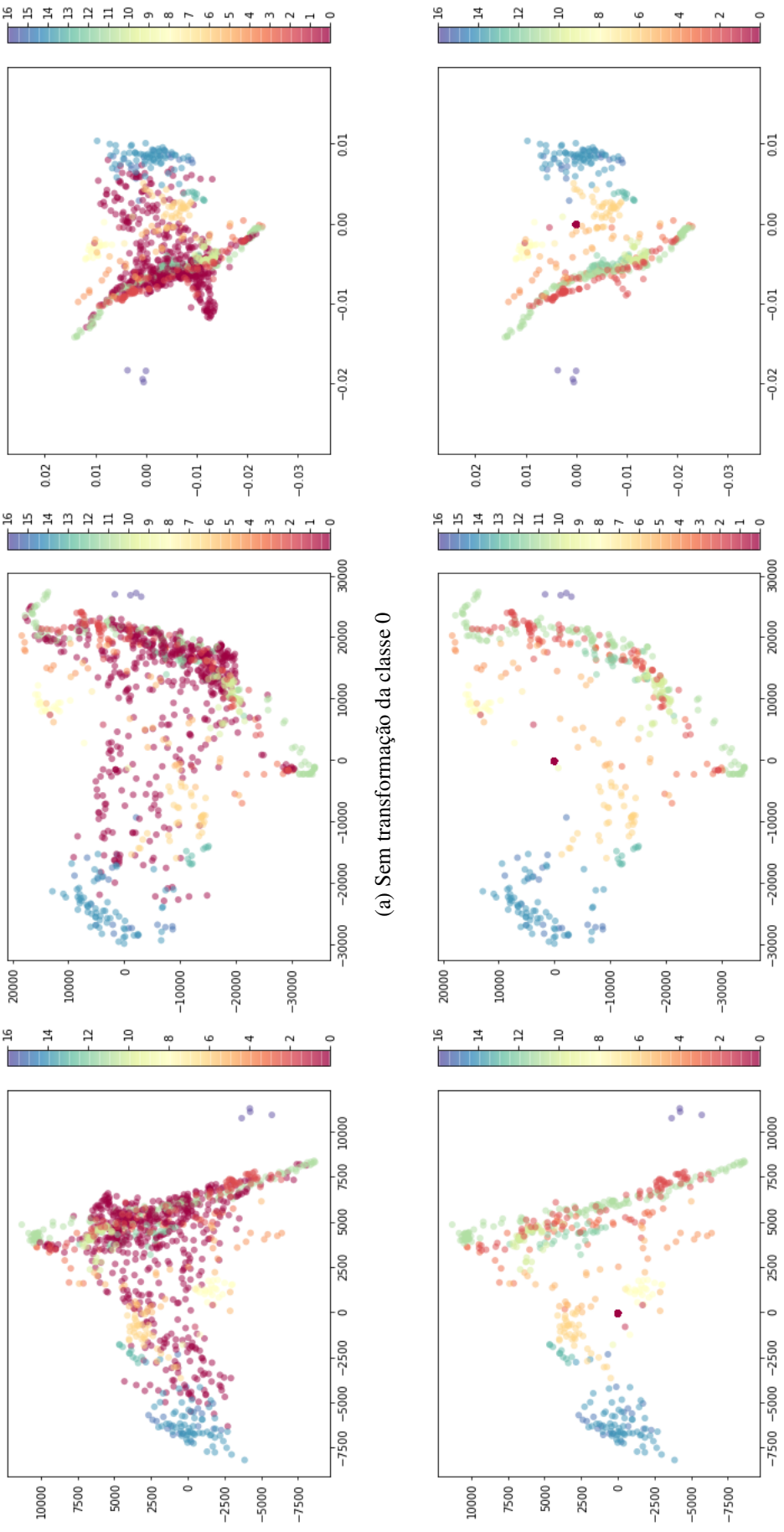
Tabela 4.7: Dimensionalidade estimada com menor número de vizinhos

4.3.1 *Indian Pines*

Com a aplicação da DR (linear ou não-linear), pode-se visualizar o GT na forma de gráficos de dispersão fixando a dimensão-alvo no valor 2. Esses gráficos não possuem função no procedimento geral de experimentação e servem somente para visualização. Esses gráficos representam o espaço bi-dimensional de características explicado na Figura 3.3. Note que o uso de todos os pixels na plotagem gera uma nuvem de pontos densa conforme exibido na Figura 4.25. Isso fez com que uma amostragem de pontos fosse necessária. Amostramos aleatoriamente 5% dos pixels de cada classe[13] (Figura 4.26). Note também na Figura 4.26 como a transformação da classe 0 melhora a visualização das demais classes. Nas demais imagens exibiremos esse tipo de gráfico somente para os métodos cuja dimensionalidade intrínseca seja 2 haja vista que uma representação em dimensão menor do que a intrínseca provavelmente não estaria mapeando o espalhamento dos pontos da forma mais fiel possível.

Figura 4.25: *Indian Pines* - gráfico de dispersão sem amostragem de pontos (esquerda PCA, direita ISOMAP)

[13] essa porcentagem foi escolhida por ser uma quantia razoável que não prejudicaria a visualização tendo em vista a quantidade original de pixels



(a) Sem transformação da classe 0

(b) Com transformação da classe 0

Figura 4.26: *Indian Pines* - gráficos de dispersão do GT (PCA esquerda, Isomap centro, LLE direita)

Em relação a detecção da dimensionalidade intrínseca, com a remoção de bandas citada na subseção 4.1.1, a dimensão original do conjunto se tornou igual a 200. De acordo com a estratégia mencionada na seção anterior, usando 400 vizinhos obteve-se os seguintes valores de dimensionalidade por método: PCA 5, ISOMAP 61, LLE 47. Utilizando 165 vizinhos obtivemos: PCA 5, ISOMAP 55, LLE 36.

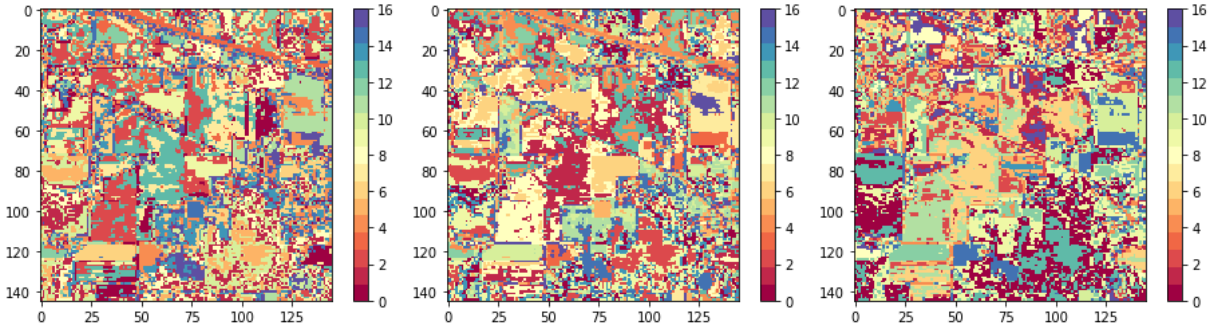
O resultado de um agrupamento também pode ser exibido por um mapa de rótulos para comparação com o GT. Para se ter uma ideia visual da influência do pareamento de rótulos, da transformação de *background* e da dimensionalidade-alvo, observe na Figura 4.27 resultados dos dois algoritmos de agrupamento para os três métodos de DR na dimensão 2 (i.e não-intrínseca). Note a dificuldade de inspeção devida ao baixo desempenho causado pela ausência de pareamento, ausência de transformação de *background* e uso de uma dimensionalidade não-intrínseca. Todos esses fatores implicam em queda de desempenho na avaliação do agrupamento. Entretanto observe na Figura 4.28 os resultados com uso da dimensionalidade estimada, transformação e pareamento. Note a melhora de discriminação, interpretação e desempenho no agrupamento.

4.3.2 *Salinas*

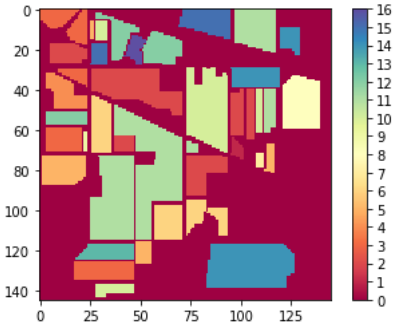
A quantidade de pixels dessa imagem não permitiu seu processamento integral por nosso hardware. Portanto, 10 reduções com amostragem aleatória de 20% dos exemplares de cada classe foram realizadas. Para se encontrar a dimensão-alvo executamos o mesmo procedimento de estimação para todas as 10 amostragens. Com isso, determinamos a dimensionalidade levando em consideração o que a maioria das estimativas indicou. Com a dimensão original igual a 204 e o número de vizinhos igual a 408 obtivemos: no caso do PCA que as dez apontaram para dimensão 2; para o ISOMAP todas para o valor 4; para o LLE obteve-se o valor 46. Com 40 vizinhos obtivemos: PCA 2, ISOMAP 10, LLE 47.

Pelo fato do PCA ser computacionalmente mais leve em comparação aos métodos não-lineares, conseguimos executá-lo para a imagem toda na dimensão 2 e visualizar o gráfico de dispersão do GT com e sem transformação de *background* (Figura 4.29). Note na Figura 4.30 como o uso do *Bagging* não prejudica a geometria da dispersão.

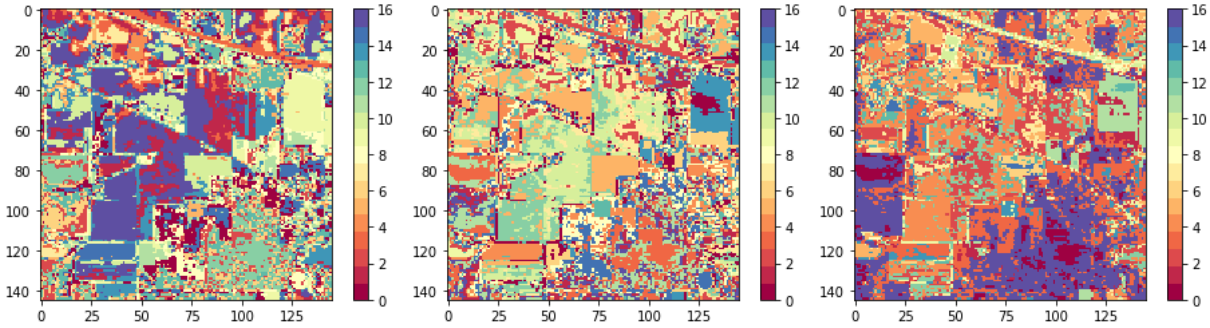
Para cada amostragem foram executados dez agrupamentos de cada algoritmo por método de DR. Como no caso dessa imagem não é possível visualizar mapas de rótulos dos agrupamentos (devido à aplicação do *Bagging*), vale a pena verificar os resultados de maneira visual usando os gráficos de dispersão para o PCA (que apontou para uma dimensão igual a 2). Esses resultados se encontram na Figura 4.31.



(a) *K-Means*

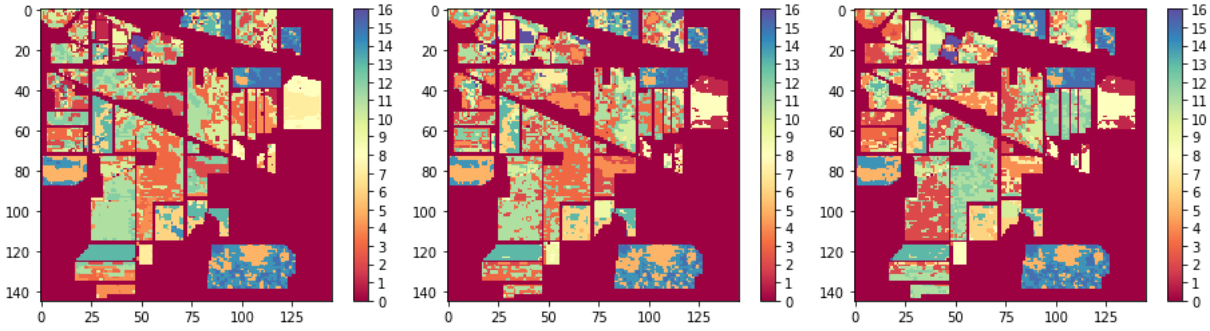


(b) GT

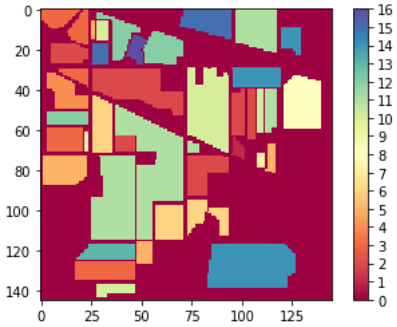


(c) EM

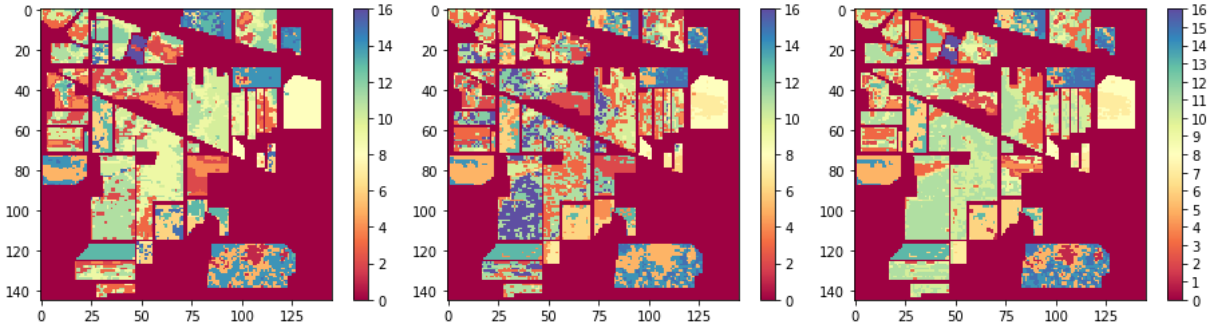
Figura 4.27: *Indian Pines* - mapas de rótulos de agrupamentos em dimensionalidade-alvo igual a 2, sem transformação de *background* e sem pareamento de rótulos com GT (PCA esquerda, ISOMAP centro, LLE direita)



(a) *K-Means*



(b) GT



(c) EM

Figura 4.28: *Indian Pines* - mapas de rótulos de agrupamentos em dimensionalidade estimada, com transformação de *background* e pareamento de rótulos (PCA esquerda, ISOMAP centro, LLE direita)

	NDR	PCA	ISOMAP	LLE
Entropy	-0,881	-0,861	-0,878	-0,874
Jaccard	0,400	0,400	0,412	0,408
Purity	0,573	0,586	0,571	0,584
Rand	0,769	0,771	0,774	0,771
Kappa	0,917	0,916	0,917	0,919
Média	0,355	0,362	0,359	0,361

Tabela 4.8: *Indian Pines* - valores das medidas de validação por método de DR

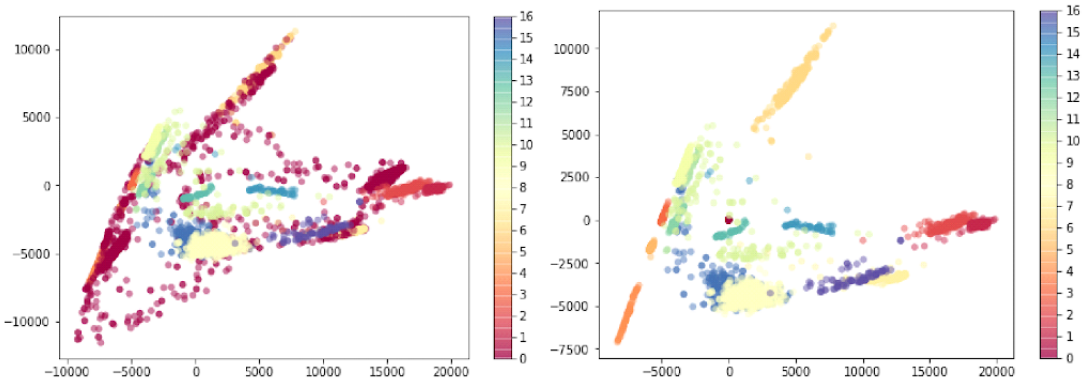


Figura 4.29: ***Salinas*** **- dispersão PCA GT da imagem toda com e sem transformação de** ***background***

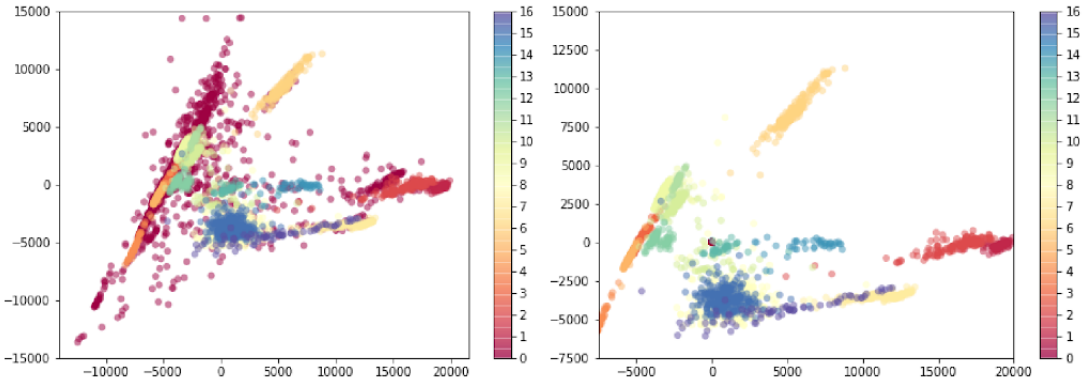


Figura 4.30: ***Salinas*** **- dispersão PCA GT do** ***Bagging*** **com e sem transformação de** ***background***

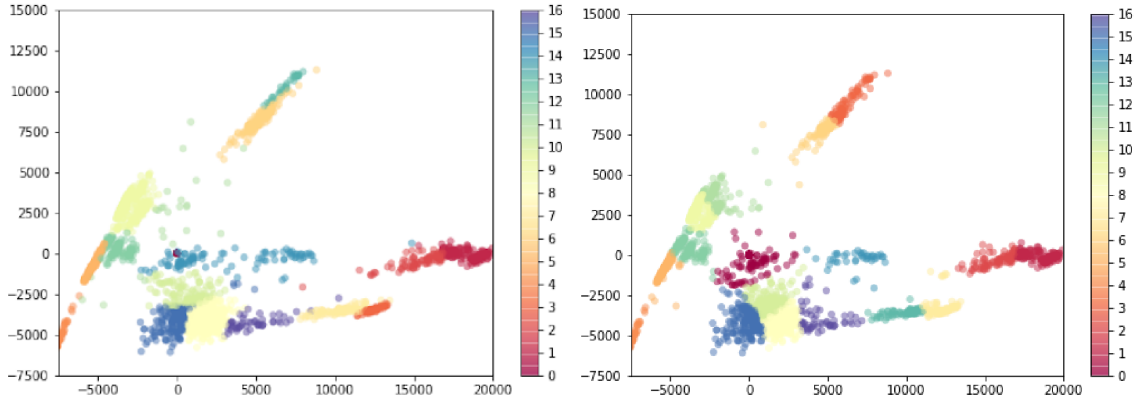


Figura 4.31: ***Salinas*** **- gráficos de dispersão PCA agrupamentos com transformação de** ***background*** **e pareamento. EM (esquerda),** ***K-Means*** **(direita)**

	NDR	PCA	ISOMAP	LLE
Entropy	-0,430	-0,303	-0,408	-0,285
Jaccard	0,463	0,460	0,467	0,459
Kappa	0,800	0,851	0,787	0,850
Purity	0,872	0,899	0,875	0,909
Rand	0,953	0,960	0,954	0,959
Média	0,531	0,573	0,534	0,578

Tabela 4.9: *Salinas* - valores das medidas de validação por método de DR

4.3.3 *SalinasA*

Esse tamanho não exigiu utilização de *Bagging*. A estimação de dimensionalidade com 204 dimensões e 408 vizinhos resultou nos valores 2 para PCA, 2 para ISOMAP e 51 para LLE. Com 40 vizinhos: PCA 2, ISOMAP 4, LLE 42. Vinte execuções de cada algoritmo de agrupamento foram realizadas para cada método de DR.

Como a dimensão utilizada para o PCA foi igual a 2, faz sentido visualizarmos o GT e o resultado do agrupamento por um gráfico de dispersão (Figura 4.32) com classe 0 transformada e pareamento de rótulos. Os pontos plotados do resultado de agrupamento foram os mesmos escolhidos para a plotagem do GT. O mapa de rótulos do PCA é exibido com e sem pareamento (Figura 4.33). Para o ISOMAP (Figura 4.34) e LLE (Figura 4.35) exibimos somente com pareamento.

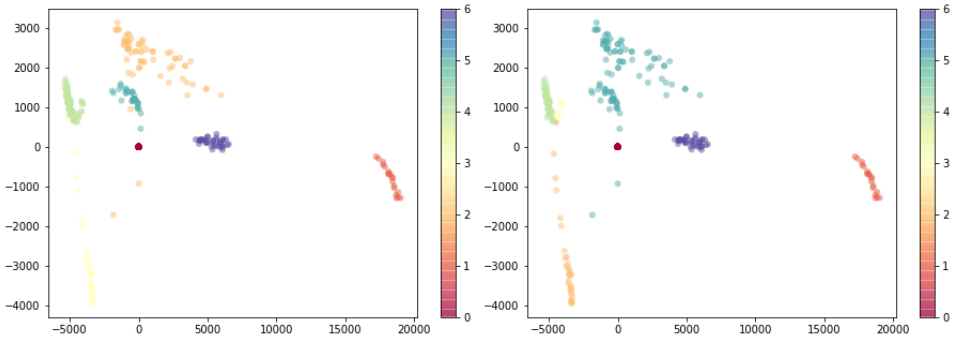
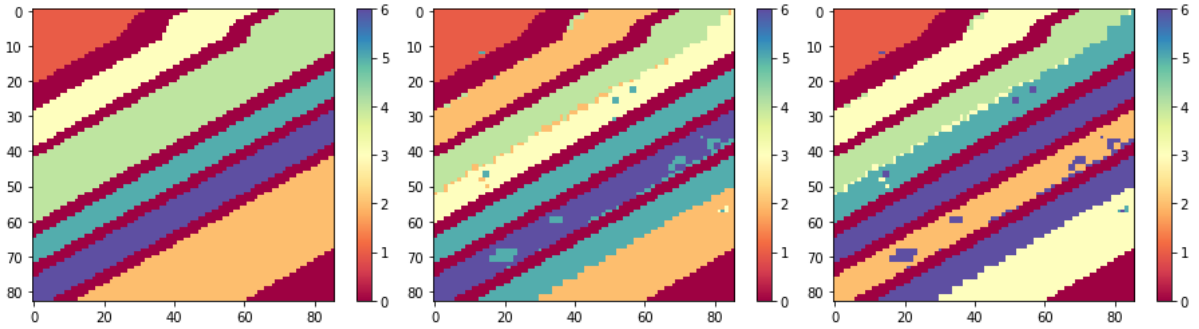
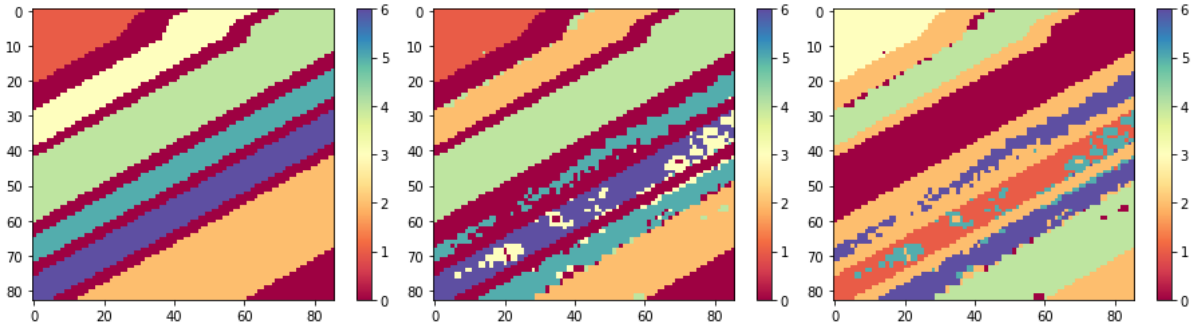


Figura 4.32: *SalinasA* - gráficos de dispersão para o PCA do GT (esquerda) e do EM



(a) EM



(b) *K-Means*

Figura 4.33: *SalinasA* - PCA mapa de rótulos GT(esquerda), com pareamento (centro), sem pareamento (direita)

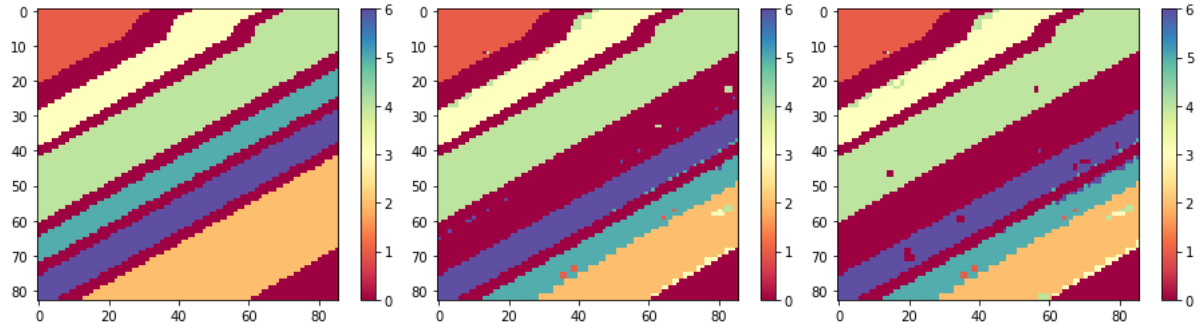


Figura 4.34: *SalinasA* - mapa de rótulos GT (esquerda) ISOMAP EM (centro) e ISOMAP *K-Means* (direita)

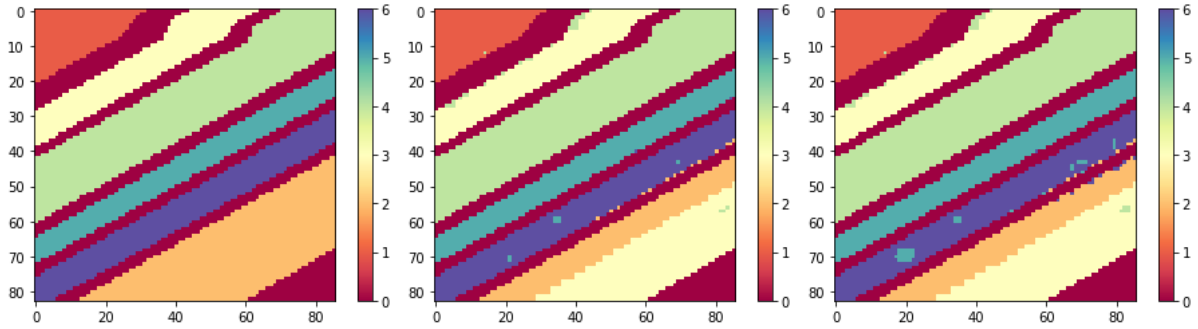


Figura 4.35: *SalinasA* - mapa de rótulos GT (esquerda), LLE EM (esquerda) e LLE *K-Means* (direita)

	NDR	PCA	ISOMAP	LLE
Entropy	-0,358	-0,228	-0,266	-0,246
Jaccard	0,302	0,559	0,458	0,347
Purity	0,882	0,910	0,906	0,931
Kappa	0,857	0,869	0,859	0,843
Rand	0,836	0,870	0,861	0,852
Média	0,503	0,596	0,563	0,545

Tabela 4.10: *SalinasA* - valores das medidas de validação por método de DR

4.3.4 *Pavia Centre*

Essa imagem exigiu o uso de *Bagging* e portanto não foi possível visualizar os mapas de rótulos dos agrupamentos, dessa forma exiberemos somente o mapa de rótulos do GT. O *Bagging* foi implantado por 100 amostragens de 2% dos pixels de cada classe. As dimensionalidades com 20 vizinhos resultaram nos valores 2 para o PCA, 20 para o ISOMAP e 22 para o LLE. Foram executados cinco agrupamentos para cada amostragem *Bagging* (e subsequente método distinto de DR). Para o PCA podemos comparar o gráfico de dispersão (Figura 4.36) de todos os pixels com e sem transformação de *background* (a) e (b), e com *Bagging* (c). Podemos também analisar os resultados dos agrupamentos EM (d) e *K-Means* (e) dada uma execução de exemplo de cada método de agrupamento.

	NDR	**PCA**	**ISOMAP**	**LLE**
Entropy	-0,090	-0,084	-0,080	-0,066
Jaccard	0,747	0,882	0,889	0,897
Kappa	0,864	0,889	0,894	0,932
Purity	0,978	0,978	0,980	0,986
Rand	0,994	0,993	0,993	0,994
Média	0,698	0,731	0,735	0,748

Tabela 4.11: *Pavia Centre* - valores das medidas de validação por método de DR

4.3.5 *Pavia University*

O *Bagging* consistiu em vinte amostragens aleatórias de 10% dos pixels de cada classe. As dimensões estimadas com 20 vizinhos são 3 para o PCA, 17 para o ISOMAP e 18 para o LLE. Para cada amostragem executou-se 10 vezes cada algoritmo de agrupamento.

	NDR	**PCA**	**ISOMAP**	**LLE**
Entropy	-0,210	-0,182	-0,196	-0,284
Jaccard	0,783	0,797	0,792	0,753
Kappa	0,758	0,826	0,814	0,771
Purity	0,939	0,949	0,944	0,936
Rand	0,980	0,984	0,982	0,951
Média	0,650	0,674	0,667	0,625

Tabela 4.12: *PaviaU* - valores das medidas de validação por método de DR

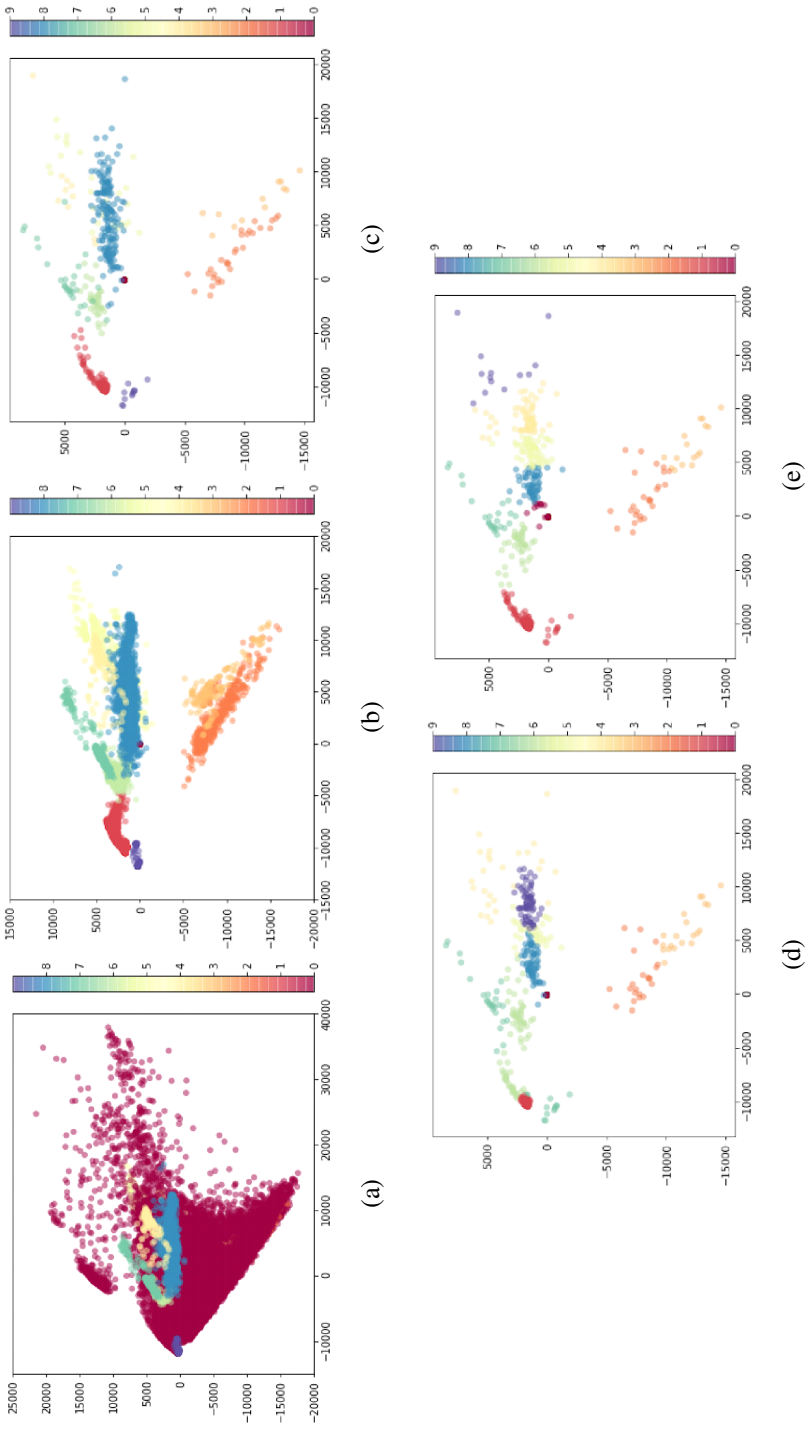


Figura 4.36: *Pavia Centre* - gráficos de dispersão PCA

4.3.6 *Kennedy Space Center*

Note como de fato as classes de interesse possuem poucas amostras em relação à classe 0. Isso nos fez adotar uma estratégia de amostragem ligeiramente diferente para essa imagem. Selecionamos todas as amostras pertencentes às classes diferentes de 0 e amostramos aleatoriamente 5% dos pixels *background,* para assim, atingir no total a quantidade máxima de pixels que conseguimos processar. Portanto essa imagem não exigiu múltiplas amostragens nem múltiplas execuções de DR. Utilizando essa abordagem, as dimensões deduzidas com 20 vizinhos foram 47 para o PCA, 110 para o ISOMAP e 65 para o LLE. Como não foram feitas múltiplas amostragens, a construção dos agrupamentos seguiu a estratégia original de 20 execuções para cada método.

	NDR	**PCA**	**ISOMAP**	**LLE**
Entropy	-0,683	-0,705	-0,707	-0,289
Kappa	0,602	0,577	0,602	0,792
Jaccard	0,655	0,656	0,666	0,698
Purity	0,830	0,819	0,829	0,929
Rand	0,946	0,941	0,944	0,989
Média	0,469	0,457	0,466	0,623

Tabela 4.13: *KSC* - valores das medidas de validação por método de DR

4.3.7 *Botswana*

Do ponto de vista da amostragem de exemplares, essa imagem possui um comportamento muito parecido com *Kennedy Space Center*, portanto as mesmas estratégias de seleção e agrupamento foram adotadas. Quanto à dimensionalidade, com 20 vizinhos, o PCA aponta para o valor 2, o ISOMAP para 30 e o LLE para 30. Mesmo o PCA apontando para dimensionalidade 2 (o que permitiria a extração de gráficos de dispersão) note que a distribuição original das classes (Figura 4.23) derivaria uma dispersão de difícil visualização devido a escassez de amostras por classe. Portanto, tais gráficos não foram gerados.

	NDR	PCA	ISOMAP	LLE
Entropy	-0,177	-0,194	-0,144	-0,159
Jaccard	0,802	0,820	0,836	0,803
Kappa	0,802	0,797	0,861	0,804
Purity	0,952	0,948	0,962	0,955
Rand	0,995	0,995	0,996	0,994
Média	0,674	0,673	0,702	0,679

Tabela 4.14: *Botswana* - valores das medidas de validação por método de DR

4.4 Resultados - testes de hipóteses

Na Tabela 4.9 apresentamos o *p-value* de *Friedman* por imagem. O teste de *Friedman* nos diz apenas se algum par de métodos diferiu. Usamos *p-value* > 0,05 para indicar equivalencia de métodos. Um teste *post-hoc* é necessário para descobrir quais pares apresentam diferença. Mostramos na Tabela 4.8 os *Nemenyi p-value* menores do que 0,05 e o correspondente par de ocorrência por imagem.

Usando os dados da Tabela 4.8 e a linha "Média" das Tabelas 4.10 - 4.14, podemos ver para *Botswana* que, para um nível de significância de 5%, há evidência de que o ISOMAP produziu melhores resultados de agrupamento do que os outros métodos, incluindo a ausência de DR. Isso implica que a DR foi benéfica para essa imagem. Também há evidência de que o ISOMAP gerou melhores atributos em comparação ao PCA. Isso justifica a melhor adequação de um método não-linear para essa imagem. Seguindo o mesmo raciocínio, em *KSC* podemos dizer que o LLE é superior aos outros métodos. Em *Pavia Centre*, LLE é superior à ausência de DR e ao PCA. Em *PaviaU* e *SalinasA*, o PCA foi significativamente distinto comparado à ausência de DR.

	p-value
Botswana	0,004
KSC	0,001
PaviaU	0,0000001
Pavia	0,003
Salinas	0,175
SalinasA	0,002
Indian Pines	0,322

Tabela 4.15: *Friedman p-value* por imagem

	Botswana	**KSC**	**PaviaU**	**Pavia**	**SalinasA**
NDR x PCA			0,007		0,001
NDR x ISOMAP	0,010				
NDR x LLE				0,004	
PCA x ISOMAP	0,004				
PCA x LLE		0,001	0,0006	0,010	
ISOMAP x LLE		0,019	0,027		

Tabela 4.16: ***Nemenyi p-values***

4.5 Conclusão

Os resultados dos testes de hipótese mostraram que: para cinco imagens dentre as sete consideradas, a DR trouxe benefício para a tarefa de agrupamento; entre esses cinco casos, em três a DR não-linear teve um melhor desempenho em comparação à linear. Com o uso de testes de hipótese não-paramétricos de comparação múltipla em diversas imagens reais e volumosas, foi encontrada diferença estatística significativa entre a aplicação e a não-aplicação de DR.

Foi exibida evidência empírica de que a extração de características aumenta significativamente o desempenho de uma tarefa de agrupamento no contexto de HIs. Apresentamos essa evidência usando DR não-supervisionada e aprendizado não-supervisionado. O GT foi usado somente para se calcular as medidas externas na etapa de validação de agrupamento.

Os resultados mostraram que, quando da presença de diferença significativa na aplicação de DR, a aplicação de métodos não-lineares de extração de características em HIs implicou na maioria dos casos em melhores resultados de agrupamento por GMM em comparação a métodos lineares. Nos demais casos, os desempenho de ambas as abordagens de DR foram equivalentes. Justificativa teórica para essa contribuição foi dada pela apresentação das estratégias e algoritmos do PCA, ISOMAP e LLE e pela discussão dos mesmos na seção 3.3, onde se vislumbrou a possibilidade da capacidade dos métodos não-lineares extraírem dados de forma mais fiel à separabilidade de classes quando da ocorrência de maldição da dimensionalidade.

A geometria de dispersão no espaço dos vetores representantes dos pixels de HIs está relacionada aos valores de refletância dos mesmos em cada imagem e, no nosso estudo, isso diz respeito somente a vizinhança espectral (e não espacial). Dado que não é possível de forma explícita relacionar teoricamente essas distintas e inúmeras possibilidades de geometrias com a hipótese de eficiência de um método não-linear em relação a um linear (ou mesmo até à eficiência da aplicação de extração de características ou sua ausência), então é necessário um estudo

empírico experimental para se verificar tais hipóteses. A dispersão de classes (GT) em cada imagem também é um fator com forte influência nos resultados do estudo apresentado e cada imagem está associada à uma dispersão particular de definida por especialistas.

Trabalhos futuros podem incluir: outras técnicas de extração de características, especialmente algoritmos mais modernos de redução não-linear; estratégias mais sofisticadas de descoberta de dimensionalidade intrínseca; outras HIs; mais métodos de agrupamento; uso de índices internos de validação de agrupamento para se testar um procedimento completamente não-supervisionado; uso de um conjunto de classificadores supervisionados para se obter uma perspectiva supervisionada e possivelmente mais evidências da maldição da dimensionalidade (poucos pixels deveriam ser usados para treinamento tendo em vista que na prática a rotulação dos especialistas é custosa). Acreditamos que tais expansões possam resultar em um estudo mais robusto em DR para HIs, compilando assim os principais pontos de interesse nessa área e servindo como guia para um primeiro contato.

REFERÊNCIAS BIBLIOGRÁFICAS

AGGARWAL, C. C.; HINNEBURG, A.; KEIM, D. A. On the surprising behavior of distance metrics in high dimensional space. In: *Lecture Notes in Computer Science*. [S.l.]: Springer, 2001. p. 420–434.

ALVARENGA, B. *Espectro Eletromagnético*. 2019. `https://www.todamateria.com.br/espectro-eletromagnetico.htm`. Accessed: 2019-11-14.

BALASUBRAMANIAN, M.; SCHWARTZ, E. L. The isomap algorithm and topological stability. *Science*, American Association for the Advancement of Science, v. 295, n. 5552, p. 7–7, 2002. ISSN 0036-8075. Disponível em: <https://science.sciencemag.org/content/295/5552/7>.

BAUMGARDNER, M. F.; BIEHL, L. L.; LANDGREBE, D. A. *220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3*. Sep 2015. Disponível em: <https://purr.purdue.edu/publications/1947/1>.

BELLMAN, R.; COLLECTION, K. M. R. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961. (Princeton Legacy Library). Disponível em: <https://books.google.com.br/books?id=POAmAAAAMAAJ>.

BENEDIKTSSON, J.; GHAMISI, P. *Spectral-Spatial Classification of Hyperspectral Remote Sensing Images*. Artech House, 2015. ISBN 9781608078134. Disponível em: <https://books.google.com.br/books?id=TtnRCgAAQBAJ>.

BERNSTEIN, M.; SILVA, V. D.; LANGFORD, J. C.; TENENBAUM, J. B. *Graph Approximations to Geodesics on Embedded Manifolds*. 2000.

BEYER, K. S.; GOLDSTEIN, J.; RAMAKRISHNAN, R.; SHAFT, U. When is "nearest neighbor" meaningful? In: *Proceedings of the 7th International Conference on Database Theory*. London, UK, UK: Springer-Verlag, 1999. (ICDT '99), p. 217–235. ISBN 3-540-65452-6. Disponível em: <http://dl.acm.org/citation.cfm?id=645503.656271>.

BIOUCAS-DIAS, J. M.; PLAZA, A.; CAMPS-VALLS, G.; SCHEUNDERS, P.; NASRABADI, N.; CHANUSSOT, J. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and Remote Sensing Magazine*, v. 1, n. 2, p. 6–36, June 2013. ISSN 2373-7468.

BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

BORG, I.; GROENEN, P. *Modern Multidimensional Scaling: theory and applications*. 2. ed. [S.l.]: Springer-Verlag, 2005.

BORODIN, A.; OSTROVSKY, R.; RABANI, Y. Lower bounds for high dimensional nearest neighbor search and related problems. In: *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, May 1-4, 1999, Atlanta, Georgia, USA*. [s.n.], 1999. p. 312–321. Disponível em: <https://doi.org/10.1145/301250.301330>.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, Aug 1996. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1018054314350>.

CAMASTRA, F. Data dimensionality estimation methods: a survey. *Pattern Recognition*, v. 36, n. 12, p. 2945 – 2954, 2003. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320303001766>.

CAMPADELLI, P.; CASIRAGHI, E.; CERUTI, C.; ROZZA, A. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, v. 2015, p. 1–21, 10 2015.

CAMPS-VALLS, G.; TUIA, D.; GÓMEZ-CHOVA, L.; JIMÉNEZ, S.; MALO, J. Remote sensing image processing. In: *Remote Sensing Image Processing*. [S.l.: s.n.], 2011.

CARREIRA-PERPINAN, M. A. *A review of dimension reduction techniques*. 1997.

CERUTI, C. *NOVEL TECHNIQUES FOR INTRINSIC DIMENSION ESTIMATION*. Tese (Doutorado) — Scuola di Dottorato in Matematica e Statistica per le Scienze Computazionali - XXVII ciclo Dipartimento di Matematica Federigo Enriques, 2014. Disponível em: <https://pdfs.semanticscholar.org/49b5/04787231cb2b7708ed7f2e6f4e3094a9f020.pdf>.

CHANG, C.-I. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. [S.l.]: Plenum Publishing Co., 2003. ISBN 0306474832.

CHANG, C.-I.; DU, Q. Estimation of number of spectrally distinct signal sources in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, v. 42, n. 3, p. 608–619, March 2004. ISSN 0196-2892.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37–46, 1960. Disponível em: <https://doi.org/10.1177/001316446002000104>.

CONGALTON, R. G. A review of assessing the accuracy of classifications of remotely sensed data. *REMOTE SENS. ENVIRON.*, 1991.

COX, T. F.; COX, M. A. A. *Multidimensional Scaling*. [S.l.]: Chapman & Hall, 2001. 295 p p. (Monographs on Statistics and Applied Probability, v. 88).

CUNNINGHAM, J. P.; GHAHRAMANI, Z. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, v. 16, p. 2859–2900, 2015.

DEMARTINES, P. *ANALYSE DE DONNEES PAR RESEAUX DE NEURONES AUTO-ORGANISES*. [s.n.], 1994. Disponível em: <https://books.google.com.br/books?id=_ObEMgEACAAJ>.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, v. 39, n. 1, p. 1–38, 1977.

DIACONIS, P.; FREEDMAN, D. Asymptotics of graphical projection pursuit. *Ann. Statist.*, The Institute of Mathematical Statistics, v. 12, n. 3, p. 793–815, 09 1984. Disponível em: <https://doi.org/10.1214/aos/1176346703>.

DING, L.; TANG, P.; LI, H. Subspace feature analysis of local manifold learning for hyperspectral remote sensing images classification. *Applied Mathematics & Information Sciences*, Citeseer, v. 8, n. 4, p. 1987, 2014.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. 2. ed. New York: Wiley, 2001. ISBN 978-0-471-05669-0.

FAN, M.; GU, N.; QIAO, H.; ZHANG, B. Intrinsic dimension estimation of data by principal component analysis. *CoRR*, abs/1002.2050, 2010. Disponível em: <http://arxiv.org/abs/1002.2050>.

FORGY, E. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, v. 21, p. 768–780, 1965.

FRANÇOIS, D. *High-dimensional data analysis: optimal metrics and feature selection*. Tese (Doutorado) — UCL - FSA/INMA - Département d'ingénierie mathématique, 2007. Disponível em: <http://hdl.handle.net/2078.1/5170>.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, Informa UK Limited, v. 32, n. 200, p. 675–701, dec 1937. Disponível em: <https://doi.org/10.1080%2F01621459.1937.10503522>.

FUKUNAGA, K. *Introduction to Statistical Pattern Recognition (2Nd Ed.)*. San Diego, CA, USA: Academic Press Professional, Inc., 1990. ISBN 0-12-269851-7.

GHAMISI, P.; YOKOYA, N.; LI, J.; LIAO, W.; LIU, S.; PLAZA, J.; RASTI, B.; PLAZA, A. Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, v. 5, n. 4, p. 37–78, Dec 2017. ISSN 2168-6831.

GOETZ, A. F.; VANE, G.; SOLOMON, J. E.; ROCK, B. N. Imaging spectrometry for earth remote sensing. *Science*, American Association for the Advancement of Science, v. 228, n. 4704, p. 1147–1153, 1985. ISSN 0036-8075. Disponível em: <http://science.sciencemag.org/content/228/4704/1147>.

GOLDBERG, Y.; RITOV, Y. Ldr-lle: Lle with low-dimensional neighborhood representation. In: BEBIS, G.; BOYLE, R.; PARVIN, B.; KORACIN, D.; REMAGNINO, P.; PORIKLI, F.; PETERS, J.; KLOSOWSKI, J.; ARNS, L.; CHUN, Y. K.; RHYNE, T.-M.; MONROE, L. (Ed.). *Advances in Visual Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 43–54. ISBN 978-3-540-89646-3.

GRANA, M.; VEGANZONS, M.; AYERDI, B. *Hyperspectral Remote Sensing Scenes - Grupo de Inteligencia Computacional (GIC)*. feb 2020. Disponível em: <http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes>.

GREEN, R. O.; EASTWOOD, M. L.; SARTURE, C. M.; CHRIEN, T. G.; ARONSSON, M.; CHIPPENDALE, B. J.; FAUST, J. A.; PAVRI, B. E.; CHOVIT, C. J.; SOLIS, M.; OLAH, M. R.; WILLIAMS, O. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (aviris). *Remote Sensing of Environment*, v. 65, n. 3, p. 227 – 248, 1998. ISSN 0034-4257. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0034425798000649>.

HALL, P.; LI, K.-C. On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, The Institute of Mathematical Statistics, v. 21, n. 2, p. 867–889, 06 1993. Disponível em: <https://doi.org/10.1214/aos/1176349155>.

HE, J.; DING, L.; JIANG, L.; LI, Z.; HU, Q. Intrinsic dimensionality estimation based on manifold assumption. *Journal of Visual Communication and Image Representation*, v. 25, n. 5, p. 740 – 747, 2014. ISSN 1047-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1047320314000078>.

HUGHES, G. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, v. 14, n. 1, p. 55–63, January 1968. ISSN 0018-9448.

HWANG, J.; LAY, S.; LIPPMAN, A. F. Nonparametric multivariate density estimation: a comparative study. *IEEE Trans. Signal Processing*, v. 42, n. 10, p. 2795–2810, 1994. Disponível em: <https://doi.org/10.1109/78.324744>.

IVEZIĆ, Ž.; CONNOLLY, A.; VANDERPLAS, J.; GRAY, A. *Statistics, Data Mining and Machine Learning in Astronomy*. [S.l.]: Princeton University Press, 2014.

JACCARD, P. The distribution of the flora in the alpine zone.1. *New Phytologist*, v. 11, n. 2, p. 37–50, 1912. Disponível em: <https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x>.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. Disponível em: <http://portal.acm.org/citation.cfm?id=46712>.

JAIN, A. K.; DUIN, R. P. W. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1, p. 4–37, Jan 2000. ISSN 0162-8828.

JIA, X.; KUO, B.; CRAWFORD, M. M. Feature mining for hyperspectral image classification. *Proceedings of the IEEE*, v. 101, n. 3, p. 676–697, March 2013. ISSN 0018-9219.

JIMENEZ, L. O.; LANDGREBE, D. A. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, v. 28, n. 1, p. 39–54, Feb 1998. ISSN 1094-6977.

JOLLIFFE, I. T. *Principal Component Analysis*. 2. ed. [S.l.]: Springer, 2002. 487 pages p.

KENDALL, M. *A Course in the Geometry of N Dimensions*. Dover Publications, 2004. (Dover books on mathematics). ISBN 9780486439273. Disponível em: <https://books.google.com.br/books?id=_dFJ6pSzRLkC>.

KESHAVA, N.; KEREKES, J. P.; MANOLAKIS, D.; SHAW, G. A. An algorithm taxonomy for hyperspectral unmixing. In: . [S.l.: s.n.], 2000.

KUHN, H. W. Variants of the hungarian method for assignment problems. *Naval Research Logistics Quarterly*, v. 3, n. 4, p. 253–258, December 1956. Disponível em: <https://ideas.repec.org/a/wly/navlog/v3y1956i4p253-258.html>.

KUHN, H. W.; YAW, B. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, p. 83–97, 1955.

LANDGREBE, D. Multispectral data analysis: a signal theory perspective. 01 1998.

LANDGREBE, D. On information extraction principles for hyperspectral data a white paper. 05 2019.

LANDGREBE, D. A. Information extraction principles and methods for multispectral and hyperspectral image data. In: . [S.l.: s.n.], 1998.

LEE, J. A.; VERLEYSEN, M. *Nonlinear Dimensionality Reduction*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2007. ISBN 0387393501, 9780387393506.

LI, S.; WU, H.; WAN, D.; ZHU, J. An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. *Knowledge-Based Systems*, v. 24, n. 1, p. 40 – 48, 2011. ISSN 0950-7051. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705110001097>.

LLOYD, S. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, v. 28, n. 2, p. 129–137, March 1982. ISSN 0018-9448.

LOWE, D. S.; POLCYN, F. C.; SHAY, J. R. Multispectral data collection program. In: *Proceedings of the Third Symposium on Remote Sensing of Environment*. [S.l.: s.n.], 1965. p. 667–780.

LUXBURG, U. von. A tutorial on spectral clustering. *Statistics and Computing*, v. 17, p. 395–416, 2007.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.: s.n.], 1967. p. 281–297.

MARINI, F.; AMIGO, J. M. Chapter 2.4 - unsupervised exploration of hyperspectral and multispectral images. In: AMIGO, J. M. (Ed.). *Hyperspectral Imaging*. Netherlands: Elsevier, 2020, (Data Handling in Science and Technology, v. 32). p. 93 – 114. ISBN 978-0-444-63977-6. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780444639776000067>.

MINKA, T. P. Automatic choice of dimensionality for pca. In: LEEN, T. K.; DIETTERICH, T. G.; TRESP, V. (Ed.). *NIPS*. [S.l.]: MIT Press, 2000. p. 598–604.

MUNKRES, J. *ALGORITHMS FOR THE ASSIGNMENT AND TRANSPORTATION PROBLEMS*. 1957.

MURPHY, J. M.; MAGGIONI, M. Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 57, n. 3, p. 1829–1845, 2018.

MURPHY, K. P. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN 9780262018029 0262018020.

MYASNIKOV, E. Evaluation of nonlinear dimensionality reduction techniques for classification of hyperspectral images. In: . [S.l.: s.n.], 2018.

NEMENYI, P. *Distribution-free Multiple Comparisons*. Princeton University, 1963. Disponível em: <https://books.google.com.br/books?id=nhDMtgAACAAJ>.

PAL, M.; FOODY, G. M. Feature selection for classification of hyperspectral data by svm. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 48, n. 5, p. 2297–2307, 2010.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PLAZA, A.; BENEDIKTSSON, J. A.; BOARDMAN, J. W.; BRAZILE, J.; BRUZZONE, L.; CAMPS-VALLS, G.; CHANUSSOT, J.; FAUVEL, M.; GAMBA, P.; GUALTIERI, A.; MARCONCINI, M.; TILTON, J. C.; TRIANNI, G. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, v. 113, p. S110 – S122, 2009. ISSN 0034-4257. Imaging Spectroscopy Special Issue. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0034425709000807>.

PLAZA, A.; MARTINEZ, P.; PLAZA, J.; PEREZ, R. Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. *IEEE Transactions on Geoscience and Remote Sensing*, v. 43, n. 3, p. 466–479, March 2005. ISSN 1558-0644.

POLITO, M.; PERONA, P. Grouping and dimensionality reduction by locally linear embedding. In: DIETTERICH, T. G.; BECKER, S.; GHAHRAMANI, Z. (Ed.). *Advances in Neural Information Processing Systems 14*. MIT Press, 2002. p. 1255–1262. Disponível em: <http://papers.nips.cc/paper/2033-grouping-and-dimensionality-reduction-by-locally-linear-embedding.pdf>.

RAMAKRISHNA, B.; PLAZA, A.; CHANG, C.-I.; REN, H.; DU, Q.; CHANG, C.-C. Spectral/spatial hyperspectral image compression. In: ____. [S.l.: s.n.], 2006. p. 309–346.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, Taylor & Francis, v. 66, n. 336, p. 846–850, 1971. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356>.

RIDDER, D. de; DUIN, R. P. *Locally Linear Embedding for Classification*. [S.l.], 2002.

ROWEIS, S.; SAUL, L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, v. 290, p. 2323–2326, 2000.

SAQUI, D. *Metodologia supervisionada para seleção de bandas de imagens hiperespectrais utilizando o NSGA2*. Tese (Doutorado) — Universidade Federal de São Carlos, January 2018.

SAUL, L.; ROWEIS, S. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, v. 4, p. 119–155, 2003.

SAUL, L. K.; ROWEIS, S. T. *An Introduction to Locally Linear Embedding*. [S.l.], 2000.

SCOTT, D. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992. (A Wiley-interscience publication). ISBN 9780471547709. Disponível em: <https://books.google.com.br/books?id=7crCUS_F2ocC>.

SCOTT, D.; THOMPSON, J. Probability density estimation in higher dimension. *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, 01 1983.

SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 7 1948. Disponível em: <https://ieeexplore.ieee.org/document/6773024>.

SUN, K.; MARCHAND-MAILLET, S. An information geometry of statistical manifold learning. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. JMLR.org, 2014. (ICML'14), p. II–1–II–9. Disponível em: <http://dl.acm.org/citation.cfm?id=3044805.3044893>.

TENENBAUM, J. B.; SILVA, V. de; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, v. 290, p. 2319–2323, 2000.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition, Fourth Edition*. 4th. ed. Orlando, FL, USA: Academic Press, Inc., 2008. ISBN 1597492728, 9781597492720.

TRUNK, G. V. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 1, n. 03, p. 306–307, jul 1979. ISSN 0162-8828.

WANG, J.; CHANG, C.-I. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, v. 44, n. 6, p. 1586–1600, June 2006. ISSN 0196-2892.

WEBB, A. *Statistical Pattern Recognition*. 2. ed. A Hodder Arnold Publication, 1999. Disponível em: </bib/webb/webb1999spr/Statistical Pattern Recognition 2nd Ed - Andrew R. Webb.pdf>.

WEGMAN, E. J. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, [American Statistical Association, Taylor & Francis, Ltd.], v. 85, n. 411, p. 664–675, 1990. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2290001>.

WEINBERGER; KILIAN; SAUL; LAWRENCE. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, v. 70, p. 77, 10 2006.

YOUNG, T.; CALVERT, T. *Classification, estimation, and pattern recognition*. American Elsevier Pub. Co., 1974. ISBN 9780444001351. Disponível em: <https://books.google.com.br/books?id=ZWhQAAAAMAAJ>.

ZHENG, Z.; CHEN, P.; ZHU, M.; HUANG, Z.; HE, Y.; FENG, Y.; LU, Y.; YU, Z.; YU, S.; WANG, S.; LI, J. The manifold learning for dimensionality reduction with hyperspectral image. In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. [S.l.: s.n.], 2016. p. 2757–2760. ISSN 2153-7003.

Apendice A

DEMONSTRAÇÕES MATEMÁTICAS

A.1 Fenômeno da ortogonalização

Seja d a dimensão do espaço, o cosseno do ângulo entre um vetor diagonal (i.e. que possui o mesmo valor em todas as componentes) e um vetor que compõe a base do espaço (i.e. um eixo coordenado Euclidiano) é:

$$\cos(\theta_d) = \pm\frac{1}{\sqrt{d}} \tag{A.1}$$

Note que $\lim_{d\to\infty}\cos(\theta) = 0$, o que implica que, em um espaço de alta dimensão, um vetor diagonal tem a tendência de se tornar ortogonal aos eixos de coordenadas.

Figura A.1 (esquerda) ilustra como o ângulo θ_d entre um vetor diagonal e um eixo da base se aproxima de 90º com o aumento da dimensionalidade d.

Seja x_{diag} um vetor diagonal qualquer em um espaço d dimensional. Seja xc_i a sua i-ésima coordenada. Qualquer ponto no espaço pode ser representado na forma:

$$p = \sum_{i=1}^{d} \alpha_i xc_i \tag{A.2}$$

A projeção de p em x_{diag}é dada por p_{diag}:

$$p_{diag} = \left(p^T x_{diag}\right) x_{diag} = \sum_{i=1}^{d} \alpha_i \left(xc_i^T x_d\right) x_d \tag{A.3}$$

Mas, conforme d aumenta $xc_i^T x_{diag} \approx 0$ o que implica $p_{diag} \approx 0$

Como consequencia p_{diag} é projetada para a origem, perdendo a informação de sua localização no espaço d dimensional. Então, a projeção de qualquer grupo de vetores em um vetor diagonal (i.e. que pode ser obtido como a média das características), poderia destruir a informação contida nos dados.

A.2 Fenômeno da concentração

Dada uma esfera E de raio ε_1 e uma esfera E_i inscrita em E de raio $\varepsilon_1 - \varepsilon_2$ dizemos que a casca de E é a estrutura definida pela diferença $E - E_i$.

Em um espaço d dimensional, a fração do volume s da casca é dada por r:

$$r = \frac{s(\varepsilon_1) - s(\varepsilon_1 - \varepsilon_2)}{s(\varepsilon_1)} = \frac{\varepsilon_1^d - (\varepsilon_1 - \varepsilon_2)^d}{\varepsilon_1^d} = 1 - \left(1 - \frac{\varepsilon_2}{\varepsilon_1}\right)^d \tag{A.4}$$

Note que em altas dimensões, $\lim_{d\to\infty} r = 1, \forall \varepsilon_2 > 0$, implicando que a maior parte do volume da hiperesfera está concentrada em sua casca

Para ilustrar esse fenômeno, observe na Figura A.1 (direita) para o caso $\varepsilon_2 = \frac{\varepsilon_1}{5}$, como o volume da hiperesfera passa a se concentrar na casca conforme a dimensão aumenta.

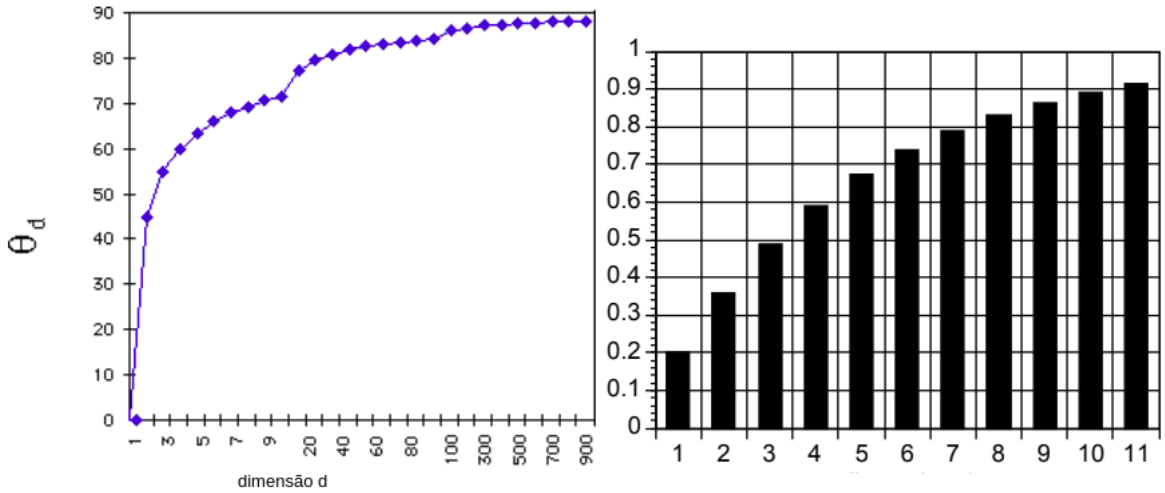


Figura A.1: Dimensionalidade vs ângulo entre vetor diagonal e eixo de coordenadas (esquerda). Volume da casca de uma hiperesfera em função da dimensionalidade para $\varepsilon = \frac{r}{5}$ (direita). Extraída de (LANDGREBE, 2019)

Apendice B

PROBLEMA DA ALOCAÇÃO ÓTIMA

Seja $P = \{p_1, ..., p_k\}$ a partição de grupos nos dados gerada pelo agrupamento e $P' = \{p'_1, ..., p'_k\}$ a partição de classes do GT. Uma instância dessas partições poderia ser apresentada pela Figura B.1. Visualmente sabemos que o agrupamento dividiu de fato o conjunto da forma esperada; porém, se tomarmos como critério o simples casamento entre rótulos, conclui-se que não houve acerto algum na estimação (note que os rótulos gerados a cada execução de agrupamento são arbitŕarios).

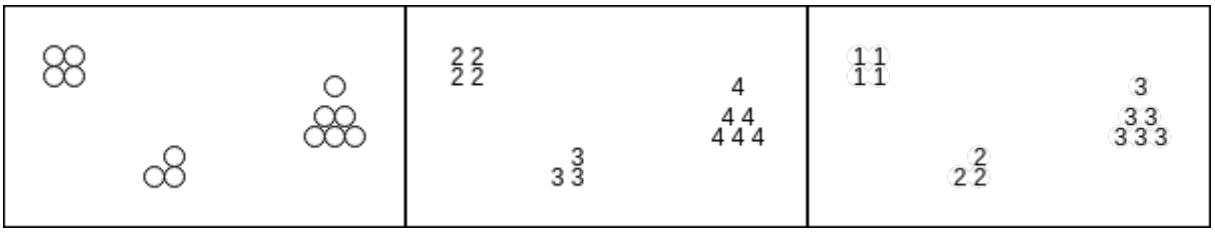


Figura B.1: Resultado do agrupamento (centro); GT (direita)

Portanto, se faz necessário emparelhar os rótulos entre as duas partições antes da aplicação da medida de avaliação *Kappa*. Por emparelhamento entende-se encontrar a correspondência entre rótulos p_i e p'_j tal que a similaridade entre os mesmos seja máxima. Por exemplo o emparelhamento a respeito da Figura B.1 seria:

- rótulo '4' do agrupamento corresponde ao '3' do GT
- rótulo '3' do agrupamento ao '2' do GT
- rótulo '2' do agrupamento ao '1' do GT

Em teoria dos grafos, o problema da alocação ótima pode ser descrito da seguinte forma: dado um grafo bipartido $G = (V, E)$ com $V = O \cup Q$ e $O \cap Q = \emptyset$ tal que $|O| = |Q| = k$ em que t_{ij} denota

o custo da aresta entre os vértices o_i e q_j, o problema consiste em encontrar o emparelhamento perfeito de custo mínimo. O problema é representado por uma matriz de custos quadrada $k \times k$ dada na Figura B.3 e o grafo é representado conforme Figura B.2.

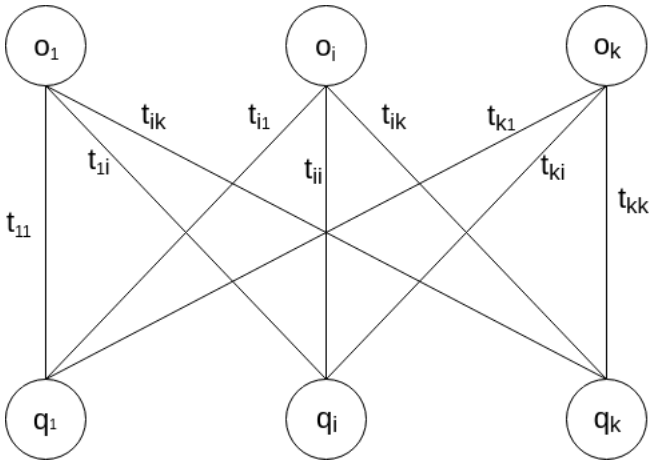


Figura B.2: Grafo genérico do problema da alocação ótima

		rótulos agrupamento		
		1	i	k
rótulos GT	1	t_{11}	t_{i1}	t_{k1}
	i	t_{1i}	t_{ii}	t_{ki}
	k	t_{1k}	t_{ik}	t_{kk}

Figura B.3: Matriz de custos do problema de alocação ótima

A solução do oroblema pode ser obtida através do algoritmo Húngaro (KUHN; YAW, 1955; KUHN, 1956; MUNKRES, 1957) que pode ser descrito informalmente pelos seguintes passos:

1. Na matriz de custos original, identifique o valor mínimo de cada linha e subtraia esse valor de todas as entradas da linha

2. Na matriz resultante do passo 1, identifique o valor mínimo de cada coluna e subtraia esse valor de todas as entradas da coluna

3. Identifique a solução ótima como a designação viável associada com as entradas iguais a zero da matriz obtida no passo 2

   Se não for possível garantir nenhuma designação viável (com todas as entradas iguais a zero), então

   (a) trace o número mínimo de linhas horizontais e verticais na última matriz reduzida que abrange todas as entradas zero

(b)selecione a menor entrada não abrangida e subtraia essa entrada de todas as entradas não abrangidas e então adicione a todas as entradas na interseção de duas linhas

(c)se não for possível encontrar nenhuma designação viável entre as entradas zero resultantes, repita o passo 3a; caso contrário, determine a designação ótima

O problema de emparelhamento de rótulos pode ser mapeado no problema de alocação ótima: cada rótulo de uma partição p_i faz o papel de um nó o_i assim como cada p'_j de um q_j. Note que $|O| = |Q| = k$ ainda, com k sendo a quantidade de rótulos (i.e. número de partições a serem definidas, parâmetro para o algoritmo de agrupamento). O custo da aresta entre um o_i e q_j passa a ser uma medida de quão dissimilar se encontra a representação de p_i em relação a p'_j. Ou seja, informalmente, quanto maior o valor desse custo, menos provável é que o rótulo de p_i esteja associando à mesma classe semântica do rótulo de p'_j. Uma forma de definir o custo t_{ij} é calculando o número de elementos na interseção das partições p_i e p'_j e subtraindo o dobro desse valor da soma do número de elementos das duas partições. A intuição por trás desse cálculo é a de que, se todos os elementos estão na interseção, o valor resultante é zero, o que significa que um rótulo corresponde ao outro. Por outro lado, quanto maior for esse valor resultante, mais distintas são as duas partições. Desse modo, em posse de todos os vértices o_i e q_j e de todos os custos t_{ij} (i.e para todo k), o algoritmo Húngaro pode ser aplicado para encontrar a correspondência de rótulos desejada.