UMA ABORDAGEM EFICIENTE PARA MÉTODOS NÃO LINEARES DE REDUÇÃO DE DIMENSIONALIDADE E UMA NOVA METODOLOGIA SUPERVISIONADA PARA REDUÇÃO DE DIMENSIONALIDADE BASEADA EM PROTÓTIPOS

Vinicius Layter Xavier

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Nelson Maculan Filho

Rio de Janeiro

Setembro de 2016

UMA ABORDAGEM EFICIENTE PARA MÉTODOS NÃO LINEARES DE REDUÇÃO DE DIMENSIONALIDADE E UMA NOVA METODOLOGIA SUPERVISIONADA PARA REDUÇÃO DE DIMENSIONALIDADE BASEADA EM PROTÓTIPOS

Vinicius Layter Xavier

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Nelson Maculan Filho, D.Sc.

_____

Prof. Jano Moreira de Souza, D.Sc.

_____

Prof. Celso da Cruz Carneiro Ribeiro, D.Sc.

_____

Prof. José Francisco Moreira Pessanha, D.Sc.

_____

Prof. Jorge Machado Damazio, D.Sc.

_____

Prof. José André Moura Brito, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2016

Xavier, Vinicius Layter

Uma Abordagem Eficiente para Métodos não Lineares de Redução de Dimensionalidade e uma Nova Metodologia Supervisionada para Redução de Dimensionalidade Baseada em Protótipos/Vinicius Layter Xavier. – Rio de Janeiro: UFRJ/COPPE, 2016.

IX, 82 p.: il.; 29,7 cm.

Orientador: Nelson Maculan Filho

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2016.

Referências Bibliográficas: p. 76-82.

1. *Sammon Mapping*. 2. *Supervised MDS*. 3. Classificação Supervisionada. 4. *Ranking* Bipartido. 5. *Local MDS* 6. *Least Absolute Residuals.* 7. Suavização Hiperbólica. I. Maculan Filho, Nelson. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Aos meus pais, Adilson e Solange.

Agradecimentos

Agradeço aos meus pais, Adilson e Solange, pois foram fundamentais na minha formação, com ensinamentos diários ao longo da minha vida. Sem o apoio deles eu não conseguiria chegar até aqui.

Agradeço a todos os professores do PESC e em especial ao meu orientador, professor Nelson Maculan Filho, pelos ensinamentos, orientação, apoio e pela relação de confiança e amizade.

Sou grato aos meus amigos de laboratório pela amizade e companheirismo nesses anos que estudamos juntos. Dentre eles, um agradecimento especial a Daniela Lubke, Renan Vicente Pinto, Caio Ribeiro de Souza, Laura De Oliveira F. Moraes e Diego Tertuliano.

Agradeço a todos os funcionários da secretaria do PESC em especial Josefina Solange Silva Santos, Gutierrez da Costa, Sônia Regina, Maria Mercedes Barreto e Claudia Helena Prata.

Agradeço a todos os integrantes da banca pelo aceite de imediato da participação na banca de defesa de tese.

E por fim, a todos aqueles que de alguma forma colaboraram com este trabalho.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

UMA ABORDAGEM EFICIENTE PARA MÉTODOS NÃO LINEARES DE REDUÇÃO DE DIMENSIONALIDADE E UMA NOVA METODOLOGIA SUPERVISIONADA PARA REDUÇÃO DE DIMENSIONALIDADE BASEADA EM PROTÓTIPOS

Vinicius Layter Xavier

Setembro/2016

Orientador: Nelson Maculan Filho

Programa: Engenharia de Sistemas e Computação

Primeiramente, nesta tese é apresentada uma abordagem eficiente para métodos não lineares de redução de dimensionalidade da classe dos métodos de escalonamento multidimensional métrico. Os métodos de Sammon, *Local MDS*, *Least Absolute Residuals* e *Supervised MDS* possuem a característica de serem não diferenciáveis. Com o emprego da suavização hiperbólica é proposta uma formulação suavizada desses métodos. Em seguida, com o método *Supervised MDS* são abordados os problemas de Classificação Supervisionada e *Ranking Bipartido*. No processo de otimização desses dois últimos problemas é proposta além da suavização das formulações o emprego de processamento paralelo. Nos problemas de redução de dimensionalidade, classificação supervisionada e *ranking* bipartido, os experimentos computacionais apresentados mostram um excelente desempenho do algoritmo proposto em comparação com outros algoritmos tradicionais. Em seguida, é proposta uma nova metodologia supervisionada para redução de dimensionalidade não-linear fundamentada em protótipos representativos de cada classe. É proposta uma generalização dos métodos de MDS, Sammon e *Least Absolute Residuals* para tratar observações novas. Na generalização dos métodos de Sammon e *Least Squares MDS* é apresentada a demonstração da propriedade de convexificação das formulações propostas. Essa metodologia tem a vantajosa característica de gerar problemas de otimização independentes e de dimensão muito baixa. Devido às características de diferenciabilidade e separabilidade, torna-se viável a resolução de problemas de grande porte.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

AN EFFICIENT APPROACH FOR NONLINEAR DIMENSIONALITY REDUCTION METHODS AND A NEW METHODOLOGY BASED ON PROTOTYPES FOR SUPERVISED DIMENSIONALITY REDUCTION

Vinicius Layter Xavier

September/2016

Advisor: Nelson Maculan Filho

Department: Systems Engineering and Computer Science

Initially, this thesis presents an efficient approach to nonlinear dimensionality reduction methods, of the class of metric multidimensional scaling methods. The methods of Sammon, Local MDS, Least Absolute Residuals and Supervised MDS have the characteristic of being non-differentiable. A smoothing formulation of these methods is proposed, using hyperbolic smoothing. Following this, the problems of Supervised Classification and Bipartite Ranking are addressed, using the Supervised MDS method. In the optimization process of the two last problems, the use of parallel processing is proposed, in addition to the smoothing of the formulations. For the problems of Dimensionality Reduction, Supervised Classification and Bipartite Ranking, the computational experiments show an excellent performance of the proposed algorithm when compared to traditional algorithms. After this, a new methodology is proposed for supervised nonlinear dimensionality reduction, based on a representative prototype for each class. A generalization of the methods of MDS, Sammon and Least Absolute Residuals is proposed, in order to deal with new observations. In the generalization of the methods of Sammon and Least Squares MDS, the property of convexification of the proposed formulations is demonstrated. This methodology has the advantage of generating low dimension independent optimization problems, since the number of observations does not influence the size of each problem. Due to the characteristics of differentiability and separability, it makes feasible the resolution of large scale problems.

.

Sumário

Capítulo 1

Introdução

A análise de conjunto de dados em um espaço de alta dimensão é uma tarefa relevante e complexa. Neste contexto os métodos de redução de dimensionalidade possuem um importante papel, pois, de um modo geral, é mais fácil extrair informações contidas em dados quando o número de variáveis é pequeno. O principal objetivo dos métodos de redução de dimensionalidade é encontrar um conjunto menor de variáveis, calculado a partir dos originais, buscando uma mínima perda de informação em relação aos dados originais.

Os métodos de redução de dimensionalidade são muitas vezes utilizados com o objetivo de visualização de dados, em especial, quando as projeções são feitas para os espaços R^2 ou R^3, casos em que é possível visualizar os dados com diagramas de dispersão. Diagramas de dispersão, de uma forma geral, são usados para a visualização de combinações de duas ou três variáveis. Entretanto, quando o número de variáveis aumenta, por mais que se consiga visualizar todas as combinações, perde-se a relação conjunta das variáveis, sendo assim torna-se difícil a interpretação dos dados como um todo. Desta forma, é natural o uso dos métodos de redução de dimensionalidade com o objetivo de visualização de dados, pois na projeção, busca-se minimizar a perda da estrutura original dos dados. Além disso, a inspeção visual do conjunto de dados de alta dimensão permite uma rápida identificação de elementos estruturais como agrupamentos, regiões homogêneas e observações atípicas (GISBRECHT, SCHULZ, HAMMER, 2015).

Tipicamente dois tipos de erro podem acontecer no processo de redução de dimensionalidade. No primeiro tipo, dados que estão muito distantes no espaço de alta dimensão podem ser erroneamente projetados para pontos próximos no espaço de baixa dimensão. De forma oposta, no segundo tipo, o erro ocorre quando observações que

estão originalmente próximas, vizinhas no espaço de alta dimensão, são representadas por pontos distantes na projeção, causando uma descontinuidade no mapeamento entre os espaços de alta dimensão e baixa dimensão (VENNA, KASKI, 2006).

O problema de redução de dimensionalidade tem inúmeras aplicações práticas em diversas áreas do conhecimento humano. SAMMON (1969) relaciona uma série dessas áreas, tais como: Estatística multivariada, Teoria da informação e Reconhecimento de padrões. A presença do uso dessas técnicas em muitas áreas é essencialmente justificada pela capacidade dos métodos de simplificar a interpretação dos dados, possibilitando entre outras vantagens, o uso de ferramentas de mineração de dados de um modo mais eficiente.

Um dos métodos mais conhecidos e tradicionais em Redução de Dimensionalidade é a Análise de Componentes Principais (PCA). Nesse método, ocorre uma mudança de base das variáveis através de uma transformação linear ortogonal que mapeia o espaço original dos atributos para um novo conjunto de coordenadas ortogonais chamadas de componentes principais. Esta transformação é definida de tal maneira que considera o critério de preservação da variância, no qual as componentes principais concentram quantidades decrescentes de variância global.

Muitos métodos novos vêm surgindo baseados em modificações ou adaptações de métodos antigos. Um exemplo é a adaptação do PCA através do uso da técnica de *kernel* para se ter um mapeamento não linear. Atualmente existe uma grande diversidade de métodos de redução de dimensionalidade e uma vasta literatura. WANG e SUN (2014) apresentam uma revisão bibliográfica sobre redução de dimensionalidade e mineração de dados. MAATEN *et al.* (2009), apresenta uma avaliação comparativa de métodos de redução de dimensionalidade. SORZANO *et al.* (2014) apresentam uma extensa revisão bibliográfica. LEE e VERLEYSEN (2007) abordam detalhadamente o problema de redução de dimensionalidade não linear. BORG e GROENEN (2005) e COX e COX (2000) abordam detalhadamente o problema de escalonamento multidimensional.

Dentre a variedade de métodos existentes na literatura, neste trabalho são abordados métodos não lineares, métodos esses que não são fundamentados em uma combinação linear das variáveis originais. Em particular, são abordados um conjunto de

métodos da classe escalonamento multidimensional. Escalonamento multidimensional (MDS) é uma classe de métodos de redução de dimensionalidade que considera o critério de preservação de distância ou dissimilaridade, em que se busca minimizar as distorções entre as distâncias ou as dissimilaridades medidas entre as observações no espaço original de alta dimensão e as distâncias medidas no espaço de baixa dimensão. Dessa forma, busca-se preservar a informação da estrutura dos dados, bem como relações de vizinhança entre as observações.

Na primeira parte do trabalho, são abordados os seguintes métodos da classe escalonamento multidimensional: Escalonamento multidimensional com métrica de mínimos quadrados e distância Minkowski, Mapeamento de Sammon, Local MDS, *Least Absolute Residuals MDS* (LAR MDS) e Supervised MDS. As formulações destes métodos possuem a significativa característica de serem não diferenciáveis. Para superar essas dificuldades decorrentes da presença de não diferenciabilidade, o método de resolução proposto adota uma estratégia de suavização usando uma função diferenciável de classe C^{∞}. A utilização desta técnica, denominada suavização hiperbólica (SH), permite superar as principais dificuldades apresentadas pelos problemas originais.

Além do problema de redução de dimensionalidade apresentado acima, na segunda parte desse trabalho a resolução dos problemas de Classificação Supervisionada e *Ranking* bipartido é abordada tendo como base o método o método Supervised MDS. O método *Supervised MDS* é um método relativamente novo e pertence a classe de métodos de redução de dimensionalidade entretanto ele tem como objetivo principal a classificação supervisionada e o *ranking* bipartido. Comparado com outros métodos tradicionais para os problemas de Classificação Supervisionada e *Ranking* bipartido o método *Supervised MDS* produz bons resultados e em alguns casos resultados até superiores, sendo assim pode ser considerado um método muito eficaz e promissor.

Com a finalidade de ilustrar o desempenho das abordagens propostas, dois tipos de experimentos computacionais foram realizados: Experimentos para uma comparação direta com algoritmos congêneres disponíveis na literatura e experimentos para uma comparação com resultados publicados na literatura. Com base no conjunto dos

resultados obtidos, pode-se verificar a confiabilidade e a eficácia da metodologia proposta.

Os métodos de redução de dimensionalidade, em sua grande maioria, processam os dados em uma única etapa e não possuem a capacidade de generalização para o mapeamento de uma observação nova. Alguns métodos possuem uma extensão especial para abordar as observações novas, enquanto outros executam o mapeamento de forma direta como é o caso de PCA (NAGARAJAN *et al.*, 2015). A extensão dessas metodologias de redução de dimensionalidade para o tratamento de observações novas, no contexto de métodos não lineares, não é direta e usualmente faz-se necessário o uso de funções auxiliares dependentes de parâmetros que devem ser ajustados.

Na terceira parte deste trabalho, é proposto o uso de protótipos representativos dos dados em articulação com uma nova abordagem para a generalização do mapeamento de um conjunto de observações novas. Os métodos de Sammon, Escalonamento Multidimensional com distância Euclidiana e *Least Absolute Residuals* são utilizados na redução de dimensionalidade dos protótipos, gerando protótipos no espaço de baixa dimensão. Em seguida, em função das coordenadas destes protótipos no espaço de baixa dimensão, todas as observações são mapeadas, sejam observações do conjunto de dados ou observações novas. Assim, deve-se destacar que o mesmo procedimento utilizado no mapeamento das observações do conjunto de dados é também utilizado nas observações novas.

No próximo capítulo, é apresentada uma revisão dos métodos de MDS. No Capítulo 3, são apresentados os problemas de classificação supervisionada e *ranking* bipartido e o método Supervised MDS. No Capítulo 4, é introduzida uma visão geral sobre a SH e a proposta da aplicação da SH em algumas formulações dos métodos apresentados nos Capítulos 2 e 3. Em seguida, no Capítulo 5, é apresentada uma revisão de algoritmos da literatura em conjunto com o algoritmo proposto. No Capítulo 6, são apresentados resultados computacionais para a metodologia proposta e uma comparação com as abordagens tradicionais não suavizadas. No Capítulo 7, é proposta uma nova metodologia não linear baseada no uso de protótipos representativos dos dados em conjunto com a generalização para observações novas. Por último, o Capítulo 8, é destinado as conclusões.

Capítulo 2

Métodos da classe Escalonamento multidimensional

Existem diferentes funções de erro para o problema de redução de dimensionalidade, muitas dessas formulações levam o nome MDS. Esse termo MDS é designado para uma família de métodos, onde as relações de distância existentes entre as observações em um espaço de alta dimensão são também idealmente preservadas ao máximo em um espaço métrico de mais baixa dimensão.

Para a formalização do problema será utilizada a seguinte notação. Suponha que tenhamos uma matriz de dissimilaridades simétrica, D_{nxn}, calculada a partir de um conjunto de n observações, $\mathbf{z}_i, i = 1, \cdots, n$, cada uma com S atributos, ou seja, cada observação pertence a um espaço de dimensão S. O processo de redução de dimensionalidade busca a representação de cada observação em um espaço de dimensão inferior d, onde $d < S$.

Os métodos de escalonamento multidimensional consistem em obter um novo conjunto de observações, $\mathbf{x}_i, i = 1, \cdots, n$, pertencentes ao espaço de menor dimensão d, ou seja $\mathbf{x}_i \in \Re^d$, de modo que gere uma matriz de distâncias $\widehat{D_{nxn}}$, que se aproxime ao máximo da matriz de dissimilaridades D_{nxn}, tendo com base a função de escalonamento a ser minimizada. A matriz de dissimilaridades pode ser uma matriz de distâncias, não necessariamente de distâncias euclidianas. COX e COX (2000) apresentam uma série de alternativas para medidas de dissimilaridade para dados de natureza quantitativa. A seguir, é apresentada uma revisão dos métodos e suas respectivas formulações que serão abordadas neste trabalho.

2.1 *Least squares MDS* com distância Minkowski

Uma forma simples e natural de se medir o ajuste entre as dissimilaridades no espaço original e as distâncias no espaço de dimensão inferior é com a tradicional função de escalonamento mínimos quadrados com distâncias Minkowski:

$$f(\mathbf{x}) = \frac{1}{c}\sum_{i<j}^{n} w_{ij}(D_{ij} - \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_p)^2 \qquad (1)$$

O caso particular da função de escalonamento com mínimos quadrados, com distância euclidiana, equivalente à distância Minkowski com $p = 2$, conhecida como *Kruscal´s raw Stress*, comumente também chamada de STRESS (COX e COX, 2000):

$$f_{STRESS}(\mathbf{x}) = \frac{1}{c}\sum_{i<j}^{n} w_{ij}(D_{ij} - \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2)^2 \qquad (2)$$

onde D_{ij} representa a dissimilaridade entre as observações $\mathbf{z}_i$ e $\mathbf{z}_j$, w_{ij} representa pesos não negativos, usualmente definido como $w_{ij} = 1$, exceto para dados faltantes onde $w_{ij} = 0$, c é uma constante normalizadora, usualmente igual a soma dos termos da matriz de dissimilaridade $c = \sum_{i<j}^{n} Dij$. Os problemas (1) e (2) são problemas de programação não-linear, não diferenciáveis, definidos em um espaço com nd dimensões. Como o número de observações é, em geral, muito grande trata-se, em geral de um problema de grande porte.

2.2 *Sammon mapping*

SAMMON (1968) em seu método propôs uma função de escalonamento que prioriza a preservação da estrutura local dos dados, pois para cada par de observações utiliza uma ponderação na função de escalonamento inversamente proporcional a distância do par. Desta forma, é dada uma importância relativa maior para os valores de erro associados a pequenas distâncias D_{ij} e menor importâncias a valores altos de D_{ij}.

$$Esamon´s(\mathbf{x}) = \frac{1}{c}\sum_{i<j}^{n} \frac{(D_{ij} - \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2)^2}{Dij} \qquad (3)$$

2.2.1. *Geodesic Nonlinear Mapping*

No método Geodesic Nonlinear Mapping (GNM) (LEE, VERLEYSEN, 2005) a função de escalonamento é a mesma utilizada no método de Sammon, equação (3). Sendo assim, o método GNLM pode ser considerado uma extensão do método de Sammon. No método de Sammon, a matriz de distância D_{nxn} é usualmente calculada com a distância euclidiana, enquanto que no método GNM a distância utilizada é uma aproximação da distância geodésica. Essa aproximação é dada pela matriz de distâncias em um grafo, de modo que cada distância é calculada pelo caminho mínimo entre todos os pares de vértices no grafo. De modo análogo, o método ISOMAP (TENENBAUM, DE SILVA, LANGFORD, 2000) utiliza exatamente a mesma matriz de distâncias em grafo, usualmente calculada pelo algoritmo de Djikstra. O método ISOMAP, diferencia-se do GNM por efetuar a projeção das observações no espaço de baixa dimensão com o método de Escalonamento Multidimensional Clássico.

2.3 *Local MDS*

CHEN e BUJA (2009) desenvolveram o método chamado de Local MDS. O método busca preservar a relação de vizinhança das observações no espaço reduzido tratando de forma diferente observações vizinhas e não vizinhas. Para as observações vizinhas no espaço de alta dimensão, considera uma força local de preservação das distâncias no espaço de baixa dimensão. Para as observações não vizinhas no espaço de alta dimensão considera uma força repulsiva com a finalidade de afastar as observações no espaço de baixa dimensão.

A intensidade da força repulsiva é ajustada através de um parâmetro t. A relação de vizinhança é dada de forma simétrica, pelos k vizinhos mais próximos em um grafo. A simetria é construída de modo que as observações $\mathbf{z}_i$ e $\mathbf{z}_j$ são vizinhas se $\mathbf{z}_j$ está entre os k vizinhos mais próximos de $\mathbf{z}_i$ ou se $\mathbf{z}_i$ está entre os k vizinhos mais próximos de $\mathbf{z}_j$. A função objetivo do método Local MDS tem a expressão:

$$f_{\text{Local MDS}}(\mathbf{x}) = \sum_{(i,j)\in N} (D_{ij} - \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2)^2 \; - t \sum_{(i,j)\notin N} \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2 \quad , (4)$$

onde o conjunto simétrico de pares vizinhos é representado por $(i,j) \in N$. O primeiro termo da equação é chamado de *local stress* e o segundo de termo de repulsão. Considerando um valor fixo para o parâmetro t, a importância relativa da repulsão diminui na medida em que $|N|$ aumenta. Por sua vez a cardinalidade do conjunto de vizinhos, $|N|$, dependente fundamentalmente do valor de k referente ao número de vizinhos mais próximos.

O ajuste do parâmetro t é fundamentado em duas propriedades desejadas:

- Invariância sob mudança na ordem de grandeza das distâncias D_{ij}, por isto tem-se a utilização da mediana das distâncias dos pares de observações vizinhas, $mediana_N(Dij)$, na expressão (5).

- Invariância sob mudança no tamanho do grafo, pois o tamanho do grafo influencia no número de somas do termo local, $|N|$, e no termo de repulsão, $|N^c| = N(N-1)/2 - |N|$, dessa forma tem-se a utilização do fator $\frac{|N|}{|N^c|}$ na expressão (5) definindo assim t:

$$t = \frac{|N|}{|N^c|}.mediana_N(Dij).\tau \quad , \quad (5)$$

onde o parâmetro τ é um parâmetro auxiliar, não dependente do conjunto de dados de entrada e do valor de k vizinhos mais próximos. O parâmetro τ tem a função de facilitar o ajuste do parâmetro t.

Para mapeamentos realizados com diferentes valores de k e de t, uma medida da qualidade da redução de dimensionalidade tem que ser aplicada para se ter mapeamentos comparáveis.

2.4 *Least Absolute Residuals* (LAR)

HEISER (1988) propôs a utilização de soma de valores absolutos entre as diferenças das distâncias no espaço de alta dimensão e as distâncias no espaço de baixa dimensão. A métrica de mínima soma de valores absolutos é muito utilizada nos métodos robustos, resistentes a efeitos dos *outlier* nos dados, que são observações que apresentam valores discrepantes em relação a outras observações do conjunto de dados.

O método LAR tem a seguinte função de escalonamento:

$$f_{LAR}(\mathbf{x}) = \frac{1}{c}\sum_{i<j}^{n} w_{ij}\left|D_{ij} - \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2\right| \quad (6)$$

Capítulo 3

Classificação Supervisionada e *Ranking* Bipartido

Em um cenário típico de classificação supervisionada, busca-se, a partir de um conjunto de atributos, prever um resultado associado a uma nova observação tendo como base um modelo de previsão ou de aprendizagem. O conjunto de treinamento, conjunto esse no qual todas as observações possuem além dos atributos um respectivo resultado associado, é utilizado para construir um modelo de previsão ou de aprendizagem. Esse modelo então, nos permite com a informação dos atributos de uma observação nova prever o seu respectivo resultado associado.

Um bom modelo é aquele que prevê com precisão o resultado associado a cada observação. O termo supervisionado é devido à presença de uma variável resultado que guia o processo de construção do modelo na fase de treinamento. O resultado associado pode ser uma variável quantitativa ou uma variável qualitativa, representando duas ou mais classes. Na aprendizagem não supervisionada utiliza-se somente os atributos não tendo a variável de resultado no processo (FRIEDMAN, HASTIE, TIBSHIRANI, 2001).

O conjunto de teste é o conjunto de observações que não foram utilizadas para construir o modelo e são utilizadas na avaliação do modelo. Cada observação desse conjunto de teste possui os atributos e o respectivos resultado associado. Para cada observação, utiliza-se o modelo somente com o conjunto de atributos para estimar ou prever o resultado associado. A avaliação do modelo se dá na comparação do resultado previsto com o conhecido resultado associado à observação.

No problema geral de *ranking* ou ordenação, busca-se no espaço de alta dimensão uma relação de ordem ou de preferência entre as observações. Uma aplicação típica e relevante está presente em sistemas de recomendação e ferramentas de busca e

recuperação da informação, onde busca-se uma ordenação que induz uma relação de preferências. O problema de *ranking* bipartido é um caso especial do problema geral de *ranking* em que as observações possuem duas classes. A informação das classes é utilizada na construção do modelo de aprendizagem fundamentado em uma função que induz o *ranking* nas observações. Esse problema esta relacionado com o problema de classificação supervisionada binária. Desse modo, alguns algoritmos de classificação supervisionada podem ser usados no contexto do problema do *ranking* bipartido (AGARWAL, 2005). Como exemplo temos o problema da ordenação de importância de e-mails utilizando como base as informações das classes *Spam* e Não *Spam*.

Os problemas de classificação supervisionada e *ranking* bipartido são abordados com o método Supervised MDS. A seguir é apresentada uma revisão desse método.

3.1 *Supervised MDS* para observações do conjunto de treinamento e visualização

Recentemente, WITTEN e TIBSHIRANI (2011) desenvolveram um novo método chamado de *Supervised Multidimensional Scaling* (SMDS) para visualização, classificação supervisionada e *ranking* bipartido. De forma eficiente, esse método considera simultaneamente os problemas de redução de dimensionalidade e o problema de classificação supervisionada, incorporando a informação da classe na determinação da localização das observações no espaço reduzido.

SMDS adota a seguinte função de escalonamento:

$$SupervMDS(\mathbf{x},\alpha)=\left\{\begin{array}{l}\frac{1}{2}(1-\alpha)\sum_{i=1}^{n}\sum_{j=1}^{n}(D_{ij}-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|_2)^2\ + \\ \alpha\sum_{i,j:y_j>y_i}(y_j-y_i)\sum_{l=1}^{d}(\frac{D_{ij}}{\sqrt{d}}-(x_{jl}-x_{il}))^2\end{array}\right\} \qquad (7),$$

onde $\mathbf{y}\in\Re^n$ é um vetor binário com a informação das classes, y_i representa a classificação ou resultado associado à observação i, $y_i\in\{1,2\}$, α é um parâmetro que balanceia o primeiro termo da função objetivo com o segundo, $\alpha\in[0,1]$. O primeiro termo corresponde à formulação *Least squares MDS*, enquanto o segundo termo contempla a componente associada às informações das classes.

A segunda parcela do somatório tende a afastar as observações de classes diferentes, $y_i \neq y_j$. Essa parcela do somatório só ocorre para os valores nos quais $y_j > y_i$. Observações na classe 2, tenderão a ter valores maiores do que as observações na classe 1 em todas as d componentes, pois ao minimizar o desvio quadrático em $(\frac{D_{ij}}{\sqrt{d}} - (x_{jl} - x_{il}))^2$, os termos $\frac{D_{ij}}{\sqrt{d}}$ são sempre positivos, portanto, no processo de minimização cada parcela $(x_{jl} - x_{il})$ tende a ser positiva. Essa parcela é positiva se e somente se $x_{jl} > x_{il}$, sendo assim as observações na classe 2 tenderão a ter valores maiores do que as observações na classe 1. Embora a formulação (6) considere a informação da classe ela aborda somente o problema de redução de dimensionalidade e não generaliza para uma observação nova.

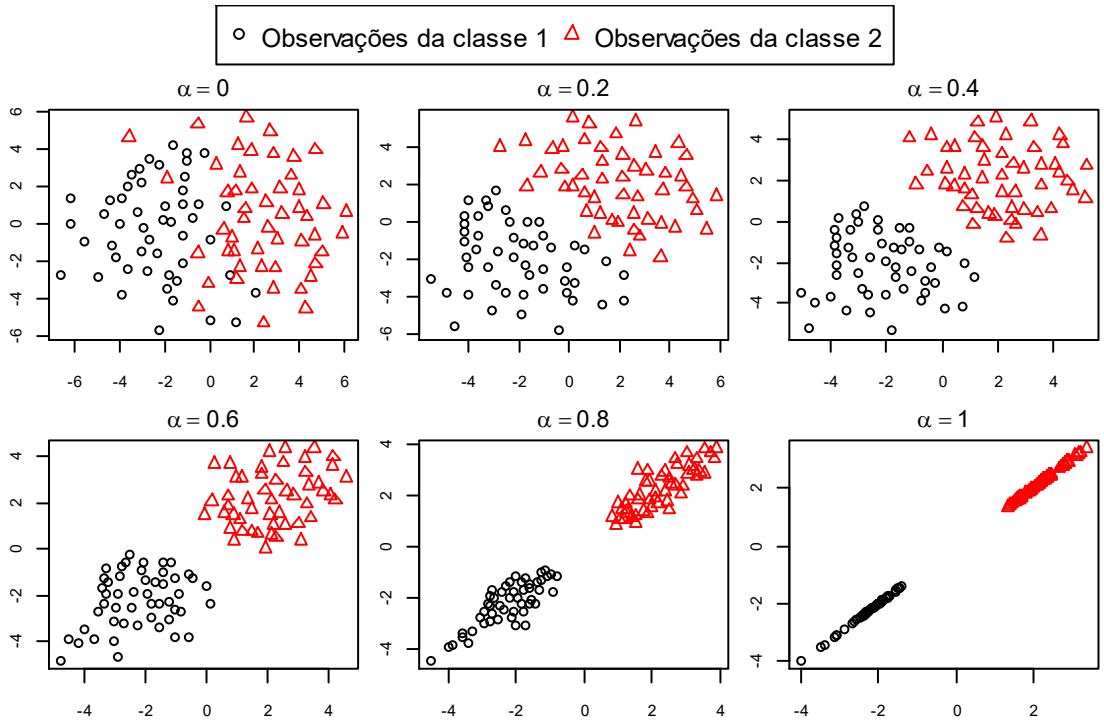

Figura 3.1: Efeito do parâmetro α.

A Figura 3.1 apresenta um exemplo sobre o efeito do parâmetro α. O conjunto de dados foi gerado por uma mistura de duas distribuições normais multivariadas de mesmo tamanho no espaço R^{15}: $x \sim N(\mathbf{0.4}, \mathbf{I})$, $x \sim N(-\mathbf{0.4}, \mathbf{I})$, com matriz de covariância igual à matriz identidade e cada uma das distribuições com o vetor de média

com todas as componentes iguais. Utilizando cada uma das distribuições foram geradas 50 observações, desta forma tem-se 50 observações em cada classe. Na medida em que o parâmetro α aumenta as observações de classes diferentes tendem a se afastar, aumentando a separabilidade das classes. Pode-se verificar no primeiro gráfico da Figura 3.1 que quando α é igual a zero existe uma clara mistura das duas classes, com um leve aumento no parâmetro as classes se separam. Nos seis gráficos da Figura 3.1 à medida que o parâmetro α é aumentado, gradualmente pode ser notada uma crescente separação entre essas duas classes.

Numa primeira fase, com a utilização da informação das classes, a formulação (7) é utilizada para redução de dimensionalidade das observações do conjunto de treinamento e fornece a base para a aplicação da formulação que considera uma nova observação e, por conseguinte, a sua classificação.

3.2 *Supervised MDS* para classificação

O critério *Supervised MDS* para o problema de classificação e de redução de dimensionalidade utiliza um conjunto de treinamento, composto de n observações, cada uma com a informação da classe, $\mathbf{y}$. Pretende-se estimar as coordenadas de uma observação nova no espaço reduzido, $\mathbf{x}_{n+1}$, e a classe desta nova observação y_{n+1}. Para isso, a seguinte notação é definida:

O índice $n+1$ faz referência a uma observação nova genérica e $\mathbf{D}_{n+1} = \left[D_{1,n+1}, \ldots, D_{n,n+1} \right]^T$ representa o vetor de distâncias no espaço de alta dimensão entre cada observação do conjunto de teste e a observação nova.

O critério *Supervised MDS* para a classificação generaliza a formulação (7) através de dois novos procedimentos considerando: a inclusão da observação nova na função objetivo e a utilização do vetor $\mathbf{x}$, resultante da minimização da formulação (7), como referência no espaço de baixa dimensão. O vetor $\mathbf{x}$ corresponde às observações do conjunto de treinamento no espaço reduzido. A generalização da formulação (7) é feita com a suposição hipotética sobre a classe da observação nova. Desta forma:

A formulação $h_1(D_{n+1}, \mathbf{x}_{n+1})$ considera a hipótese de a observação nova pertencer a classe 1, ou seja, $y_{n+1} = 1$.

$$h_1(D_{n+1}, \mathbf{x}_{n+1}, \alpha) = \left\{ \begin{array}{l} (1-\alpha)\sum_{i=1}^{n} (D_{i,n+1} - \left\| \mathbf{x}_i - \mathbf{x}_{n+1} \right\|_2)^2 \; + \\ \alpha \sum_{i:y_i=2, 1 \le i \le n} \sum_{l=1}^{d} (\frac{D_{i,n+1}}{\sqrt{d}} - (x_{il} - x_{n+1,l}))^2 \end{array} \right\} \quad (8)$$

A formulação $h_2(D_{n+1}, \mathbf{x}_{n+1})$ considera a hipótese de a observação nova pertencer a classe 2 ou seja, $y_{n+1} = 2$.

$$h_2(D_{n+1}, \mathbf{x}_{n+1}, \alpha) = \left\{ \begin{array}{l} (1-\alpha)\sum_{i=1}^{n} (D_{i,n+1} - \left\| \mathbf{x}_i - \mathbf{x}_{n+1} \right\|_2)^2 \; + \\ \alpha \sum_{i:y_i=1, 1 \le i \le n} \sum_{l=1}^{d} (\frac{D_{i,n+1}}{\sqrt{d}} - (x_{n+1,l} - x_{il}))^2 \end{array} \right\} \quad (9)$$

Para cada função $h_1(D_{n+1}, \mathbf{x}_{n+1})$ e $h_2(D_{n+1}, \mathbf{x}_{n+1})$, minimiza-se em relação a $\mathbf{x}_{n+1}$, obtendo duas alternativas para as coordenadas da observação nova no espaço reduzido.

Como exemplo um conjunto de teste foi gerado utilizando o mesmo modelo apresentado na Figura 3.1, com 50 observações em cada classe. Cada observação do conjunto de teste é considerada como uma observação nova. A Figura 3.2 apresenta resultados da minimização de $h_1(D_{n+1}, \mathbf{x}_{n+1})$ e $h_2(D_{n+1}, \mathbf{x}_{n+1})$ considerando o vetor $\mathbf{x}$ obtido pela minimização de (7) para o conjunto de treinamento e considerando $\alpha = 0,8$. É importante ressaltar que nos dois gráficos da Figura 3.2 são utilizadas todas as 100 observações do conjunto de treinamento.

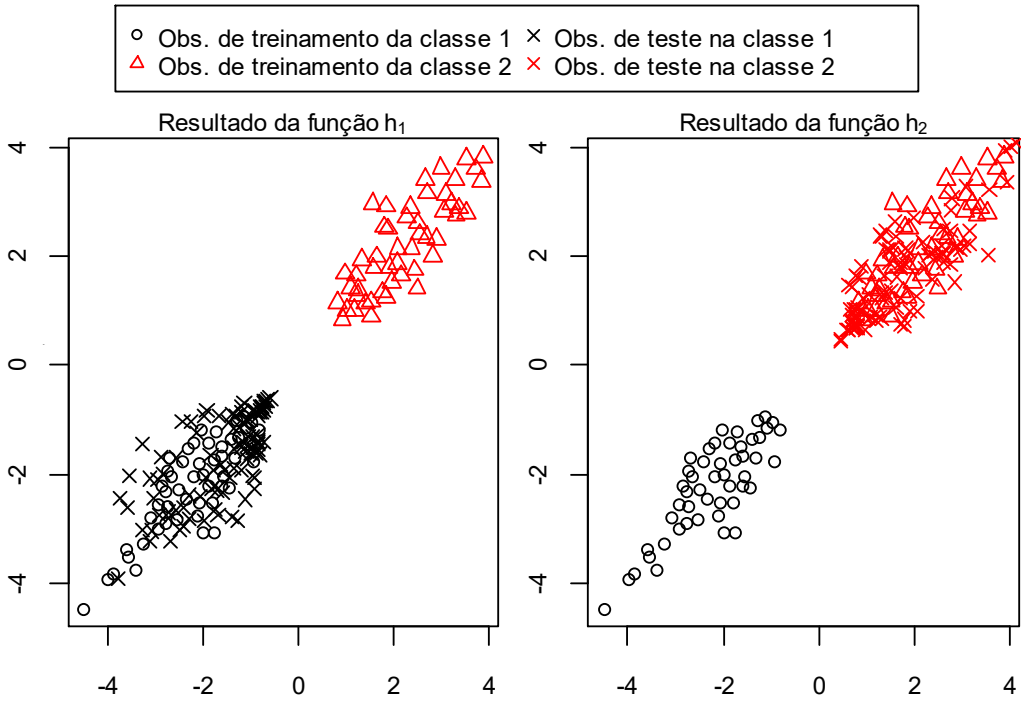

Figura 3.2: Resultados para o conjunto de teste com $\alpha = 0{,}8$

A regra de classificação mais simples atribui a classe da observação nova de acordo com o índice da função $h_k, k \in \{1,2\}$ correspondente à função com o menor valor no processo de otimização. Sendo assim, a classe imputada corresponde àquela que produziu um menor resíduo no processo de redução de dimensionalidade. As coordenadas geradas pela minimização da função h_k que definiu a classe correspondem às coordenadas da observação nova no espaço reduzido. Desta forma tem-se:

$$\hat{y}_{n+1} = \underset{k \in \{1,2\}}{\arg\min} \left\{ \min_{x_{n+1}} h_k(D_{n+1}, x_{n+1}) \right\}$$

A regra de classificação pode ser representada de forma equivalente a:

$$\begin{cases} se \ \min_{\mathbf{x}_{n+1} \in \mathbb{R}^d} h_1(D_{n+1}, \mathbf{x}_{n+1}) - \min_{\mathbf{x}_{n+1} \in \mathbb{R}^d} h_2(D_{n+1}, \mathbf{x}_{n+1}) < 0 \ \rightarrow \ \hat{y}_{n+1} = 1 \\ caso\ contrário \ \hat{y}_{n+1} = 2 \end{cases}$$

De uma forma coerente, as coordenadas da observação nova são obtidas por:

$$\hat{\mathbf{x}}_{n+1} = \arg\min_{\mathbf{x}_{n+1}\in\mathbb{R}^d} \left\{ h_1(D_{n+1}, \mathbf{x}_{n+1}), h_2(D_{n+1}, \mathbf{x}_{n+1}) \right\}$$

A classificação pode ser feita de um modo mais geral, substituindo o valor zero por um valor fixo, v , representando um ponto de corte genérico:

$$\begin{cases} se \ \min_{\mathbf{x}_{n+1}\in\mathbb{R}^d} h_1(D_{n+1}, \mathbf{x}_{n+1}) - \min_{\mathbf{x}_{n+1}\in\mathbb{R}^d} h_2(D_{n+1}, \mathbf{x}_{n+1}) < v \ \rightarrow \ \hat{y}_{n+1} = 1 \\ caso\ contrário \ \ \hat{y}_{n+1} = 2 \end{cases}$$

A Figura 3.3 apresenta as 100 observações do conjunto de teste com suas classificações obtidas pelo procedimento acima descrito com o valor do parâmetro $\alpha = 0{,}8$. Deve ser observado que nesse exemplo o classificador obteve uma acurácia de 94%, pois das 100 observações 6 foram classificadas erradas.

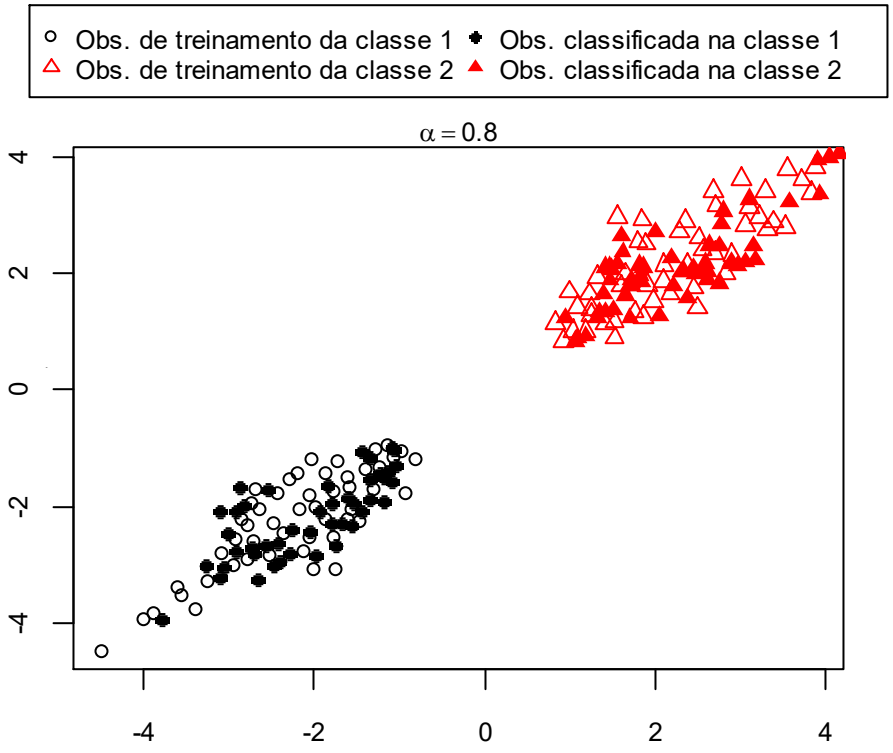

Figura 3.3: Resultado da Classificação das observações do conjunto de teste

O parâmetro α pode ser ajustado em função do erro de classificação fora da amostra utilizando validação cruzada. Deve ser observado que quando $\alpha = 0$ SMDS é equivalente a formulação *Least Squares MDS*.

3.3 *Supervised MDS* para *Ranking* Bipartido

O problema de *ranking* bipartido é um caso especial do problema de *ranking* e é diretamente relacionado com o problema de classificação supervisionada. De modo análogo ao problema de classificação supervisionada, considere um conjunto de n observações, pertencentes a um espaço de dimensão S, no qual cada observação possui uma classe associada, $y_i \in \{1,2\}$. No problema de *ranking* bipartido, busca-se uma função f, com imagem no espaço de dimensão 1, $f : R^S \to R^1$. A ordenação dos valores obtidos pela função *f* aplicada nas n observações produz um *ranking*. De forma análoga ao problema de classificação, no problema de *ranking* bipartido deseja-se um separar as classes através de um *ranking*, ou ordenamento. Dessa forma, alguns algoritmos de classificação podem ser utilizados para produzir o *ranking*.

Para qualquer par de observações $\mathbf{x}_i$ e $\mathbf{x}_j$, tem-se a seguinte propriedade desejada para a função f: $f(\mathbf{x}_i) > f(\mathbf{x}_j)$ se $y_i > y_j$. O número de pares onde essa propriedade não é satisfeita é um indicador da qualidade do *ranking*. Uma medida comum de avaliação do *ranking* é o erro do *ranking* bipartido, definido por:

$$\frac{1}{n_1 n_2} \sum_{i:y_i=1} \sum_{j:y_j=2} \left(1_{f(\mathbf{x}_j)>f(\mathbf{x}_i)} + \frac{1}{2} 1_{f(y_i)=f(y_i)} \right)$$

No método *Supervised MDS* o *ranking* bipartido é construído pela ordenação das diferenças $\min_{\mathbf{x}_{n+1}} h_1(D_{n+1}, \mathbf{x}_{n+1}) - \min_{\mathbf{x}_{n+1}} h_2(D_{n+1}, \mathbf{x}_{n+1})$.

Capítulo 4

Suavização Hiperbólica

Com o objetivo de obter uma alternativa suave ou diferenciável para as formulações (2), (3), (4), (6), (7), (8), (9) o conceito da Suavização Hiperbólica (SH) é apresentada neste capítulo. A abordagem SH tem sido aplicada com sucesso num grande número de problemas difíceis da classe Np-hard, incluindo problemas não diferenciáveis e não convexos com um grande número de mínimos locais. Dentre essa classe de problemas difíceis resolvidos utilizando a SH pode-se citar por exemplo: recobrimento de uma região planar (XAVIER, OLIVEIRA, 2005), recobrimento de um corpo sólido (VENCESLAU *et al.*, 2014), distância geométrica, formulação (2) (XAVIER, 2003), (SOUZA *et al.*, 2011), *clustering* (XAVIER, 2010, XAVIER, XAVIER, 2011, BAGIROV *et al.*, 2014), *Multisource Fermat-Weber* (XAVIER *et al.* 2014b) e *hub location* (XAVIER *et al.* 2015). Um conjunto de aplicações bem-sucedidas pode ser encontrado em (XAVIER, XAVIER, 2014).

Dados dois pontos x_i e x_j no R^n e u sendo a distância euclidiana entre x_i e x_j, $u = \left\| \mathbf{x}_i - \mathbf{x}_j \right\|_2$. A distância euclidiana é não diferenciável quando $x_i = x_j$. A função $\theta(u, \gamma)$ usada para suavização é definida por:

$$\theta(u, \gamma) = \sqrt{u^2 + \gamma^2}$$

A função $\theta(u, \gamma)$ tem as seguintes propriedades (Xavier, Xavier, 2014):

(a) $\theta(u, \gamma) > u, \quad \forall \gamma > 0;$

(b) $\lim_{\gamma \to 0} \theta(u, \gamma) = u;$

(c) $\theta(u, \gamma)$ pertence à classe c^{∞} de funções diferenciáveis.

(d) $\theta^{`}(u, \gamma) = u / (u^2 + \gamma^2)^{1/2}$

(e) $\theta^{``}(u, \gamma) = \gamma^2 / (u^2 + \gamma^2)^{3/2}$

(f) $\theta^{``}(0,\gamma) = 1/|\gamma|$

(g) $\lim_{\gamma \to 0} \theta^{``}(0,\gamma) = \infty$

Note que a propriedade (b) implica que a função $\theta(u,\gamma)$ é uma aproximação assintótica da função u. A propriedade (c) permite a utilização de métodos de otimização poderosos baseados em aproximação da série de Taylor de primeira ou de segunda ordem. Devido à propriedade (e), a curvatura da função $\theta(u,\gamma)$ é crescentemente atenuada na medida em que se aumenta o parâmetro γ. Pela propriedade (f), a curvatura assume o valor máximo quando $u = 0$. De forma similar pelas propriedades (b) e (g), a curvatura da função $\theta(u,\gamma)$ no ponto $u = 0$ tende a infinito na medida em que $\theta(u,\gamma)$ se aproxima de u.

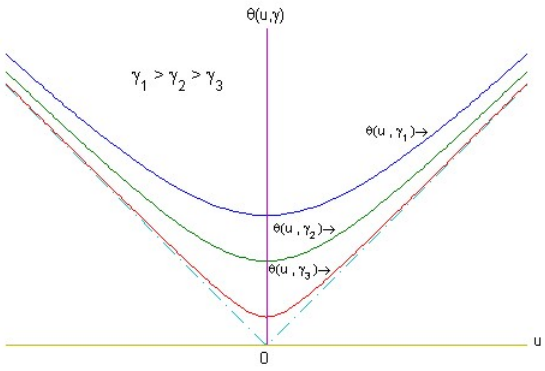

Figura 4.1: Função $\theta(u,\gamma)$ com sequência de valores decrescentes do parâmetro γ

A Figura 4.1 ilustra as propriedades (a) e (b). Pode-se ver que para a sequência de valores decrescente do parâmetro γ, $\gamma_1 > \gamma_2 > \gamma_3$, a função $\theta(u,\gamma)$ se aproxima da função u. Pela definição de $\theta(u,\gamma)$ pode-se verificar facilmente a propriedade adicional:

$$\max_u(\theta(u,\gamma) - u) = \theta(0,\gamma) = \gamma$$

4.1 Aplicando a Suavização Hiperbólica em MDS

Voltando a atenção para as formulações de MDS expressas pelas funções (2), (3), (4), (6), (7), (8) e (9) nota-se que a não diferenciabilidade é uma propriedade comum nessas sete formulações. Tem-se como proposta aplicar SH e substituir essas funções por um conjunto de problemas sucedâneos bem-comportados. Com este propósito em mente, problemas intrinsecamente não diferenciáveis são aproximados por alternativas completamente diferenciáveis.

Por conseguinte, a função *Least Squares MDS* (2) será substituída por (XAVIER, 2003, XAVIER, XAVIER, 2015):

$$\tilde{f}_{STRESS}(\mathbf{x},\gamma) = \frac{1}{c}\sum_{i<j}^{n} w_{ij}\,(D_{ij} - \theta(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{2,}\,\gamma))^2 \qquad (10)$$

A função *Sammon mapping* (3) será substituída por:

$$\tilde{f}_{\,Esamon's}(\mathbf{x},\gamma) = \frac{1}{c}\sum_{i<j}^{n} \frac{(D_{ij} - \theta(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{2,}\,\gamma))^2}{Dij} \qquad (11)$$

A função *Local MDS* (4) será substituída por:

$$\tilde{f}_{\text{Local MDS}}(\mathbf{x},\gamma_1,\gamma_2) = \sum_{(i,j)\in N}(D_{ij} - \theta(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{2,}\,\gamma_1))^2 \; - t\sum_{(i,j)\notin N}\theta(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{2,}\,\gamma_2) \qquad (12)$$

A função do *Least Absolute Residuals* (6) será substituída por:

$$\tilde{f}_{LAR}(\mathbf{x},\gamma_1,\gamma_2) = \frac{1}{c}\sum_{i<j}^{n} w_{ij}\;\theta_1\,(D_{ij} - \theta_2(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{2,}\,\gamma_2)\;,\;\gamma_1) \qquad (13)$$

A função *Supervised MDS* (7) será substituída por:

$$\tilde{f}_{SupervMDS}(\mathbf{x},\alpha,\gamma) = \left\{ \begin{array}{l} \frac{1}{2}(1-\alpha)\sum_{i=1}^{n}\sum_{j=1}^{n}(D_{ij} - \theta(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{2,}\,\gamma))^2 \; + \\ \alpha \sum_{i,j:y_j>y_i}(y_j - y_i)\sum_{l=1}^{d}(\frac{D_{ij}}{\sqrt{d}} - (x_{jl} - x_{il}))^2 \end{array} \right\} \qquad (14)$$

As funções $h_1(D_{n+1}, \mathbf{x}_{n+1})$ e $h_2(D_{n+1}, \mathbf{x}_{n+1})$ serão substituídas respectivamente por:

$$\tilde{h}_1(D_{n+1}, \mathbf{x}_{n+1}, \gamma) = \left\{ \begin{array}{l} (1-\alpha)\sum_{i=1}^{n}(D_{i,n+1} - \theta(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{2,} \gamma))^2 \; + \\ \alpha \sum_{i:y_i=2, 1\le i\le n} \sum_{l=1}^{d}(\frac{D_{i,n+1}}{\sqrt{d}} - (x_{il} - x_{n+1,l}))^2 \end{array} \right\} \quad (15)$$

$$\tilde{h}_2(D_{n+1}, \mathbf{x}_{n+1}, \gamma) = \left\{ \begin{array}{l} (1-\alpha)\sum_{i=1}^{n}(D_{i,n+1} - \theta(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{2,} \gamma))^2 \; + \\ \alpha \sum_{i:y_i=1, 1\le i\le n} \sum_{l=1}^{d}(\frac{D_{i,n+1}}{\sqrt{d}} - (x_{n+1,l} - x_{il}))^2 \end{array} \right\} \quad (16)$$

Capítulo 5

Métodos numéricos para MDS

Diferentes abordagens para resolver problemas de redução de dimensionalidade da classe MDS têm sido propostas na literatura. Métodos de primeira ordem são abordagens tradicionais para resolver problemas de otimização não-linear. As primeiras contribuições relativas à minimização da função de *Least squares MDS* utilizaram método de primeira ordem, KRUSKAL (1964b) e GUTTMAN (1968) utilizam o algoritmo do tipo *steepest descent*.

SAMMON (1969), a fim de assegurar uma diminuição do valor da função de escalonamento, utiliza o algoritmo de Newton em conexão com as estratégias de busca linear com passo fixo, entre 0,3 e 0,4, sendo chamado de fator mágico. Kearsley et al. (1998) apresentam um algoritmo com base em métodos de segunda ordem para resolver problemas de *Least squares* MDS, cujos resultados computacionais obtidos mostram um bom desempenho do algoritmo de Newton.

Como as funções de MDS são não convexas, com um grande número de mínimos locais, esses métodos de otimização tradicionais têm um sucesso limitado para resolver problemas de MDS. Este fato tem motivado o aparecimento de alternativas como o algoritmo SMACOF (*Scaling by MAjorizing a COmplicated Function*) (LEEUW, MAIR, 2009) para a função de escalonamento STRESS.

SMACOF é um algoritmo moderno que se baseia no princípio da majoração iterativa (MI), também conhecido como Majorização. MI foi introduzido no contexto de MDS por De LEEUW (1977). De forma muito simplificada, MI adota uma função objetivo auxiliar, $g(x, y)$, que satisfaz a seguinte desigualdade: $g(x, y) \geq f(x)$, ou seja, é sempre maior ou igual do que a função objetivo original. A relação de igualdade é satisfeita quando $y = x$, sendo assim, $g(x, x) = f(x)$. Além disso, a função objetivo auxiliar deve ser uma alternativa mais fácil para o procedimento de otimização, caso contrário não valeria a pena substituir a função original. Para maiores detalhes sobre MI

uma abordagem didática é apresentada em BORG e GROENEN (2005). O algoritmo SMACOF tem a propriedade de eliminação de mínimos locais, sendo esta propriedade ilustrada em BORG e GROENEN (2005).

WITTEN e TIBSHIRANI (2011) publicaram o método *Supervised MDS* e um algoritmo fundamentado no princípio da MI, sendo a função de majorização derivada da desigualdade de Cauchy-Schwarz. No algoritmo apresentado, o processo de minimização da função de majorização é realizado através do método do gradiente.

Na literatura existem diferentes abordagens de suavização em algoritmos de redução de dimensionalidade. Para suavização do problema escalonamento unidimensional métrico com função Least squares MDS, PLINER (1996) propõe a utilização da distância diferenciável de segunda ordem:

$$g(u,\xi) = \begin{cases} u^2(3\xi - |u|)/3\xi^2 + \xi/3 & if\ |u| < \xi \\ |u| \in & if\ |u| \geq \xi \end{cases}$$

GROENEN *et al.* (1999) propuseram um algoritmo fundamentado na MI e o uso da função Huber como uma forma alternativa para suavizar a função objetivo Least squares MDS. Eles também estenderam o algoritmo de majorização para o problema de *Least squares MDS* com distância Minkowski.

A utilização da Suavização Hiperbólica na formulação (2) foi apresentada originalmente em (XAVIER, 2005). Nesta tese é proposta a generalização do uso da Suavização Hiperbólica em diversos métodos de redução de dimensionalidade, tendo como contribuições inovadoras as abordagens nos métodos de Sammon, Local MDS e *Least Absolute Residuals* e Supervised MDS.

5.1 Algoritmo HSDR

Os problemas definidos pelas funções (2), (3), (4), (6), (7), (8) e (9) podem ser resolvidos por meio de algoritmos de programação não-linear. Nota-se que, para estas formulações, as funções objetivo são não lineares, não convexas e não diferenciáveis com um grande número de mínimos locais. Ademais, não existe um algoritmo determinístico para o qual uma solução pode ser obtida em tempo polinomial.

Para resolver as formulações suavizadas (10) (11) (12) (13) (14), propõe-se um algoritmo único, chamado de *Hyperbolic Smoothing Dimensionality Reduction* (HSDR). A idéia central por trás do algoritmo HSDR é resolver uma seqüência de

problemas diferenciáveis, onde o parâmetro de suavização é gradualmente reduzido por um fator de redução, princípio este apresentado em (XAVIER, 2003, XAVIER, OLIVEIRA, 2005, XAVIER, XAVIER, 2011, BAGIROV *et al.*, 2014). Assim, é gerada uma sequência de problemas suavizados que se aproximam gradativamente dos problemas originais (2), (3), (4), (6) e (7). Cada problema nessa sequência de aproximação é completamente diferenciável, permitindo assim a aplicação de métodos eficientes e robustos de otimização sem restrições, tais como o Gradiente Conjugado e os métodos da família Quasi-Newton. O algoritmo proposto HSDR, em sua varredura principal, gera uma sequência de pontos intermediários que têm uma convergência assintótica a um ponto de mínimo local.

Algoritmo HSDR

Fase de inicialização

Calcule a matriz de distância D

Defina o ponto inicial $\mathbf{x}^0$

Defina o parâmetro inicial de suavização $\gamma > 0$

Defina o fator de redução $0 < \rho < 1$

Inicialize contador de iterações $\iota = 1$

Varredura principal: Repita até que uma regra arbitrária de parada seja satisfeita.

Resolva o problema irrestrito, uma das formulações (10) (11) (12) (13) e (14), a partir do ponto inicial $\mathbf{x}^{\iota-1}$ com o parâmetro γ^{ι}, obtendo uma solução intermediária $\mathbf{x}^{\iota}$

Atualize o parâmetro $\gamma^{\iota+1} = \rho\gamma^{\iota}$

Incremente o contador $\iota = \iota + 1$

Fim da varredura principal

Os métodos de preservação de distância requerem $n(n-1)/2$ posições de memória para armazenar todos os pares de distancias. A complexidade de tempo para calcular as distâncias é da ordem de $O(Rn^2)$ para o caso de distâncias euclidianas,

enquanto para o caso de distâncias em Grafo a complexidade cresce ainda mais, para $O\left(Rn^2 \log n\right)$ (LEE, VERLEYSEN, 2007). Para obter o mapeamento no espaço de baixa dimensão, os métodos não lineares, que utilizam o método do gradiente descendente, requerem $O\left(dn^2\right)$ operações por uma única iteração (LEE, VERLEYSEN, 2007).

Do ponto de vista prático, vale a pena destacar que o possível grande número de atributos S, especialmente no contexto de *big data*, não é usado nos procedimentos dominantes em termos de tempo total que correspondem à resolução do problema de otimização. O número de atributos S influencia unicamente no cálculo da matriz de distância de entrada dos métodos, sendo essa calculada uma única vez. O processo de otimização é dependente do número de observações, n , e da dimensão do espaço reduzido, d , definindo assim o número de variáveis do problema igual a nd . Em termos computacionais esse número de variáveis inviabiliza a aplicação prática dos métodos de redução de dimensionalidade não lineares em conjuntos de dados com um número muito grande de observações.

5.2 Aspectos da implementação

A especificação do parâmetro inicial de suavização, γ , é feita por uma estratégia automática e dependente unicamente do conjunto de dados de entrada. Esse valor é tomado como uma fração do valor máximo das distâncias da matriz de dissimilaridade:

$$\gamma = \frac{1}{20} \max_{ij}(D_{ij})$$

Note que a escolha de um número excessivamente grande para o fator de redução, ρ , pode introduzir grandes perturbações na seqüência de pontos intermediários de mínimos e fazer com que a convergência desta seqüência não seja harmônica. Por outro lado, se a escolha for muito pequena, o tempo de processamento pode aumentar de uma forma desnecessária. O valor de $\rho = 1/2$ mostrou-se empiricamente adequado nos experimentos numéricos. No que se diz respeito à regra de parada, adotando as especificações acima dos parâmetros ρ e γ um número fixo de oito iterações para a varredura principal mostrou-se empiricamente adequado.

A metodologia proposta foi implementado utilizando a linguagem estatística R (R Core Team, 2014). As tarefas de otimização foram realizadas por meio do método de Gradiente Conjugado, através da rotina *optim* da biblioteca *stats*. Em relação à regra de parada, em cada problema irrestrito utilizou-se o máximo de 150 iterações do método Gradiente Conjugado.

No método Supervised MDS, as observações do conjunto de treinamento projetadas no espaço reduzido são obtidas através do algoritmo HSDR, considerando a minimização da equação (14). Para a resolução dos problemas de classificação supervisionada e *ranking* bipartido, a redução de dimensionalidade de cada observação nova é independente uma das outras. As coordenadas destas observações novas no espaço de baixa dimensão são obtidas através da minimização das funções na forma original (8) e (9) ou das correspondentes funções propostas (15) e (16).

O algoritmo apresentado em (WITTEN, TIBSHIRANI, 2011) para minimização das funções (8) e (9) não faz uso de processamento paralelo. Por serem processos de otimização independentes, nesta tese é proposta a utilização de computação paralela na minimização das funções (8), (9) na forma original e nas sucedâneas suavizadas (15) e (16) para cada observação do conjunto de teste. Foi implementado o processamento paralelo em CPUs utilizando a estrutura de processamento paralelo da biblioteca *doSNOW* para a minimização das funções suavizadas (15) e (16). Para essas formulações não se utilizou o processo de decrescimento do parâmetro γ e sim um valor de γ fixo baixo, para evitar erro numérico no gradiente decorrente da operação da divisão por zero.

Capítulo 6

Resultados computacionais

6.1 Comparação direta com outros algoritmos disponíveis na literatura.

Nesta seção, trata-se de apresentar e de discutir os experimentos computacionais com o objetivo de se ter uma validação do desempenho da metodologia proposta, através da comparação direta do algoritmo proposto com outros algoritmos disponíveis na literatura. Esses experimentos foram realizados primeiramente com conjuntos de dados sintéticos e posteriormente com conjuntos de dados reais apresentados na literatura.

6.1.1. Experimentos com conjuntos de dados sintéticos

Os conjuntos de dados sintéticos foram gerados do modo apresentado em WITTEN e TIBSHIRANI (2011), consistindo uma mistura de duas classes de tamanhos iguais provenientes das distribuições normais multivariadas no R^{10}: $\mathbf{z} \sim N(\mathbf{0.4}, \mathbf{I})$ e $\mathbf{z} \sim N(-\mathbf{0.4}, \mathbf{I})$, com matriz de covariância igual à identidade.

Nos experimentos foram realizadas projeções em espaços de 2 e 3 dimensões. Foram gerados dez tipos de conjuntos de dados de diferentes tamanhos, possuindo 10, 20,..., 100 observações. Para cada tipo de conjunto de dados foram simulados 100 conjuntos de dados diferentes.

Para efeito de avaliação de desempenho do HSDR, realizou-se comparações com as seguintes rotinas disponíveis no *software* R (R Core Team, 2014):

- Para a formulação de mínimos quadrados, equação (2), foi utilizada a função smacofSym da biblioteca *smacof* (LEEUW, MAIR, 2009);

- Para o mapeamento de Sammon, equação (3), foi utilizada a função de *Sammon* da biblioteca *MASS* (VENABLES, RIPLEY, 2013);

- Para *SupervisedMDS*, equação (7), foi utilizada a função TrainSuperMDSOnce da biblioteca *superMDS* (WITTEN, TIBSHIRANI, 2011).

Problemas da classe MDS são computacionalmente dispendiosos, particularmente quando o número de observações é grande, fato este que determina problemas de otimização com muitas variáveis. Trata-se de um problema de otimização global com um grande número de mínimos locais. Por conseguinte, a obtenção de um ponto de mínimo local profundo é bastante sensível à escolha do ponto inicial. Em relação aos pontos iniciais, foram utilizados dois procedimentos habituais: ponto de inicialização com base na fatoração da matriz de distância e ponto por inicialização aleatória. O Escalonamento Multidimensional Clássico, também conhecido como *principal coordinates analysis* (GOWER, 1966) fornece o ponto inicial baseado na fatoração matricial, sendo também o ponto inicial padrão para as funções *Sammon* e *SupervisedMDS*, funções estas utilizadas em nossos procedimentos de comparação. Utilizou-se a função *CMDSCALE* da biblioteca *stats* do R para gerar o ponto inicial pelo Escalonamento Multidimensional Clássico.

O desvio relativo, definido abaixo, foi utilizado como medida de comparação do HSDR com os outros algoritmos:

$$D(x^*, x^*_{HS}) = \frac{f(x^*) - f(x^*_{HS})}{f(x^*_{HS})}, \quad (17)$$

onde f é uma das funções objetivo (2) , (3) e (7), x^* é o ponto ótimo no espaço reduzido obtido por um dos algoritmos comparativos (*smacofSym*, *TrainSuperMDSOnce* ou *Sammon*), x^*_{HS} é o ponto ótimo no espaço reduzido obtido por HSDR. A fim de produzir uma comparação justa, todas as funções foram testadas com os mesmos pontos iniciais.

Os resultados das Tabelas 1-6 apresentam o desvio relativo percentual e o número de vezes em que o algoritmo proposto, HSDR, obteve soluções iguais, piores e melhores do que os algoritmos comparados, com base em 100 simulações, considerando as duas inicializações: aleatória e pela função *CMDSCALE*. De acordo com a definição do desvio relativo em (17), o método proposto obtém uma solução melhor quando o desvio relativo é positivo. Na primeira coluna das tabelas encontra-se o número de observações de cada conjunto de dados. Para cada caso, foram simulados 100 diferentes conjuntos de dados ou instâncias.

A comparação do desempenho é apresentada seguindo dois critérios. No primeiro critério, são apresentadas estatísticas sobre o desvio relativo, compreendendo três medidas: Desvio Relativo Médio, Desvio Relativo Máximo e Desvio Relativo Mínimo. No segundo critério são apresentadas as frequências em que o algoritmo HSDR obteve soluções do tipo Igual, Melhor ou Pior do que os algoritmos utilizados na comparação (*SMACOF*, *SuperMDS* e Sammon). Uma solução é considerada como Igual quando o módulo do desvio relativo for menor do que 10^{-4}, Melhor quando o desvio relativo for maior do que 10^{-4} e Pior quando desvio relativo for menor do que -10^{-4}. Além disso, em todos os casos, utilizou-se duas estratégias independentes para gerar os pontos iniciais: Ponto inicial aleatório e ponto inicial por CMDSCALE.

6.1.1.1 Comparação 1 - SMACOF e HSDR

Nessa comparação, o algoritmo SMACOF (LEEUW, MAIR, 2009) executa a minimização da formulação (3), enquanto o algoritmo HSDR considera a minimização da formulação (11). A Tabela 6.1 mostra uma comparação do desempenho entre os algoritmos SMACOF e HSDR para resolver problemas de redução de dimensionalidade para o espaço com duas dimensões, ou seja, R^2.

Pelo primeiro critério, desvio relativo, o algoritmo proposto, HSDR, claramente obteve melhores resultados do que o algoritmo SMACOF. Pela equação (17), desvio relativo positivo indica a favor do algoritmo proposto. Na Tabela 6.1 a coluna Desvio Relativo Médio, representados em percentual, apresenta para todos os casos e para ambas as estratégias de inicialização, todos os valores maiores do que zero. Essa ocorrência mostra incontestavelmente que o desempenho médio do HSDR foi melhor

do que o algoritmo SMACOF. Além disso, outra medida que mostra a superioridade da proposta HSDR é a ordem de grandeza dos valores nas colunas Desvio Relativo Máximo e Desvio Relativo Mínimo, sendo os valores apresentados na coluna Desvio Relativo Máximo muito mais elevados.

Pelo segundo critério, últimas seis colunas da Tabela 6.1, para ambas as estratégias de inicialização, pode ser visto que o HSDR tem um desempenho melhor do que o SMACOF, uma vez que, para todos os casos, a frequência de soluções na coluna Melhor é maior do que a frequência de soluções na coluna Pior para ambas as estratégias de inicialização. Deve-se notar que esse desempenho superior torna-se mais evidente à medida que o número de observações aumenta. Em contraste, para o menor caso, com 10 observações, essa superioridade não é tão significativa, uma vez que ambos os algoritmos tiveram uma frequência expressiva de resultados na coluna Igual: frequência de 33 usando a estratégia de inicialização aleatória e frequência de 81 com a estratégia de inicialização por CMDSCALE.

Tabela 6.1: Desvio Relativo percentual em R^2 e o número de vezes que o algoritmo proposto, HSDR, obteve soluções Igual, Pior e Melhor do que o Smacof com base em 100 simulações considerando a iniciação aleatória e a inicialização por CMDSCALE.

SMACOF x HSDR em R^2												
Número	Desvio relativo (%)						Resultado Comparativo					
De	Inicialização aleatória			CMDSCALE			Inicialização aleatória			CMDSCALE		
Observações	Médio	Máximo	Mínimo	Médio	Máximo	Mínimo	Igual	Pior	Melhor	Igual	Pior	Melhor
10	8,48	188,76	-29,52	0,57	18,56	-0,62	33	12	55	81	4	15
20	8,97	53,72	-8,81	0,79	9,14	-3,10	3	11	86	42	4	54
30	4,74	34,83	-13,43	0,71	6,68	-0,58	2	14	84	15	9	76
40	5,05	36,73	-5,81	0,63	5,94	-2,35	0	11	89	16	5	79
50	3,79	23,31	-3,85	0,42	2,51	-0,29	1	12	87	6	6	88
60	2,76	29,14	-6,33	0,37	2,35	-0,28	1	14	85	5	3	92
70	2,87	23,90	-3,08	0,34	1,88	-0,85	1	16	83	6	3	91
80	1,81	13,95	-5,09	0,26	1,72	-0,71	0	18	82	2	7	91
90	1,80	12,02	-6,43	0,30	2,24	-0,55	1	16	83	11	4	85
100	1,82	9,69	-3,81	0,26	1,87	-0,66	0	27	73	3	7	90

A Figura 6.1 apresenta o Boxplot dos desvios relativos para os diferentes casos. Através do Boxplot pode-se ver a dispersão do desvio relativo. Deve ser lembrado que o

desvio relativo positivo indica um melhor desempenho do HSDR. A linha vermelha representa o desvio relativo igual a zero, sendo o ponto onde os algoritmos produzem resultados iguais.

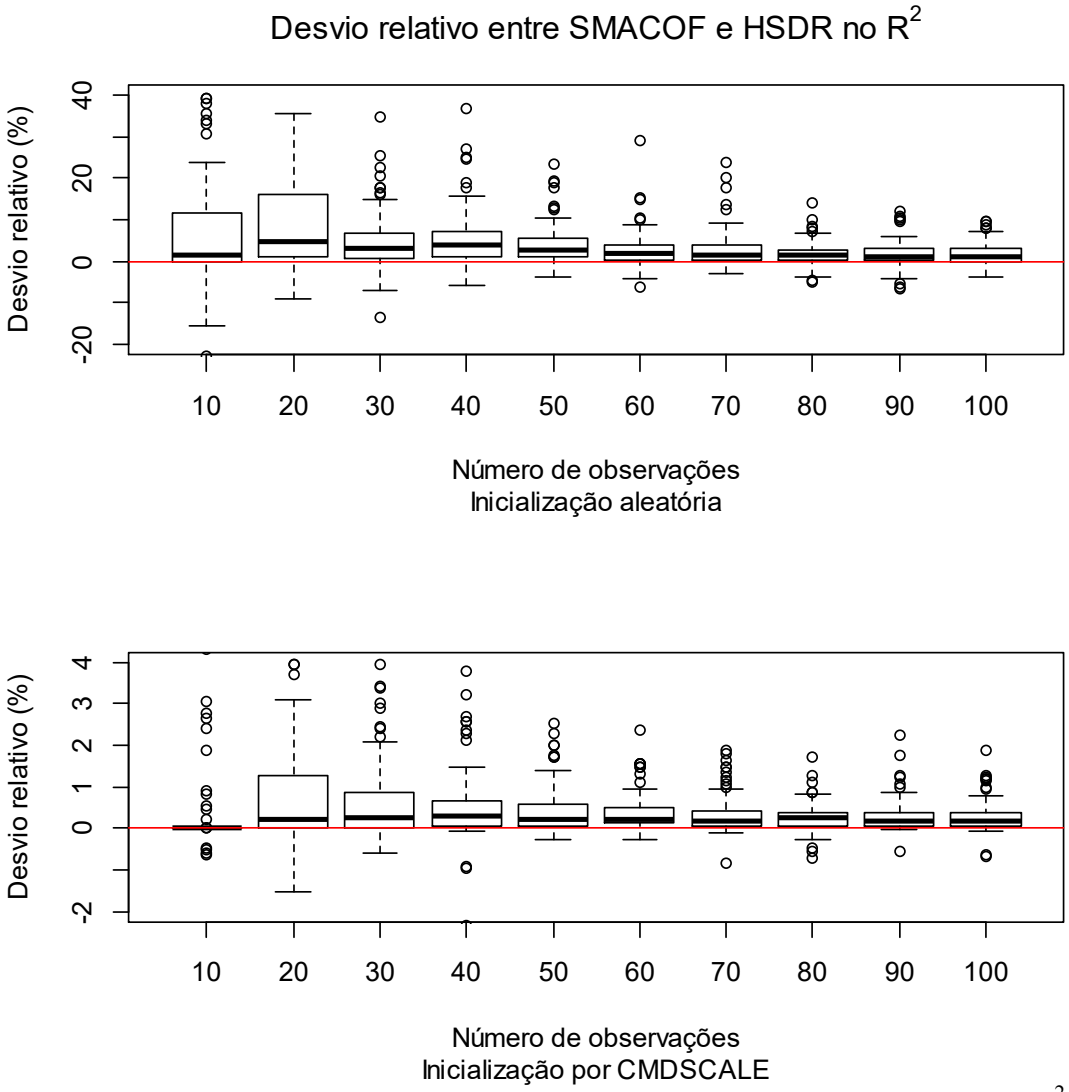

Figura 6.1: Boxplot do Desvio Relativo (%) entre SMACOF e o HSMDS, no R^2 com inicialização aleatória e inicialização por CMDSCALE, para casos de 10 até 100 observações, utilizando como base 100 simulações.

Para a estratégia de inicialização aleatória nos casos de 20 a 90 observações, o percentil 25 está acima da linha vermelha significando um melhor desempenho do HSDR em pelo menos 75% dos casos. Para a estratégia de inicialização por *CMDSCALE,* o percentil 25 está acima da linha vermelha em todos os casos com mais de 20 observações, mostrando a superioridade do algoritmo HSDR.

Comparando as estratégias de inicialização, a inicialização com *CMDSCALE*, parte inferior da Figura 6.1, gera melhores pontos iniciais e produz valores de desvio relativo com uma menor variabilidade em comparação com a estratégia de ponto de início aleatório, parte superior da Figura 6.1.

A Tabela 6.2 mostra uma comparação do desempenho entre os algoritmos SMACOF e HSMDS para resolver problemas de redução de dimensionalidade para o espaço com três dimensões, ou seja, R^3. Nesta tabela as colunas têm o mesmo significado daquelas da Tabela 6.1. Em quase todos os casos, exceto no caso com 10 observações e com inicialização por CDMSCALE, os erros médios são maiores do que zero, o que significa um desempenho médio superior do HSDR. A frequência na coluna Melhor é sempre maior do que na coluna Pior, mostrando uma clara superioridade do HSDR.

Tabela 6.2: Desvio relativo percentual no R^3 e o número de vezes que o algoritmo proposto, HSDR, obteve soluções Igual, Pior e Melhor do que o *Smacof* com base em 100 simulações considerando a iniciação aleatória e inicialização por CMDSCALE.

SMACOF x HSDR no R^3												
Número	Desvio relativo (%)						Resultado Comparativo					
De	Inicialização aleatória			CMDSCALE			Inicialização aleatória			CMDSCALE		
Observações	Médio	Máximo	Mínimo	Médio	Máximo	Mínimo	Igual	Pior	Melhor	Igual	Pior	Melhor
10	4,34	82,88	-28,24	-0,05	2,91	-5,84	65	9	26	93	3	4
20	1,77	26,54	-9,06	0,37	6,60	-1,08	55	11	34	68	4	28
30	1,83	20,10	-7,32	0,38	3,39	-0,99	29	19	52	64	3	33
40	1,76	13,40	-6,10	0,23	4,90	-1,04	16	20	64	42	13	45
50	1,50	19,46	-6,36	0,37	4,72	-1,61	19	15	66	41	8	51
60	1,14	12,76	-3,51	0,11	3,11	-1,88	23	16	61	56	9	35
70	1,21	12,14	-3,31	0,16	1,41	-0,10	14	15	71	41	5	54
80	0,93	9,17	-3,27	0,17	3,52	-0,83	19	15	66	43	6	51
90	0,93	9,80	-4,22	0,14	1,99	-1,09	19	15	66	38	5	57
100	0,62	4,55	-3,15	0,18	2,04	-0,53	23	11	66	36	4	60

A Figura 6.2 apresenta o Boxplot dos desvios entre SMACOF e HSDR no R^3. Pode-se ver que para quase todos os casos que o desempenho de HSDR foi melhor ou igual em pelo menos 75% das vezes. Há apenas um caso de exceção: 10 observações e

inicialização por CMDSCALE, entretanto a dispersão do desvio relativo é muito pequena para este caso.

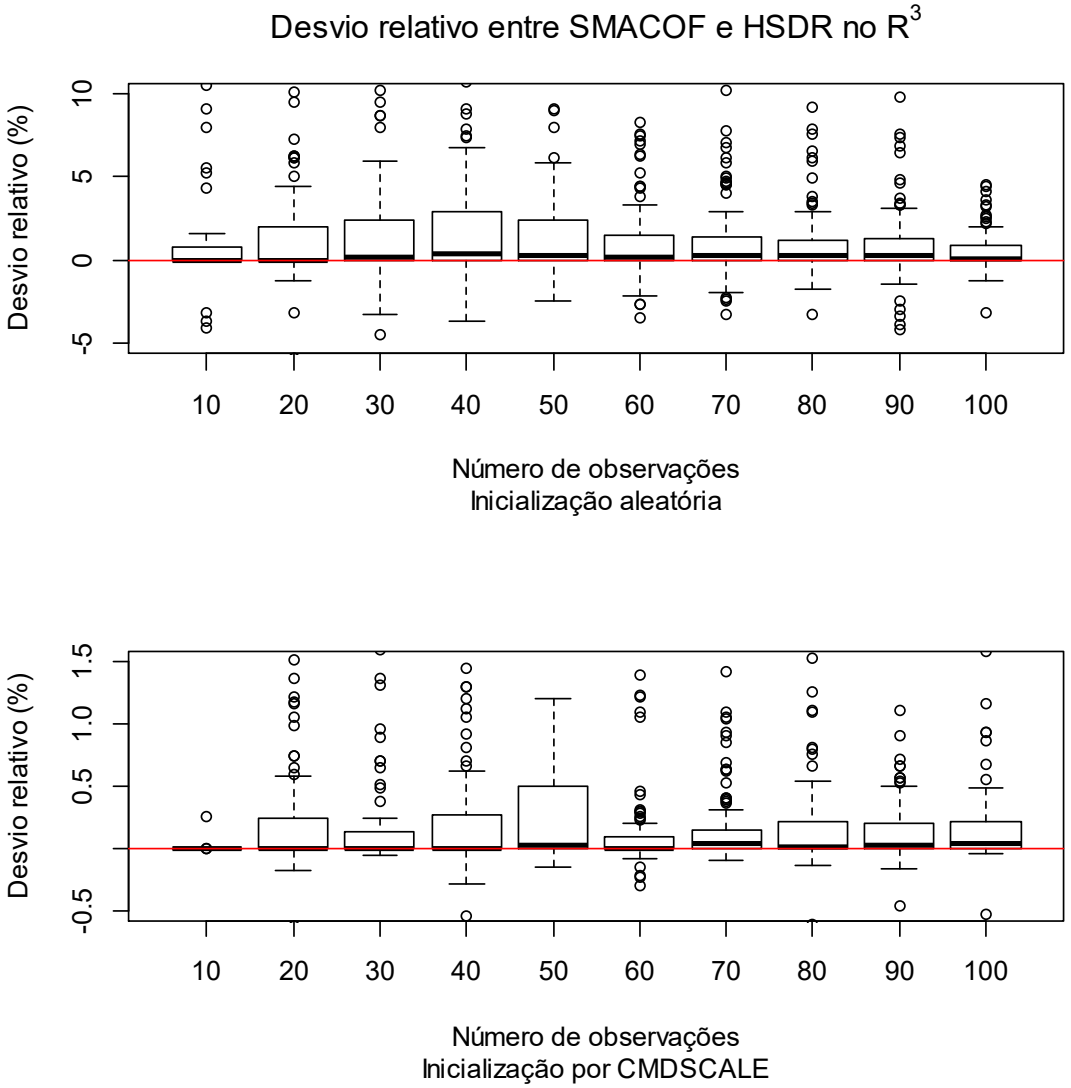

Figura 6.2: Boxplot do Desvio relativo (%) entre SMACOF e o HSMDS, no R^2 com inicialização aleatória e inicialização por CMDSCALE, para os casos de 10 até 100 observações, utilizando como base 100 simulações.

6.1.1.2 Comparação 2 - SuperMDS e HSDR

Nesta comparação, o algoritmo SuperMDS (WITTEN, TIBSHIRANI, 2011) efetua a minimização da formulação (7), enquanto o algoritmo HSDR considera a minimização de formulação (14), utilizando $\alpha = 0$, para ambos os algoritmos.

A Tabela 6.3 mostra uma comparação de desempenho entre superMDS e algoritmo HSDR para resolver problemas de teste para a redução no R^2. Segundo o primeiro critério de comparação, os Desvios relativos médios em todos os casos, são maiores que zero, principalmente para casos pequenos, o que significa que o desempenho de HSDR foi melhor.

Tabela 6.3: Desvio Relativo percentual em R^2 e o número de vezes que o algoritmo proposto, HSDR, obteve soluções Igual, Pior e Melhor do que o superMDS com base em 100 simulações considerando a iniciação aleatória e inicialização por CMDSCALE.

superMDS X HSDR no R^2												
Número	Desvio relativo (%)						Resultado Comparativo					
De	Inicialização aleatória			CMDSCALE			Inicialização aleatória			CMDSCALE		
Observações	Médio	Máximo	Mínimo	Médio	Máximo	Mínimo	Igual	Pior	Melhor	Igual	Pior	Melhor
10	10,10	396,02	-30,05	0,49	18,56	-6,67	26	16	58	82	5	13
20	7,93	37,05	-7,77	0,95	11,81	-3,10	6	5	89	38	4	58
30	5,20	34,05	-10,99	0,75	6,68	-0,45	2	12	86	19	6	75
40	4,20	19,32	-7,68	0,67	5,15	-0,96	0	9	91	15	4	81
50	2,98	19,08	-9,56	0,46	2,73	-0,08	0	16	84	4	5	91
60	2,55	16,14	-6,21	0,37	1,85	-0,28	0	14	86	4	3	93
70	2,97	27,63	-10,28	0,35	1,82	-0,83	1	17	82	6	5	89
80	1,99	9,88	-4,90	0,27	1,85	-0,72	1	17	82	3	9	88
90	1,70	10,81	-5,74	0,32	2,25	-0,55	0	22	78	6	5	89
100	1,89	10,43	-4,50	0,26	1,81	-0,66	0	18	82	2	7	91

De acordo com o segundo critério de comparação, a Tabela 6.3 mostra que o algoritmo HSDR em relação ao *SuperMDS* obteve uma maior frequência de soluções na coluna Melhor do que na coluna Pior, para ambas as estratégias de inicialização, mostrando um desempenho superior do HSDR.

A figura 6.3 apresenta o Boxplot dos desvios entre SuperMDS e HSDR para a redução para o R^2. Nessa figura, para a estratégia de inicialização aleatória nos casos com 20 ou mais observações, pode ser visto que o percentil 25 está acima da linha vermelha significando um melhor desempenho do HSDR em pelo menos 75% dos casos. Sobre a estratégia inicialização por CMDSCALE, o percentil 25 está acima da linha vermelha em todos os casos com 30 ou mais observações, mostrando a

superioridade do algoritmo HSDR.

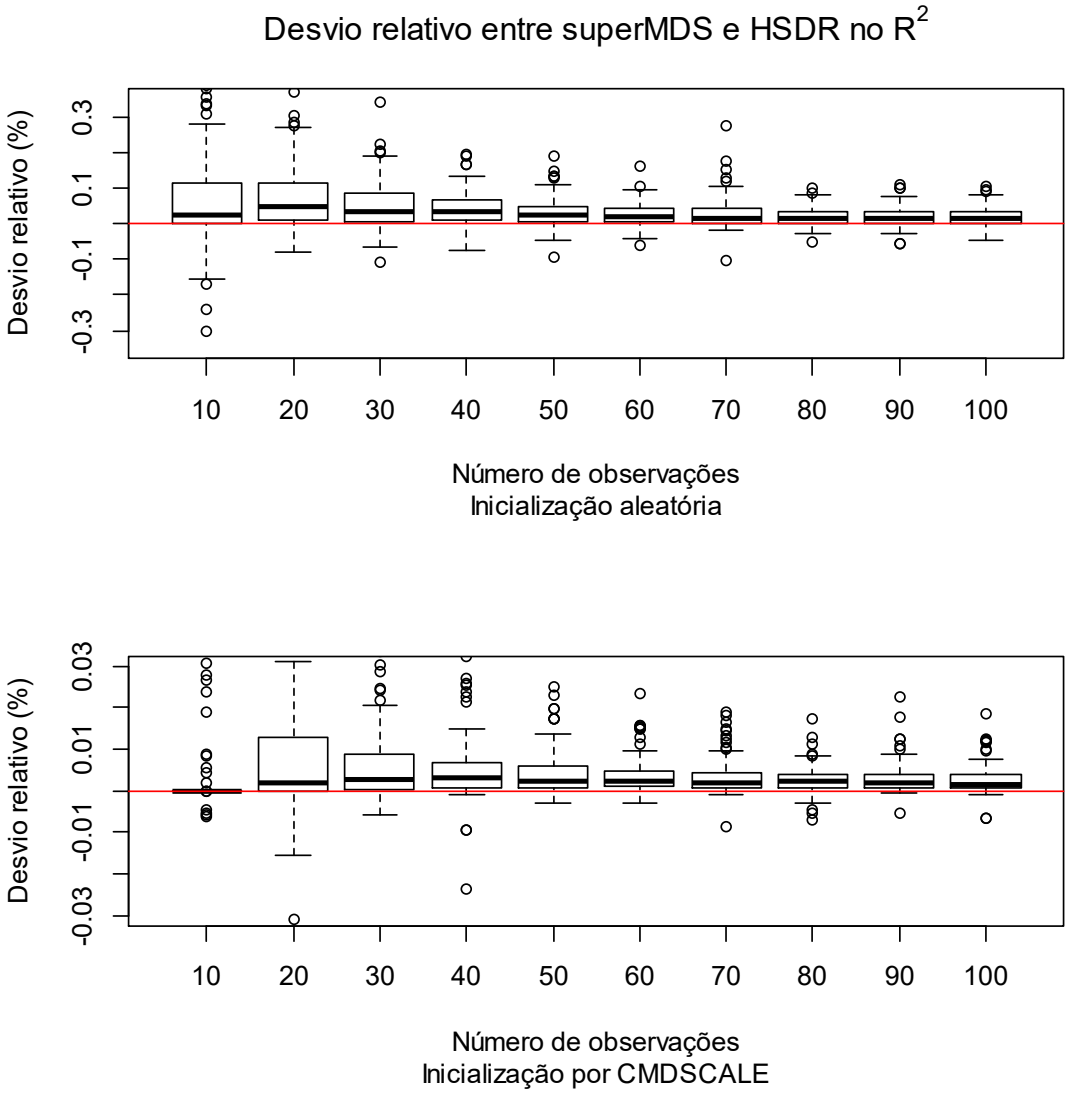

Figura 6.3: Boxplot do Desvio relativo (%) entre HSDR e o superMDS, no R^2 com inicialização aleatória e inicialização por CMDSCALE, para os casos de 10 até 100 observações, utilizando como base 100 simulações.

A Tabela 6.4 mostra uma comparação de desempenho entre superMDS e algoritmo HSDR para resolver problemas de teste no R^3. Em quase todos os casos, exceto com 10 observações e inicialização por CMDSCALE, os Desvios relativos médios são maiores que zero, o que significa que o desempenho de HSDR foi melhor.

A Figura 6.4 apresenta o Boxplot do desvio entre SupervMDS e HSDR no R^3. Pode-se ver que o desempenho de HSDR foi melhor ou igual em pelo menos 75% das

vezes. O único caso em que não há uma superioridade categórica, caso com 10 observações, a dispersão do erro relativo é muito pequena.

Tabela 6.4: Desvio Relativo percentual em R^3 e o número de vezes que o algoritmo proposto, HSDR, obteve soluções igual, pior e melhor do que o superMDS com base em 100 simulações considerando a iniciação aleatória e inicialização por CMDSCALE.

superMDS x HSDR no R^3												
Número	Desvio relativo (%)						Resultado Comparativo					
De	Inicialização aleatória			CMDSCALE			Inicialização aleatória			CMDSCALE		
Observações	Médio	Máximo	Mínimo	Médio	Máximo	Mínimo	Igual	Pior	Melhor	Igual	Pior	Melhor
10	3,82	72,04	-41,33	-0,05	2,91	-5,84	66	10	24	93	3	4
20	1,87	26,55	-9,42	0,33	6,60	-1,08	52	10	38	70	4	26
30	1,62	18,13	-7,30	0,47	9,16	-0,98	27	21	52	63	2	35
40	1,97	18,10	-3,28	0,26	4,90	-1,04	14	15	71	43	13	44
50	1,42	19,46	-2,00	0,38	4,73	-1,61	19	12	69	37	8	55
60	0,92	12,76	-4,91	0,11	3,11	-1,88	24	19	57	54	9	37
70	0,93	10,19	-2,84	0,19	1,41	-0,10	11	18	71	33	7	60
80	1,20	9,98	-2,05	0,19	3,58	-0,83	13	12	75	38	7	55
90	0,99	8,66	-4,22	0,16	1,99	-1,09	14	13	73	35	5	60
100	0,73	5,20	-3,79	0,19	2,04	-0,03	12	9	79	34	2	64

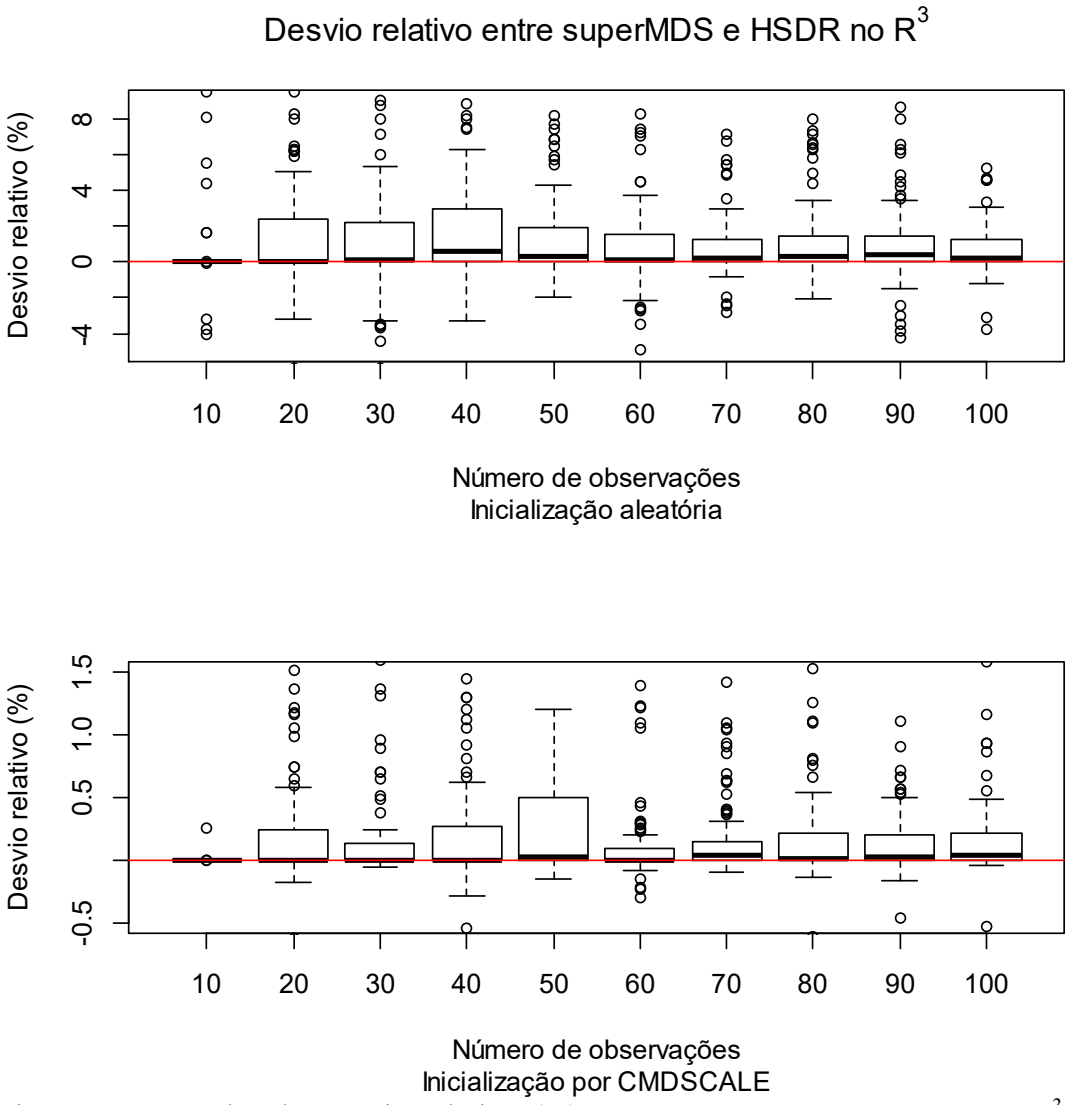

Figura 6.4: Boxplot do Desvio relativo (%) entre HSDR e o superMDS, no R^3 com inicialização aleatória e inicialização por CMDSCALE, para os casos de 10 até 100 observações, utilizando como base 100 simulações.

6.1.1.3 Comparação 3 - Sammon e HSDR

Nesta seção, o algoritmo de Sammon (VENABLES, RIPLEY, 2013) efetua a minimização da formulação (3), enquanto o algoritmo HSDR considera a minimização da formulação (11). De uma maneira análoga às análises anteriores, as Tabelas 6.5 e 6.6 apresentam uma comparação entre o desempenho do Sammon e do HSDR, para resolver problemas no R^2 e R^3.

Em relação ao primeiro critério de comparação, as Tabelas 6.5 e 6.6 mostram valores do desvio relativo médio superiores a zero para todos os casos. Essa ocorrência mostra que o algoritmo proposto HSDR obteve um desempenho médio melhor do que o algoritmo de Sammon. Ao contrário dos problemas anteriores, o erro médio aumenta na medida em que o número de observações aumenta.

Na Tabela 6.5, onde o espaço é R^2, HSDR tem um desempenho muito superior ao de Sammon. Os desvios relativos médios em ambas as estratégias de inicialização mostram uma consistente superioridade do algoritmo HSDR. Pode-se ver que o desvio relativo médio tem tendência crescente assumindo valor máximo para o caso com 100 observações, valor de 131,06% com inicialização aleatória e o valor de 57,62% com a inicialização por CMDSCALE. A comparação com base em frequências observadas, últimas colunas da Tabela 6.5, leva a conclusões análogas às anteriores.

Tabela 6.5: Desvio Relativo percentual em R^2 e o número de vezes que o algoritmo proposto, HSDR, obteve soluções Igual, Pior e Melhor do que o Sammon com base em 100 simulações considerando a iniciação aleatória e inicialização por CMDSCALE.

Sammon X HSDR em R^2												
Número	Desvio relativo (%)						Resultado Comparativo					
De	Inicialização aleatória			CMDSCALE			Inicialização aleatória			CMDSCALE		
observações	Médio	Máximo	Mínimo	Médio	Máximo	Mínimo	Igual	Pior	Melhor	Igual	Pior	Melhor
10	14,35	568,25	-43,80	6,12	278,07	-4,28	28	15	57	70	9	21
20	32,21	577,40	-10,86	7,84	153,43	-3,88	3	14	83	20	8	72
30	56,94	583,09	-7,50	19,07	152,04	-0,65	0	15	85	5	4	91
40	57,18	454,25	-5,96	25,34	174,43	-0,64	0	8	92	3	3	94
50	66,57	447,34	-8,05	21,65	155,92	-0,30	0	11	89	3	5	92
60	88,30	379,79	-5,98	34,18	161,83	-0,21	0	16	84	3	3	94
70	114,84	412,95	-6,20	31,75	164,85	-0,47	0	8	92	1	2	97
80	104,86	423,59	-1,37	47,58	157,75	-0,62	0	5	95	2	2	96
90	111,04	411,10	-4,10	40,68	158,58	-0,37	0	12	88	1	5	94
100	131,06	378,10	-3,31	57,62	161,97	-0,06	0	4	96	0	1	99

Em relação ao segundo critério de comparação, a Tabela 6.5 mostra os valores da coluna Melhor expressivamente maiores que os valores da coluna Pior, para ambas

as estratégias de inicialização. Essas duas ocorrências mostram de uma forma consistente um desempenho significativamente superior do HSDR sobre Sammon

A Figura 6.5 apresenta o Boxplot do desvio entre SAMMON e HSDR para a redução para o R^3. Nessa figura, pode-se ver que o desempenho do HSDR foi melhor ou igual em pelo menos 75% das vezes para todos os casos com mais do que 10 observações para ambas as alternativas de inicialização. A dispersão do desvio relativo para o caso com 10 observações é muito pequena.

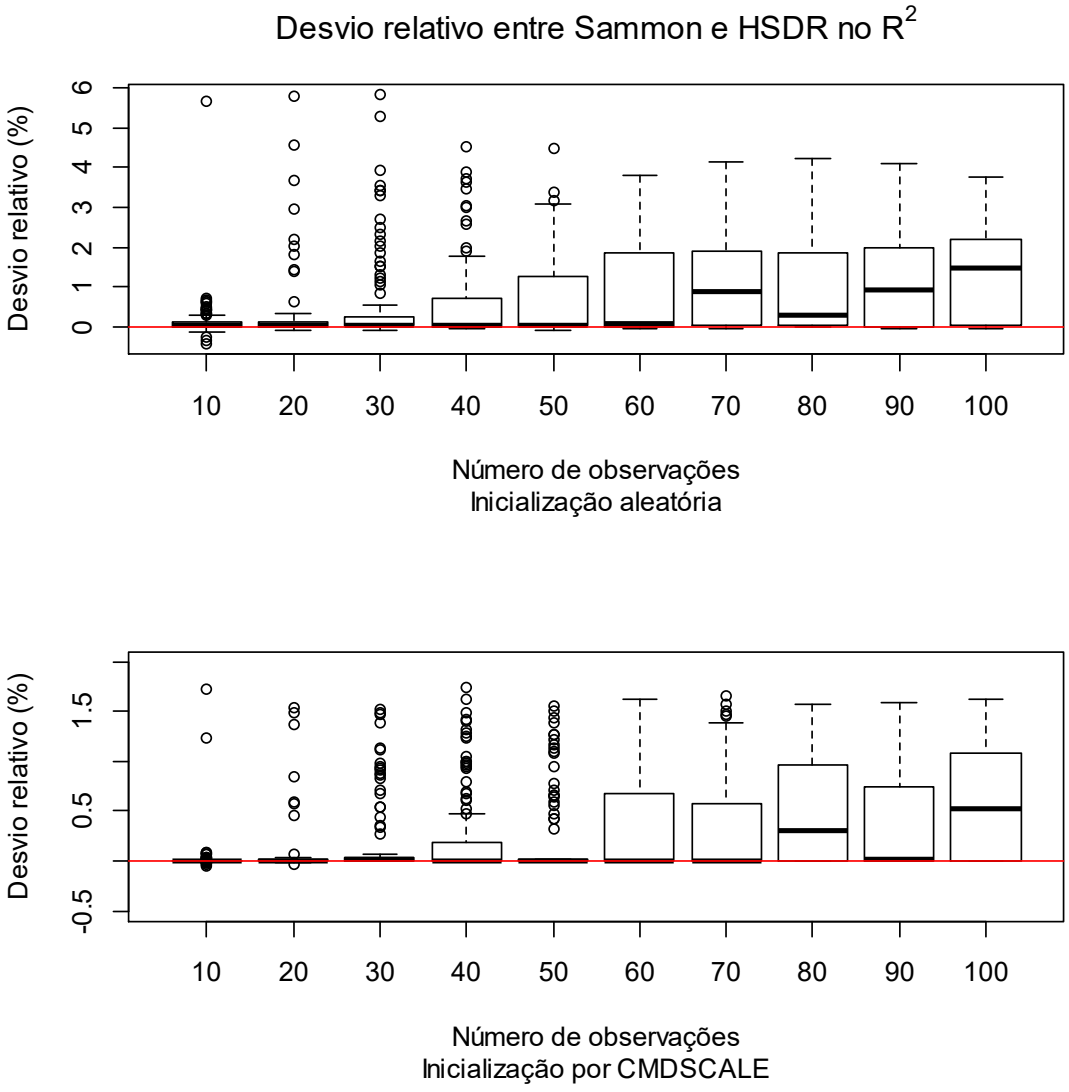

Figura 6.5: Boxplot do Desvio relativo (%) entre Sammon e o HSMDS, no R^2 com inicialização aleatória e inicialização por CMDSCALE, para os casos de 10 até 100 observações, utilizando como base 100 simulações.

A Tabela 6.6, na qual o espaço reduzido é o R^3, mostra resultados semelhantes aos apresentados na Tabela 6.5, levando assim a conclusões análogas referentes à superioridade do HSDR. De forma geral, destacam-se altos erros relativos médios e em particular para os casos com maior número de observações, 278,68% com a estratégia de inicialização aleatória para o caso com 90 observações e 103,19% com a inicialização por CMDSCALE para o caso com 80 observações.

Tabela 6.6: Erro Relativo percentual no R^3 e o número de vezes que o algoritmo proposto, HSDR, obteve soluções Igual, Pior e Melhor do que o Sammon com base em 100 simulações considerando a iniciação aleatória e inicialização por CMDSCALE.

Sammon x HSDR no R^3												
Número	Desvio relativo (%)						Resultado Comparativo					
De	Inicialização aleatória			CMDSCALE			Inicialização aleatória			CMDSCALE		
observações	Médio	Máximo	Mínimo	Médio	Máximo	Mínimo	Igual	Pior	Melhor	Igual	Pior	Melhor
10	122,73	2631,09	-28,10	7,98	301,24	-4,45	47	10	43	83	2	15
20	56,66	923,89	-17,55	16,63	266,42	-4,12	33	12	55	56	6	38
30	121,72	995,61	-12,76	36,16	252,67	-1,66	13	12	75	45	5	50
40	148,59	774,02	-3,14	36,27	220,20	-2,19	12	11	77	27	7	66
50	186,41	764,26	-2,57	55,44	223,79	-0,96	10	12	78	19	6	75
60	237,23	760,31	-8,90	46,40	215,98	-2,65	6	11	83	25	6	69
70	229,03	671,77	0,00	57,29	221,93	-1,81	7	0	93	11	6	83
80	273,55	625,05	-0,73	103,19	220,45	-0,17	6	4	90	11	1	88
90	278,68	643,13	-3,45	93,94	234,49	-1,23	3	8	89	9	2	89
100	236,11	651,96	-2,14	99,89	219,37	-0,03	3	11	86	11	1	88

A Figura 6.6 apresenta o Boxplot do desvio relativo entre Sammon e HSDR no R^3. Nessa figura pode-se ver que o desempenho de HSDR foi melhor ou igual em pelo menos 75% das vezes para os casos com mais do que 20 observações e inicialização aleatória e para os casos com mais do que observações 40 observações e inicialização por CMDSCALE

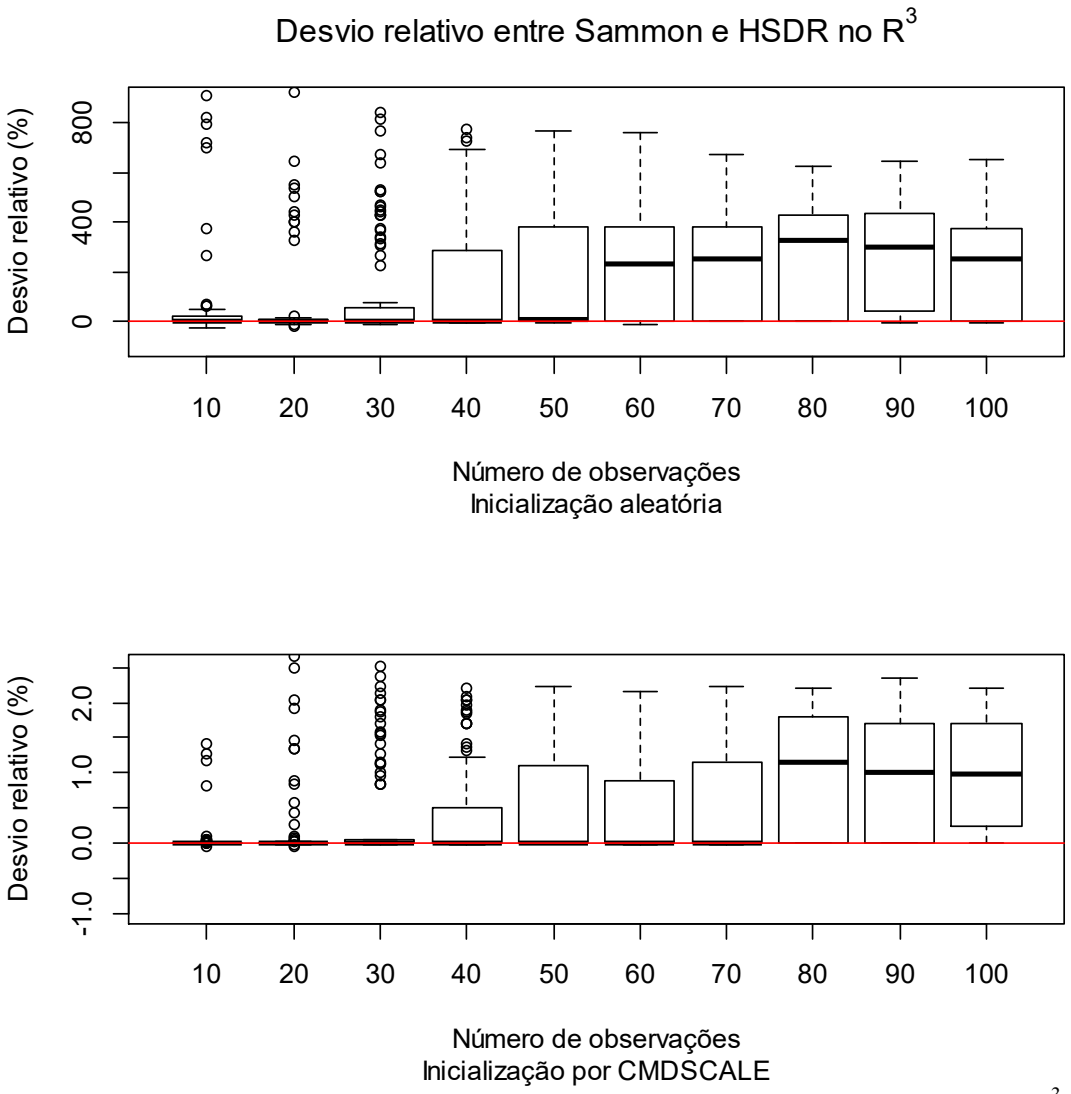

Figura 6.6: Boxplot do Desvio relativo (%) entre HSDR e o Sammon, no R^3 com inicialização aleatória e inicialização por CMDSCALE, para os casos de 10 até 100 observações, utilizando como base 100 simulações.

6.1.2. Experimentos com conjuntos de dados reais

Adicionalmente aos experimentos com dados sintéticos anteriormente descritos, realizou-se outro conjunto de experimentos de redução de dimensionalidade usando um conjunto de dados reais, para avaliar o desempenho do algoritmo proposto HSDR através da comparação direta do algoritmo proposto com outros algoritmos consagrados pela literatura. Para tal fim, foram selecionados sete conjuntos de dados publicamente disponíveis no UCI Machine Learning Repository. Todos os sete conjuntos de dados

selecionados estão originalmente associados à tarefa de classificação supervisionada, a saber relacionados:

O conjunto de dados *Lung Cancer Data Set* (*lung-cancer.data*) contém dados sobre três tipos de cânceres de pulmão, com 32 observações e 56 atributos. Somente as observações que não possuem valores ausentes forma utilizadas, totalizando 27 observações completas.

O conjunto de dados *SPECTF Heart Data Set* (*SPECTF.train*) é composto por imagens de *Single Proton Emission Computed Tomography (SPECT*) contendo 79 observações, correspondentes a imagens de doentes pre-processadas com 44 atributos. Cada um dos doentes é classificado em duas categorias: normal e anormal.

O conjunto de dados *Statlog Vehicle Silhouettes* (*xaa.dat*) é composto por um conjunto de características extraídas da silhueta de veículos produzidas a partir de imagens de quatro tipos de veículos: OPEL, SAAB, BUS e VAN. Há um total de 94 observações, cada uma com 18 atributos.

O conjunto de dados *Urban Land Cover* (*training.csv*) é composto por imagens aéreas de alta resolução contendo 168 observações, cada uma com 147 atributos, classificadas em nove tipos de cobertura do solo urbano: árvores, grama, solo, concreto, asfalto, prédios, carros, piscinas e sombras.

O conjunto de *Wine Data Set* (*wine.data)* contém dados provenientes de análises químicas de vinhos produzidos em uma mesma região da Itália, mas derivados de três espécies de uva diferentes. A análise mensura as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos. Há um total de 177 observações cada uma com 13 atributos.

O conjunto de dados *Parkinsons Data Set* (*parkinsons.data*) é composto por uma série de medições de voz de 31 pessoas, sendo 23 com a doença de Parkinson. Cada atributo é uma medida de voz particular e cada linha corresponde uma das 195 gravações de voz a partir desses indivíduos.

O conjunto de dados *Connectionist Bench Sonar, Mines vs. Rocks* (*sonar.all-data*) contém dados correspondentes a sinais de sonar, aplicados em vários ângulos e

sob várias condições. Contém 111 observações provenientes da aplicação do sonar em cilindro de metal simulando uma mina terrestre e 97 observações provenientes da aplicação do sonar sobre rochas. Cada observação contém 60 atributos, com valores no intervalo de 0,0 a 1,0.

A Tabela 6.7 mostra uma síntese dos conjuntos de dados adotados nos experimentos. Em todas as tabelas adotou-se nomes reduzidos referentes aos conjuntos de dados. A síntese apresentada na tabela segue a mesma ordem dos conjuntos de dados apresentada no texto.

Conjunto de Dados	Número de observações	Número de atributos
Lung	27	56
SPECTF	79	44
Vehicle	94	18
Trainurban	168	147
Wine	177	13
Parkison	195	22
Sonar	208	60

Tabela 6.7: Síntese dos conjuntos de dados reais

Previamente, foi realizado um pré-processamento para eliminar registros com valores ausentes e desconsiderar as informações do rótulo da classe. Em seguida, foi realizada uma transformação de escala para cinco conjuntos de dados: Parkinson, Urban Land Cover, Statlog, Wine e SPECTF Heart Data Set. A transformação adotada faz com que os valores de cada atributo tenha média zero e variância unitária (para cada variável foi subtraída a sua média e dividida pelo seu desvio padrão).

Novamente, foram realizadas comparações entre o algoritmo proposto (HSDR) usando a formulação (10), o algoritmo SMACOF (LEEUW, MAIR, 2009) e o algoritmo superMDS (WITTEN, TIBSHIRANI, 2011) com parâmetro $\alpha = 0$, aplicados na resolução da formulação *Least squares MDS* (LSMDS) (2). Do mesmo modo, o algoritmo proposto usando a formulação (11) foi comparado com o algoritmo de Sammon (VENABLES, RIPLEY, 2013) aplicado à resolução da formulação Sammon (3). Em todos os experimentos procedeu-se como na seção anterior, para cada conjunto

de dados gerou-se 100 pontos iniciais aleatórios. A escolha do ponto inicial influencia na solução final, desta forma para se ter uma comparação justa e eliminar a variabilidade das solução, decorrentes da escolha do ponto inicial, todos os algoritmos foram inicializados com os mesmos pontos iniciais.

Tabela 6.8 e Tabela 6.9 mostram os resultados obtidos, na redução de dimensionalidade, respectivamente para 2-D e 3-D. Nestas tabelas, a primeira coluna faz referência ao conjunto de dados reais, nas quatro subsequentes são apresentados resultados para a formulação LSMDS e nas três últimas colunas são apresentados resultados para a formulação Sammon. Na tabela pode-se encontrar uma quantidade nomeada "Fmin". "Fmin" é, por definição, o melhor resultado obtido executando 100 vezes cada algoritmo comparado. Na segunda coluna, temos Fmin LSMDS registrando o melhor resultado entre os três algoritmos: SMACOF, HSDR formulação (10) e o superMDS. O número de vezes em que cada algoritmo obteve o resultado igual à Fmin ou equivalente igual à Fmin é mostrado nas próximas três colunas. Considera-se um resultado equivalentemente igual à Fmin se o desvio relativo à Fmin for menor do que 10^{-4}. A coluna Fmin SAMMON mostra o melhor resultado obtido executando 100 vezes cada um dos dois algoritmos, Sammon e HSDR formulação (11), para a formulação de Sammon.

Tabela 6.8: Resultados para os dados reais em 2-D com base em 100 inicializações, para cada um dos três algoritmos. Fmin LSMDS representa o valor da função objetivo da melhor solução encontrada, considerando os três algoritmos SMACOF, HSDR formulação (10) e superMDS. Fmin Sammon representa o valor da função objetivo da melhor solução encontrada, considerando Sammon e HSDR formulação (11).

Conjunto de Dados	Fmin LSMDS	Frequência observada de Fmin			Fmin SAMMON	Frequência observada de Fmin	
		SMACOF	HSDR LSMDS	superMDS		SAMMON	HSDR SAMMON
Lung	1288,81	3	8	3	216,684	0	21
SPECTF	18149,58	0	40	0	8244,082	0	48
Vehicle	3373,51	1	43	1	817,507	0	40
Trainurban	270304,30	0	27	1	17429,239	0	39
Wine	20541,73	1	4	0	11244,000	0	4
Parkison	20427,59	1	0	0	56716,875	0	1
Sonar	4551,46	0	24	0	6036,338	0	32

Na Tabela 6.8, por exemplo, considerando o Fmin LSMDS para o conjunto de dados SPECTF, o algoritmo proposto obteve um resultado equivalente ao mínimo em 40% das realizações do experimento, enquanto os outros métodos não obtiveram nenhuma vez um valor equivalente ao mínimo. A frequência de Fmin associada ao algoritmo proposto, HSDR, é significativamente superior aos outros dois algoritmos em quase todos os conjuntos de dados, com exceção do conjunto Parkson.

Na redução de dimensionalidade para 3-D, o algoritmo HSDR obteve valores equivalentes a Fmin em quase todos os conjuntos de dados para a formulação LSMDS. Além disso, é o único algoritmo que obteve Fmin no conjunto de dados de Sonar. No entanto, HSDR se apresentou pior do que os outros métodos para o conjunto de dados Parkison. Para a formulação Sammon Mapping, o algoritmo HSDR mais uma vez obteve valores Fmin em todos os conjuntos de dados, apresentando claramente um desempenho melhor.

Tabela 6.9: Resultados para os dados reais em 3-D com base em 100 inicializações, para cada um dos três algoritmos. Fmin LSMDS representa o valor da função objetivo da melhor solução encontrada, considerando os três algoritmos SMACOF, HSDR e superMDS. Fmin Sammon representa o valor da função objetivo da melhor solução encontrada, considerando Sammon e HSDR.

Conjunto de Dados	Fmin LSMDS	Frequência observada de Fmin			Fmin SAMMON	Frequência observada de Fmin	
		SMACOF	HSDR LSMDS	superMDS		SAMMON	HSDR SAMMON
Lung	647,63	3	23	4	434,310	0	7
SPECTF	9401,02	25	100	9	5637,281	0	78
Vehicle	1239,65	19	6	14	1246,245	0	11
Trainurban	113542,89	16	36	16	26217,809	0	31
Wine	8437,63	0	3	2	6080,301	0	16
Parkison	8153,19	13	0	12	6145,160	0	71
Sonar	2046,33	0	7	0	3756,102	0	4

6.1.1.4 SuperMDS x HSDR

Neste conjunto de comparações a seguir apresentados, o algoritmo SuperMDS (WITTEN, TIBSHIRANI, 2011) efetua a minimização da formulação (7), enquanto o algoritmo HSDR considera a minimização de formulação (14), utilizando os valores de 0,25, 0,5 e 0,75 para o parâmetro α para ambos os algoritmos. Novamente, em todos os experimentos foram gerados 100 pontos iniciais aleatórios e cada ponto inicial foi usado

nos dois algoritmos. Deve ser notado que na seção anterior foram feitas comparações entre esses algoritmos considerando o caso particular com o parâmetro α assumindo o valor zero, caso este equivalente a formulação Least Square MDS, formulação (2).

No método SupervisedMDS o número de classes é suposto ser igual a dois. Dessa forma, para os conjuntos de dados com mais de duas classes foram selecionadas somente duas classes, conforme explicado a seguir caso a caso. Os conjuntos de dados SPECTF, Parkson, Sonar foram utilizados da mesma forma como apresentada na seção anterior. Para o conjunto de dados Lung Classe, a classe que contém 8 observações foi definida como a classe 1 e a classe que contém 10 observações como a classe 2. Para o conjunto de dados Trainurban Urban Land Cover a classe *concrete* foi definida como sendo a classe 1, com 23 observações, e classe *grass* como a classe 2, com 29 observações. Para o conjunto de dados *wine* a classe que contém 58 observações foi definida como a classe 1 e a classe que contém 71 observações como a classe 2. Para o conjunto de dados *vehicle* a classe *bus* foi definida como a classe 1 com 26 observações e classe *opel* com 20 observações como a classe 2.

Tabela 6.10 e Tabela 6.11 mostram os resultados obtidos, na redução de dimensionalidade, respectivamente para 2-D e 3-D, pelos algoritmos de WITTEN e TIBSHIRANI (2011) e por HSDR usando a formulação (14). Para cada valor do parâmetro α uma comparação do desempenho é apresentada seguindo os critérios de desvio relativo e de frequência com que o algoritmo HSDR obteve soluções do tipo Igual, Pior e Melhor.

Nas duas tabelas, em todos os conjuntos de dados e para todos os três valores do parâmetro α os desvios relativos médio são maiores do que zero, mostrando uma clara superioridade do HSDR. Além disso, a frequência na coluna Melhor é sempre maior do que na coluna Pior, confirmando essa superioridade.

Tabela 6.10: Desvio Relativo percentual em R^2 e o número de vezes que o algoritmo proposto, HSDR, obteve soluções do tipo Igual, Pior e Melhor do que o superMDS com base em 100 simulações considerando a iniciação aleatória.

superMDS X HSDR em R^2							
Conjunto de Dados	Alpha	Desvio Relativo (%)			Frequencia Observada (%)		
		Médio	Máximo	Mínimo	Igual	Pior	Melhor
Lung	0,25	1,22	11,37	-7,84	0	19	81
	0,50	1,03	4,00	-0,62	0	8	92
	0,75	0,19	0,78	-0,43	2	21	77
SPECTF	0,25	1,62	5,71	-1,61	0	9	91
	0,50	0,86	2,28	-0,16	0	1	99
	0,75	0,23	0,42	0,08	0	0	100
Vehicle	0,25	0,96	28,73	-1,68	4	8	88
	0,50	0,30	11,54	-0,67	4	9	87
	0,75	0,22	1,19	-0,61	1	11	88
Trainurban	0,25	1,24	11,22	-7,24	0	25	75
	0,50	0,89	3,50	-1,46	1	12	87
	0,75	0,20	0,68	-0,22	0	10	90
Wine	0,25	1,33	6,66	-1,93	0	11	89
	0,50	0,75	2,84	-0,54	0	6	94
	0,75	0,22	0,43	0,07	0	0	100
Parkison	0,25	1,78	7,97	-3,76	0	15	85
	0,50	1,08	3,40	-0,54	1	5	94
	0,75	0,41	1,02	0,03	0	0	100
Sonar	0,25	3,20	17,85	-12,95	0	6	94
	0,50	1,74	7,16	0,03	0	0	100
	0,75	0,62	1,39	0,04	0	0	100

Tabela 6.11: Desvio Relativo percentual em R^3 e o número de vezes que o algoritmo proposto, HSDR, obteve soluções do tipo Igual, Pior e Melhor do que o superMDS com base em 100 simulações considerando a iniciação aleatória.

superMDS X HSDR em R^3							
Conjunto de Dados	Alpha	Desvio Relativo (%)			Frequencia Observada (%)		
		Médio	Máximo	Mínimo	Igual	Pior	Melhor
Lung	0,25	1,07	7,01	-2,61	0	12	88
	0,50	0,54	2,87	-0,37	0	8	92
	0,75	0,26	0,79	-0,09	0	1	99
SPECTF	0,25	1,38	4,55	-0,80	0	2	98
	0,50	0,82	2,03	0,20	0	0	100
	0,75	0,41	0,51	0,23	0	0	100
Vehicle	0,25	0,20	5,38	0,02	0	0	100
	0,50	0,24	1,37	0,03	0	0	100
	0,75	0,14	0,45	0,01	1	0	99
Trainurban	0,25	0,96	4,67	-2,78	0	12	88
	0,50	0,60	2,38	-0,50	1	4	95
	0,75	0,32	0,69	0,10	0	0	100
Wine	0,25	0,97	3,52	-0,06	0	1	99
	0,50	0,67	2,06	0,16	0	0	100
	0,75	0,47	0,71	0,11	0	0	100
Parkison	0,25	1,02	5,07	-1,17	0	6	94
	0,50	0,60	1,99	-0,15	0	2	98
	0,75	0,49	1,46	0,10	0	0	100
Sonar	0,25	0,29	0,69	0,08	0	0	100
	0,50	0,59	1,55	-0,29	0	1	99
	0,75	0,22	1,39	0,10	0	0	100

Tabela 6.12 e Tabela 6.13 mostram outro enfoque na comparação dos algoritmos SuperMDS e HSDR formulação 14 na redução de dimensionalidade, respectivamente em 2-D e 3-D. Nestas tabelas, a primeira coluna mostra o conjunto de dados e a segunda os três diferentes valores para o parâmetro α. A seguir, apresenta-se novamente para cada algoritmo a Frequência Observada da melhor solução (Fmin) e a melhor solução encontrada pelos dois algoritmos. Em quase todos os casos, o único algoritmo que obteve uma solução equivalente a Fmin foi o Supervised HS.

Tabela 6.12: Resultados para os dados reais em 2-D com base em 100 inicializações, para cada um dos três algoritmos. Fmin representa o valor da função objetivo da melhor solução encontrada, considerando os dois algoritmos superMDS e HSDR

superMDS X HSDR em R^2				
Conjunto de Dados	Alpha	Frequencia Observada da melhor solução (Fmin)		Fmin
		Supervised	Supervised HS	
Lung	0,25	0	7	717,20
	0,50	0	4	705,22
	0,75	0	10	478,05
SPECTF	0,25	0	1	25204,26
	0,50	0	1	23347,15
	0,75	0	4	14928,65
Vehicle	0,25	0	32	4196,23
	0,50	1	68	4892,93
	0,75	1	36	3705,15
Trainurban	0,25	0	15	26320,14
	0,50	0	9	26096,28
	0,75	0	7	17422,39
Wine	0,25	0	1	15242,65
	0,50	0	1	15198,10
	0,75	0	2	10153,02
Parkison	0,25	0	1	34189,14
	0,50	0	2	35389,09
	0,75	0	1	26433,82
Sonar	0,25	0	61	9698,42
	0,50	0	76	9937,02
	0,75	0	62	7049,96

Tabela 6.13: Resultados para os dados reais em 3-D com base em 100 inicializações, para cada um dos três algoritmos. Fmin representa o valor da função objetivo da melhor solução encontrada, considerando os dois algoritmos Supervised.e SupervisedHS

superMDS X HSDR em R^3				
Conjunto de Dados	Alpha	Frequencia Observada da melhor solução (Fmin)		Fmin
		Supervised	Supervised HS	
Lung	0,25	0	11	579,84
	0,50	0	7	656,81
	0,75	0	11	471,38
SPECTF	0,25	0	3	21429,04
	0,50	0	3	22203,59
	0,75	0	13	14809,32
Vehicle	0,25	0	100	3835,39
	0,50	0	99	4787,41
	0,75	0	70	3698,79
Trainurban	0,25	0	38	22580,84
	0,50	0	17	25075,48
	0,75	0	42	17331,14
Wine	0,25	0	18	12791,57
	0,50	0	35	14491,18
	0,75	0	35	10086,40
Parkison	0,25	0	22	28251,50
	0,50	0	32	33007,19
	0,75	0	67	26142,49
Sonar	0,25	0	100	8285,04
	0,50	0	95	9528,67
	0,75	0	30	7006,90

6.2 Comparação com resultados da literatura

Nesta seção, trata-se de apresentar e de discutir um conjunto de experimentos computacionais com o objetivo de se ter uma validação adicional do desempenho da metodologia proposta, através da comparação dos resultados produzidos pelo algoritmo HSDR com resultados publicados na literatura.

6.2.1. Resultados para Classificação e *Ranking* Bipartido

Witten e Tibshirani (2011) apresentam duas tabelas de resultados computacionais obtidos em dois experimentos independentes para dois objetivos distintos: resolução do problema de classificação supervisionada e resolução do problema de *ranking* bipartido. Foram realizados experimentos, seguindo os mesmos critérios descritos no artigo e as comparações de resultados. A fim de detalhar os experimentos, apresenta-se, a seguir, os três diferentes modelos utilizados para gerar os dados sintéticos:

1- Modelo *constant*, assim chamado porque há uma média constante para as observações em cada classe.

$$\mathbf{z}_i \sim \begin{cases} N(\mathbf{0.4}, \mathbf{I}_s) & \textit{se a observação } i \textit{ for da classe } 1 \\ N(-\mathbf{0.4}, \mathbf{I}_s) & \textit{se a observação } i \textit{ for da classe } 2 \end{cases}$$

$\mathbf{I}_{sxs}$, matriz de covariância tem a forma da matriz identidade, e vetor de média com todas as componentes iguais.

No exemplo da Figura 3.1 os dados foram gerados por este modelo e projetados para o R^2.

2- Modelo *two-sided*, assim chamado porque as observações da classe dois formam dois grupos distintos.

$$\mathbf{z}_i \sim \begin{cases} N(\mathbf{0}, \mathbf{I}_s) & \textit{se a observação } i \textit{ for da classe } 1 \\ N(-1, \mathbf{I}_s) \textit{ ou } N(1, \mathbf{I}_s) & \textit{com igual probabilidade se a observação } i \textit{ for da classe } 2 \end{cases}$$

3- Modelo *linear*, assim chamado porque há uma tendência linear nas observações, como uma função do índice da observação

$$\mathbf{z}_i \sim \left\{ N\left(\frac{3i}{n}, \mathbf{I}_s\right) \right.$$

As primeiras $n/2$ observações são da classe um e as seguinte da classe dois. A observação i tem vetor de média com todas as componentes iguais a $\frac{3i}{n}$. Novamente, n representa o número de observações, $\mathbf{z}_i$, $i = 1, \cdots, n$, cada uma com S atributos.

A figura 6.7 ilustra um exemplo do efeito produzido pelo parâmetro α, no conjunto de dados gerado pelo modelo *two-sided*, com $S = 15$ e $n = 100$. Cada classe tem 50 observações. Na medida em que o parâmetro α aumenta as observações de classes distintas tendem a se afastar, aumentando a separação das classes. No caso em que α é igual a zero existe uma nítida mistura das duas classes.

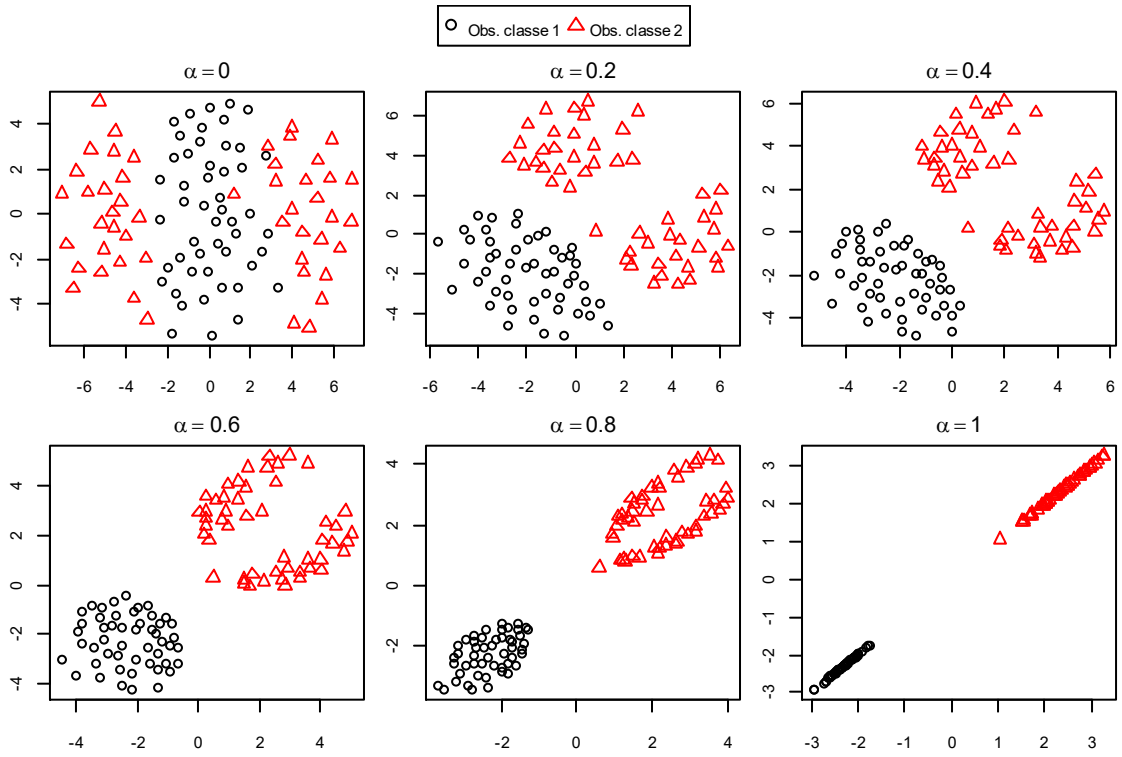

Figura 6.7: *Supervised mds* aplicado no modelo *two-sided*, $\mathbf{R}^{15} \rightarrow \mathbf{R}^2$

Witten e Tibshirani, usando as bases de dados sintéticas conforme descritas, apresentam resultados para os seguintes métodos: *L1-penalized logistic regression,* SMDS com $\alpha = 1$, SMDS com validação cruzada, método de Cox e Ferry (1993) (CF) , *LDA, k-nearest neighbors* (K-NN), support vector machine (SVM) com *linear kernel* (LK) e SVM com *polynomial kernel* (PK) *of degree* três. Os resultados foram gerados com os conjuntos de treinamento e de teste de tamanhos iguais. Todos os métodos

foram treinados com o conjunto de treinamento e o desempenho avaliado com o conjunto de teste. Todos os parâmetros de ajuste foram selecionados utilizando validação cruzada. Os erros de classificação e do *ranking* bipartido são calculados no conjunto de teste, com base em 50 simulações de conjuntos de treinamento e de conjuntos de teste. Os conjuntos de teste foram gerados com o mesmo modelo e com o mesmo número de observações do conjunto de treinamento.

6.2.1.1 *Supervised MDS* para classificação

Na seção 3.2 foi apresentada a regra geral de classificação para o modelo *Supervised MDS* onde o ponto de corte com valor zero é substituído por um valor fixo, v, representando assim um ponto de corte genérico:

$$\begin{cases} se \min\limits_{\mathbf{x}_{n+1}\in\mathbb{R}^d} h_1(D_{n+1},\mathbf{x}_{n+1}) - \min\limits_{\mathbf{x}_{n+1}\in\mathbb{R}^d} h_2(D_{n+1},\mathbf{x}_{n+1}) < v \quad \rightarrow \hat{y}_{n+1} = 1 \\ caso\ contrário\ \hat{y}_{n+1} = 2 \end{cases}$$

Na comparação dos resultados, utilizou-se a mesma regra de classificação adotada no artigo de WITTEN e TIBSHIRANI (2011), onde o ponto de corte é definido pelo valor v dado por:

$\frac{\sum_{i=1}^{n} 1_{y_i=1}}{n}$ quantil de $\left\{\min\limits_{\mathbf{x}_j\in\mathbb{R}^d} h_1(D_j,\mathbf{x}_j) - \min\limits_{\mathbf{x}_j\in\mathbb{R}^d} h_2(D_j,\mathbf{x}_j)\right\}_{j=1}^{n}$, ou seja, ordena-se todas as diferenças das funções do conjunto de teste, e a seguir os valores abaixo do quantil $\frac{\sum_{i=1}^{n} 1_{y_i=1}}{n}$ são classificados na classe um e os valores acima na classe dois.

A Tabela 6.14 apresenta resultados comparativos com os apresentados na Tabela1 de WITTEN e TIBSHIRANI (2011). Os resultados obtidos pela metodologia proposta correspondem às colunas HSSMDS $\alpha = 1$ e HSSMDS CV. As demais colunas são exatamente iguais àquelas apresentadas por Witten e Tibshirani (2011). Nas primeiras três colunas da Tabela 6.14 tem-se o modelo de geração da base de dados sintéticos utilizado, a dimensão do espaço e o número de observações. Nas quatro subsequentes tem-se o erro de classificação para o método Supervised MDS. Nas

demais colunas tem-se o erro de classificação para os métodos *L1-penalized logistic regression,* CF, *LDA, KNN, SVM* com *linear kernel* (LK) e SVM com *polynomial kernel of degree 3* (PK).

Os resultados para o método *Supervised MDS* estão divididos em dois blocos. No primeiro bloco tem-se resultados com o parâmetro $\alpha = 1$, para SMDS e a versão suavizada HSSMDS. No segundo bloco tem-se resultados obtidos por validação cruzada, para SMDS e a versão suavizada HSSMDS. Dentro de cada bloco é destacado em negrito o melhor resultado obtido.

Os resultados obtidos nos dois blocos são relevantes, pois o método proposto obteve um erro médio menor em 10 casos dentre os 12 casos apresentados.

Tabela 6.14: Considerando o número de observações n e o número de atributos S, são apresentados os erros médios de classificação obtidos em 50 simulações pelos seguintes métodos: SMDS $\alpha = 1$, HSSMDS $\alpha = 1$, *L1-penalized logistic regression,* o método de CF, *LDA, KNN, SVM* com *linear kernel* (LK) e SVM com *polynomial kernel of degree 3* (PK). Os três modelos utilizados foram: *two-sided, linear e constant.*

Modelo	S	n	SMDS $\alpha = 1$	HSSMDS $\alpha = 1$	SMDS CV	HSSMDS CV	L_1 logistic	CF	LDA	K-NN	SVM LK	SVM PK
Two-sided	5	20	0,264	**0,2460**	0,295	**0,2800**	0,502	0,505	0,501	0,336	0,501	0,422
Two-sided	5	50	0,2288	**0,2088**	0,2432	**0,2328**	0,5012	0,4976	0,5044	0,2904	0,4996	0,4388
Two-sided	15	20	0,089	**0,0580**	0,115	**0,1060**	0,512	0,482	0,489	0,164	0,487	0,467
Two-sided	15	50	0,0664	**0,0480**	**0,0664**	0,0736	0,5016	0,5096	0,522	0,0996	0,51	0,4224
Linear	5	20	0,131	**0,1300**	0,142	**0,1220**	0,212	0,134	0,176	0,183	0,151	0,24
Linear	5	50	0,1324	**0,1232**	0,13	**0,1184**	0,1984	0,1252	0,1336	0,154	0,1388	0,19
Linear	15	20	0,115	**0,0700**	0,107	**0,0540**	0,164	0,145	0,275	0,142	0,133	0,253
Linear	15	50	**0,0752**	0,0824	0,0744	**0,0728**	0,11	0,0796	0,1248	0,086	0,0796	0,2564
Constant	5	20	0,289	**0,2580**	0,264	**0,2120**	0,312	0,239	0,254	0,314	0,257	0,318
Constant	5	50	**0,2172**	0,2216	**0,2024**	0,2048	0,2812	0,1996	0,206	0,2536	0,204	0,29
Constant	15	20	0,249	**0,1980**	0,144	**0,0820**	0,218	0,2	0,261	0,165	0,124	0,278
Constant	15	50	0,1472	**0,1160**	0,0996	**0,0688**	0,1236	0,0972	0,112	0,124	0,0932	0,238

6.2.1.2 *Supervised MDS* para *Ranking* Bipartido

De modo análogo ao apresentado para o problema de classificação são apresentados, a seguir, resultados para o problema de *Ranking* Bipartido. A Tabela 6.15 apresenta resultados comparativos com os apresentados na Tabela 2 do artigo (WITTEN, TIBSHIRANI, 2011). Novamente os resultados obtidos pela metodologia proposta correspondem às colunas HSSMDS $\alpha = 1$ e HSSMDS CV, as demais colunas

são exatamente iguais àqueles apresentados por WITTEN e TIBSHIRANI (2011). Nas primeiras três colunas especifica-se o modelo utilizado, a dimensão do espaço e o número de observações. Nas quatro subsequentes, o erro do *ranking* bipartido para o método Supervised MDS. Nas demais colunas apresenta-se o erro do *ranking* bipartido para os métodos *L1-penalized logistic regression,* CF, *LDA, KNN, SVM* com *linear kernel* (LK) e SVM com *polynomial kernel of degree três* (PK).

Tabela 6.15: Para o número de observações n e número de atributos s, o erro do *ranking* bipartido dos métodos SMDS $\alpha = 1$, HSSMDS $\alpha = 1$, *L1-penalized logistic regression,* o método de CF, *SVM* com *linear kernel* (LK), SVM com *polynomial kernel of degree três* (PK) e *LDA.* Os três modelos utilizados foram: *two-sided, linear e constant* . Os resultados são médias de 50 simulações.

Modelo	S	n	SMDS $\alpha = 1$	HSSMDS $\alpha = 1$	SMDS CV	HSSMDS CV	L_1 logistic	CF	SVM LK	SVM PK	LDA
Two-sided	5	20	0,2454	**0,1778**	0,2788	**0,2084**	0,502	0,5252	0,4341	0,4214	0,4956
Two-sided	5	50	0,193	**0,1534**	0,2069	**0,179072**	0,5012	0,4968	0,4478	0,4423	0,5004
Two-sided	15	20	0,1162	**0,0128**	0,1462	**0,0386**	0,512	0,4882	0,4145	0,4179	0,4836
Two-sided	15	50	0,0646	**0,0116**	0,0646	**0,025696**	0,5016	0,503	0,4525	0,4444	0,5164
Linear	5	20	**0,058**	0,0588	0,0736	**0,045**	0,212	0,2756	0,1084	0,1121	0,1032
Linear	5	50	0,0565	**0,0499**	0,0538	**0,048544**	0,1984	0,2823	0,0563	0,0635	0,0594
Linear	15	20	0,0296	**0,0182**	0,027	**0,0136**	0,164	0,495	0,117	0,136	0,2118
Linear	15	50	0,0173	**0,0172**	0,0148	**0,014656**	0,11	0,4604	0,0169	0,0433	0,0484
Constant	5	20	0,203	**0,1868**	0,1904	**0,1574**	0,312	0,2576	0,2028	0,2134	0,1586
Constant	5	50	0,1593	**0,1267**	0,1432	**0,110471**	0,2812	0,2912	0,1207	0,1633	0,1232
Constant	15	20	**0,1306**	0,1354	**0,048**	0,0496	0,218	0,374	0,135	0,1088	0,1943
Constant	15	50	0,0669	**0,0600**	0,0456	**0,02864**	0,1236	0,4206	0,0271	0,1029	0,0416

Os resultados para o método *Supervised MDS* estão divididos em dois blocos. No primeiro apresenta-se resultados com o parâmetro $\alpha = 1$, para SMDS e a versão suavizada HSSMDS. No segundo apresenta-se resultados obtidos por validação cruzada, para SMDS e a versão suavizada HSSMDS. Dentro de cada bloco destaca-se em negrito o melhor resultado obtido.

Comparando dentro de cada bloco, pode-se ver no primeiro bloco, método *Supervised MDS* com $\alpha = 1$, o método proposto obteve um valor do erro médio do *ranking* bipartido menor em 10 dentre os 12 casos. No segundo bloco, método *Supervised MDS* com validação cruzada, o método proposto obteve o erro médio do *ranking* bipartido menor em 11 dentre os 12 casos.

6.2.2. Resultados do Artigo de CHEN e BUJA (2009)

No artigo de CHEN e BUJA (2009), o critério local continuity ou LC meta-criterion (CHEN, BUJA, 2006) foi utilizado para comparar projeções do método Local MDS (CHEN, BUJA, 2009) com outros métodos. Além disso o critério local continuity foi utilizado para comparar projeções do método Local MDS com diferentes valores dos parâmetros t e k , veja expressão (4) . O critério utiliza a relação de vizinhança nos espaços de alta dimensão e de baixa dimensão, tendo como idéia central a interseção dos K-NN de uma observação no espaço de alta dimensão e os K-NN no espaço de baixa dimensão.

O critério *local continuity* ou LC *meta-criterion* é fundamentado nas três definições:

- $N_{k'}^{S}(i) = \{j_1, \ldots, j_{k'}\}$, conjunto dos k' vizinhos mais próximos da observação i com respeito às observações no espaço de alta dimensão.

- $N_{k'}^{d}(i) = \{k_1, \ldots, k_{k'}\}$, conjunto dos k' vizinhos mais próximos da observação i com respeito às observações no espaço de baixa dimensão.

- $N_{k'}(i) = \left| N_{k'}^{S}(i) \cap N_{k'}^{d}(i) \right|$, cardinalidade da interseção dos dois conjuntos anteriores.

A cardinalidade $N_{k'}(i)$ é uma clara medida de sobreposição pontua para uma observação genérica i. Desta forma, uma medida de sobreposição global é dada por:

$$N_{k'} = \frac{1}{n} \sum_{i=1}^{n} N_{k'}(i)$$

Adota-se a padronização do valor $N_{k'}$ de modo a permitir a comparação entre diferentes valores de k':

$M_{k'} = \frac{1}{k'} N_{k'}$,desta forma, $M_{k'}$ assume valores no intervalo $[0,1]$.

Com o objetivo de obter uma configuração inicial estável no método Local MDS, CHEN e BUJA (2009) argumentam que uma boa estratégia para otimização consiste na inicialização do parâmetro τ, da expressão (5), com um valor grande, como τ = 1, e sucessivamente reduzir o parâmetro até um valor baixo, como 0,01, sempre usando a última configuração das observações no espaço reduzido na inicialização seguinte. Ao longo de todas as configurações obtidas escolhe-se uma ideal em algum

sentido específico, que pode ser medido através de uma função. Para isto Lisha Chen e Andreas Buja utilizaram o critério *LC meta-criterion.*

CHEN e BUJA (2009) utilizaram dois conjuntos de dados em seus experimentos computacionais: *Sculpture Face* e *Frey Faces.*

O conjunto de dados *Sculpture Face* (TENENBAUM *et al.*, 2000), inclui 698 imagens de uma escultura de um rosto humano. As imagens são de uma mesma escultura, variando três condições: perfil da esquerda para direita, de cima para baixo e a direção de iluminação, parâmetro angular que assume três valores. Cada imagem com 64X64 pixels, portanto a dimensão do espaço da imagem é 4096.

A Tabela 6.16 apresenta valores do critério LC medidos, com parâmetro k' com valor 6. O resultado obtido pela metodologia proposta corresponde a colunas $HSLMDS_{K=6,}$ as demais colunas são exatamente iguais àqueles apresentados por CHEN e BUJA (2009). Da segunda até a quarta coluna tem-se os métodos PCA, MDS, Isomap e *local linear embedding* (LLE) (ROWEIS, SAUL, 2000). Nas duas últimas colunas tem-se os resultado obtidos pelo método Local MDS (CHEN, BUJA, 2009), coluna $LMDS_{K=6}$, e o resultado obtidos pela metodologia proposta, Local MDS suavizado, coluna $HSLMDS_{K=6,}$ ambos os métodos com parâmetro k com valor 6. Os valores $N_{k`=6} = 5,216332$ e $M_{k`=6} = 0,8693887$ foram obtidos pelo método $HSLMDS_{K=6}$ com parâmetro $t = 0,0003929626$.

Tabela 6.16: *Sculpture Face Data: LC meta-criteria* para $k' = 6$

Métodos	PCA	MDS	Isomap	LLE	$LMDS_{K=6}$	$HSLMDS_{K=6}$
$N_{k`=6}$	2,6	3,1	4,5	2,8	5,2	5.2
$M_{k`=6}$	0,43	0,52	0,75	0,47	0,87	0.87

O conjunto de dados *Frey Faces* (ROWEIS, SAUL, 2000), inclui 1965 imagens da face de uma única pessoa cujo nome é Brendan Frey, tomadas em quadros sequenciais a partir da fragmentação de imagens de um vídeo. Cada imagem com 20X28 pixels, portanto a dimensão do espaço da imagem é 560.

A tabela 6.17 apresenta valores do critério LC medidos, com parâmetro k' com valor 6. São apresentados os resultado obtidos pelo método Local MDS, coluna $LMDS_{K=12}$, e o resultado obtidos pela metodologia proposta, Local MDS suavizado, coluna $HSLMDS_{K=6,}$ ambos os métodos com parâmetro k com valor 12. Os valores

$N_{k`=12} = 4,63257$ e $M_{k`=12} = 0,3860475$ foram obtidos pelo método $HSLMDS_{K=12}$ com parâmetro $t = 0,8047611$.

Na Tabela 6.17, são apresentados os resultado obtidos pelo método Local MDS, coluna $LMDS_{K=4}$, e o resultado obtidos pela metodologia proposta, Local MDS suavizado, coluna $HSLMDS_{K=4,}$ ambos os métodos com parâmetro k com valor 4. Os valores $N_{k`=12} = 5,163359$ e $M_{k`=12} = 0,4302799$ foram obtidos pelo método $HSLMDS_{K=4}$ com parâmetro $t = 828,5301$.

Tabela 6.17: *Frey Faces Data: LC meta-criteria* para $k' = 12$

Métodos	PCA	MDS	Isomap	LLE	KPCA	$LMDS_{K=12}$	$HSLMDS_{K=12}$	$LMDS_{K=4}$	$HSLMDS_{K=4}$
$N_{k`=12}$	3,6	4,8	4,2	3,2	3,7	4,6	4,6	5,1	5,1
$M_{k`=12}$	0,3	0,4	0,35	0,27	0,31	0,38	0,38	0,43	0,43

6.3 Conclusões

Nesta tese é proposta uma nova abordagem para os problemas de redução de dimensionalidade, classificação supervisionada e *ranking* bipartido tendo como base a utilização da Suavização Hiperbólica nas formulações (3), (4), (6), (7), (8) e (9).

A minimização das distorções entre a projeção no espaço reduzido e no espaço original dos dados não é uma tarefa trivial, tendo a abordagem metodológica proposta se mostrado eficaz na busca de um mínimo local profundo. Um único algoritmo foi proposto, chamado de *Hyperbolic Smoothing Dimensionality Reduction* (HSDR), para resolver os métodos da classe de MDS: *Sammon mapping*, *Supervised MDS*, *Local MDS* e *Least Absolute Residuals.* A idéia central do algoritmo HSDR é resolver uma sequência de problemas diferenciáveis que gradualmente se aproximam do problema original. Cada um destes problemas é completamente diferenciável permitindo a aplicação dos métodos de otimização sem restrições mais robustos e eficientes.

Para uma comparação direta com o algoritmo proposto, testes computacionais foram feitos com outros algoritmos disponíveis na linguagem R. Além disso, foram apresentados testes para se ter uma comparação com resultados publicados na literatura. O algoritmo HSDR apresentou um desempenho bem superior na maioria das

comparações. Assim, pode-se cogitar que a proposta apresentada nesta tese seja uma boa alternativa frente a outros algoritmos tradicionais da literatura.

Para os problemas de classificação supervisionada e *ranking* bipartido, no método *Supervised MDS* além da suavização das formulações, foi proposto também a utilização de processamento paralelo na redução de dimensionalidade das observações novas nos problemas de classificação supervisionada e *ranking* bipartido.

Em decorrência da complexidade, os métodos não lineares não são adequados para conjuntos de dados com muitas observações, sendo esta uma grande limitação do seu uso em grandes bases de dados. A metodologia apresentada na próxima seção tem como uma das vantagens a habilidade de processar conjuntos de dados com muitas observações.

Capítulo 7

Uma nova metodologia supervisionada para redução de dimensionalidade baseada em protótipos

7.1 Introdução

Nesta seção é apresentada uma nova metodologia para redução de dimensionalidade não linear supervisionada fundamentada em protótipos representativos de cada classe. A metodologia proposta engloba três conceitos principais que se articulam: redução de dimensionalidade, supervisão através da utilização da informação da classe à qual a observação pertence e generalização para observações novas.

A maioria dos métodos de redução de dimensionalidade processa os dados em uma única etapa (*batch process*), de modo que não possuem a capacidade de generalização para uma observação nova. Alguns métodos lineares possuem um modelo ou função explícita para o mapeamento entre os espaços de alta e baixa dimensão, permitindo uma generalização direta para uma nova observação. Todos os seguintes métodos lineares: análise de componentes principais, Análise de Discriminantes Lineares (*LDA*), *Neighborhood Components Analysis*, *Maximally Collapsing Metric Learning*, *Locality Preserving Projection* e *Neighborhood Preserving Embedding* possuem esta propriedade (MAATEN, 2007).

A extensão para o tratamento de observações fora da amostra não é uma operação direta no uso de métodos não lineares. Sherpad Interpolation (SHEPARD, 1968) e Generalized Radial Basis Function Neural Network Function Approximation

(GRBF-FA) (MOODY, DARKEN, 1989) são duas abordagens para a generalização de observações novas. Estas duas abordagens são dependentes de parâmetros livres cujas especificações podem ser não-triviais. Por exemplo, NAGARAJAN *et al* (2015) utiliza estas duas abordagens executado os ajustes dos parâmetros por validação cruzada. Schachter (1978) propõe uma regra de generalização para mapeamentos não lineares, somente válida para o R^2, baseada na divisão do espaço em um *grid*, combinando pesos entre observações do *grid* para o mapeamento.

Considerando o método de Sammon, PEKALSKA et al (1999) propõe uma transformação linear para a generalização do mapeamento para observações novas. Agrafiotis (2001) propõe a utilização de amostras dos dados em mapeamentos não lineares e utiliza redes neurais para a generalização para o conjunto inteiro das observações. Myasnikov (2011) propõem a utilização da estrutura hierárquica gerada por clustering hierárquico em conjunto com uma amostra das observações para a projeção de observações novas no método de Sammon.

NAUD e DUCH (2000) mapeiam uma amostra dos dados e fixam as coordenadas da amostra mapeada no espaço de baixa dimensão. As observações que não fazem parte da amostra são mapeadas por meio de uma modificação na função Stress com duas parcelas. Uma parcela para preservação das distâncias entre a amostra mapeada e as observações que não fazem parte da amostra. A outra parcela considera a preservação das distâncias entre as observações que não fazem parte da amostra. Na formulação de NAUD e DUCH (2000) o número de variáveis no problema é grande sendo proporcional ao número de observações que serão mapeadas e a dimensão do espaço onde as observações são projetadas. Além disso, as observações que serão mapeadas não são independentes.

A exceção do método SMDS, os métodos de escalonamento multidimensional são métodos de redução de dimensionalidade que não utilizam qualquer informação da classe de cada observação, sendo assim são métodos não supervisionados. A metodologia proposta inova ao introduzir o conceito de supervisão em métodos tradicionalmente não supervisionados. A informação da classe da observação é incorporada no modelo através da utilização de protótipos. Por protótipo se entende um elemento representativo de cada grupo, contendo alguma informação geométrica do grupo.

Pode-se adotar diversos critérios para a escolha ou definição dos protótipos. Como exemplo, pode-se citar o centro de gravidade, ou seja, a média, a mediana geométrica, uma observação que esteja no centro do grupo ou uma observação que tenha uma importância particular. O processo supervisionado ocorre no conjunto treinamento, composto pelas observações utilizadas no cálculo dos protótipos. Todas as observações que não foram utilizadas no conjunto de treinamento são chamadas de observações do conjunto de teste, podendo essas possuir ou não a identificação da classe a qual a observação é associada.

Na presente metodologia os métodos de Sammon, *Least Squares MDS* e *Least Absolute Residuals* são utilizados de forma supervisionada. Estes métodos são aplicados nos protótipos, gerando uma configuração de pontos em baixa dimensão contendo uma informação representativa de cada uma das classes nesse espaço reduzido. A seguir, para cada um dos métodos e, com exatamente o mesmo princípio de preservação de distância, propõe-se a generalização para a redução de dimensionalidade de uma observação nova ou do conjunto de treinamento. Esta generalização para uma observação nova é realizada sem a necessidade de quaisquer ajustes de parâmetros.

7.2 Utilização de protótipos nos métodos de Sammon, *Least Squares MDS* e *Least Absolute Residuals*

Para a formalização do emprego dos protótipos nos métodos de Sammon, *Least Squares MDS* e Least Absolute Residuals, suponha que se tenha um conjunto de n observações, $\mathbf{z}_i, i = 1,\cdots,n$, cada uma com S atributos, ou seja, cada observação pertence a um espaço de dimensão S. Cada observação contém uma classe associada com um respectivo rótulo representado por $y_i, i = 1,\cdots,n$, onde y_i assume um dos valores $1,\cdots,k$ referentes à sua classe, desta forma tem-se k diferentes classes. Para cada classe, um protótipo $\overline{\mathbf{z}}_k$ pertencente ao espaço de dimensão S é calculado ou escolhido segundo algum critério de preferência. Denomina-se de conjunto de treinamento, o conjunto de observações, $\mathbf{z}_i, i = 1,\cdots,n$, utilizadas no cálculo dos protótipos.

A matriz das distâncias entre os protótipos é representada por DP_{kxk}. Os protótipos no espaço de baixa dimensão são obtidos através da aplicação direta dos métodos de Sammon ou *Least Squares MDS* ou Least Absolute Residuals e são representados por $\overline{\mathbf{x}}_i$, $i = 1, \cdots, k$.

Desta forma, tomando o método de Sammon suavizado (11) como exemplo aplicado aos protótipos, tem-se a correspondente formulação:

$$Fsammon(\overline{\mathbf{x}}) = \frac{1}{c} \sum_{i<j}^{k} \frac{(DP_{ij} - \theta(\left\| \overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j \right\|_2, \gamma))^2}{DP_{ij}} \qquad (18)$$

Para a generalização do emprego dos protótipos nas formulações de Sammon (3), *Least Squares MDS* (2) (10) e Least Absolute Residuals (6) (13), basta substituir em cada método a matriz de distância D_{nxn} por DP_{kxk}, para se obter o conjunto de protótipos no espaço de baixa dimensão, $\overline{\mathbf{x}}_i$, $i = 1, \cdots, k$.

7.3 Redução de dimensionalidade para observações do conjunto de teste

Dada uma observação nova ou pertencente ao conjunto de teste, observação esta sem a classe de pertinência associada, pertencente ao espaço de dimensão S, deseja-se reduzir a dimensão utilizando o mesmo princípio de preservação de distância aplicado no mapeamento prévio dos protótipos. Para isto definiremos a notação:

O índice $n+1$ faz referência invariavelmente a uma observação nova genérica, sendo assim, a observação nova é representada por $\mathbf{z}_{n+1}$. O vetor de distâncias entre $\mathbf{z}_{n+1}$ e os protótipos $\overline{\mathbf{z}}_k$, ambos no espaço de alta dimensão, é representado por $\mathbf{DZ}_{n+1} = \left[DZ_{n+1,1}, \ldots, DZ_{n+1,k} \right]^T$. Com os valores calculados das distâncias $\mathbf{DZ}_{n+1}$ e dos protótipos no espaço reduzido $\overline{\mathbf{x}}_p$, $p = 1, \cdots, k$, calculados pela formulação (18) propomos a função de Sammon para uma observação nova (*FSnova*).

$$FSnova(\mathbf{x}_{n+1},\gamma)=\frac{1}{c}\sum_{p=1}^{k}\frac{(DZ_{n+1,p}-\theta(\left\|\overline{\mathbf{x}}_p-\mathbf{x}_{n+1}\right\|_2,\gamma))^2}{DZ_{n+1,p}} \qquad (19)$$

As coordenadas de uma observação nova genérica $\mathbf{z}_{n+1}$ no espaço de baixa dimensão, $\mathbf{x}_{n+1}$, são calculadas pela minimização da função suavizada $FSnova(\mathbf{x}_{n+1})$. Nota-se que o mesmo princípio de Sammon foi utilizado no mapeamento dos protótipos e no mapeamento de uma observação nova.

De modo análogo, quando o mapeamento dos protótipos for feito pelo método de Least Square MDS (2) (10), propõe-se uma função suavizada para o mapeamento de uma observação nova tendo como base o mesmo princípio de preservação de distância utilizado nos protótipos.

$$FLSMDSnova(\mathbf{x}_{n+1},\gamma)=\frac{1}{c}\sum_{p=1}^{k}(DZ_{n+1,p}-\theta(\left\|\overline{\mathbf{x}}_p-\mathbf{x}_{n+1}\right\|_2,\gamma))^2 \qquad (20)$$

Quando o mapeamento dos protótipos for feito pelo método *Least Absolute Residuals* (6) (13), tem-se a correspondente função suavizada de mapeamento para uma observação nova:

$$FLARnova(\mathbf{x}_{n+1},\gamma_1,\gamma_2)=\frac{1}{c}\sum_{p=1}^{k}w_p\ \theta_1(DZ_{n+1,p}-\theta_2(\left\|\overline{\mathbf{x}}_p-\mathbf{x}_{n+1}\right\|_2,\gamma_2),\gamma_1) \qquad (21)$$

Nota-se que na formulação (21) há dois procedimentos de suavização. O mais interior para a suavização da distância euclidiana, com parâmetro γ_2, e o mais exterior para suavizar o valor absoluto das diferenças entre as distâncias com parâmetro γ_1.

Em todos os casos (19), (20) e (21), as coordenadas de uma observação nova genérica $\mathbf{z}_{n+1}$ são obtidas no espaço de baixa dimensão através da minimização de cada função em relação a $\mathbf{x}_{n+1}$, onde $\mathbf{x}_{n+1}\in\mathbb{R}^d$.

7.4 Redução de dimensionalidade das observações do conjunto de treinamento

O mesmo procedimento adotado para as observações do conjunto de teste é adotado para a redução de dimensionalidade de cada observação do conjunto de treinamento. Sendo assim, para o conjunto de treinamento, tem-se n problemas de baixa dimensão, com somente d variáveis em cada, onde d é a dimensão do espaço das observações projetadas. O conjunto de treinamento no espaço de baixa dimensão é representado por $\mathbf{x}_i$, $i = 1, \cdots, n$, $\mathbf{x}_i \in R^d$.

É importante enfatizar que a redução de dimensionalidade de cada observação é completamente independente, seja esta observação do conjunto de treinamento ou do conjunto de teste. Além disso, a redução de dimensionalidade de cada observação consiste em um problema de otimização com somente d variáveis. Essas duas características viabilizam a resolução de problemas de redução de dimensionalidade com um número de observações extremamente grande, problemas estes que são absolutamente intratáveis sob ponto de vista numérico pelos métodos de redução de dimensionalidade não lineares congêneres.

Além disso, o processo de redução de dimensionalidade de cada observação pode ser paralelizado de uma forma muito direta e simples pelo fato de cada observação ser tratada independentemente.

7.5 Efeito de convexificação pela Suavização Hiperbólica

Todas as três formulações, (19), (20) e (21), apresentadas para a redução de dimensionalidade de uma observação nova, apresentam em comum a presença da suavização hiperbólica. Sem perda de generalidade essas formulações poderiam ser apresentadas sem a suavização hiperbólica, ou de forma equivalente, suavizada com parâmetro γ fixo em zero.

Entretanto a Suavização Hiperbólica (SH) tem a propriedade de eliminar pequenos mínimos locais viabilizando a obtenção de mínimos locais profundos. Essa ocorrência foi verificada empiricamente na resolução de problemas de recobrimento, *clustering*, *Multisource Fermat-Weber* e *hub location*. No caso particular do problema de geometria de distâncias, a SH tem a extraordinária propriedade de convexificação do problema original, conforme provado analiticamente em XAVIER (2003) e SOUZA *et al* (2011). Nas presentes formulações (19) e (20) a propriedade da Suavização Hiperbólica (SH) de convexificação e por conseguinte de eliminar pequenos mínimos foi verificado empiricamente.

A Figura 7.1 apresenta como exemplo ilustrativo o gráfico da função $FSnova(\mathbf{x}_{n+1}, \gamma)$, considerando um mapeamento para o espaço de dimensão um, para vários valores de $\mathbf{x}_{n+1}$ e com parâmetro γ fixo em zero. Destaca-se em azul o ponto de mínimo global, além disso, pode-se verificar que há claramente dois pontos de mínimos locais.

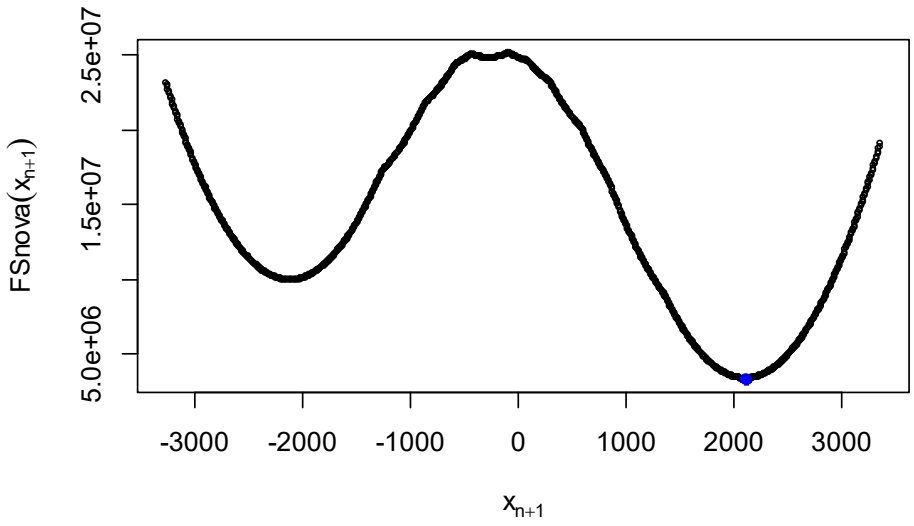

Figura 7.1: Função $FSnova(\mathbf{x}_{n+1}, \gamma = 0)$

O mesmo exemplo de redução de dimensionalidade para o espaço de dimensão um, com os mesmos dados apresentados no gráfico da Figura 7.1, é apresentado no processo típico de suavização na Figura 7.2. No processo típico de suavização, a cada

iteração o parâmetro γ decresce. A Figura 7.2 apresenta seis gráficos correspondentes a seis iterações. Em vermelho está destacado em cada gráfico o mínimo encontrado no processo de minimização da função suavizada (19) e em azul o mínimo global sem a suavização ou correspondente a forma suavizada com o parâmetro $\gamma = 0$.

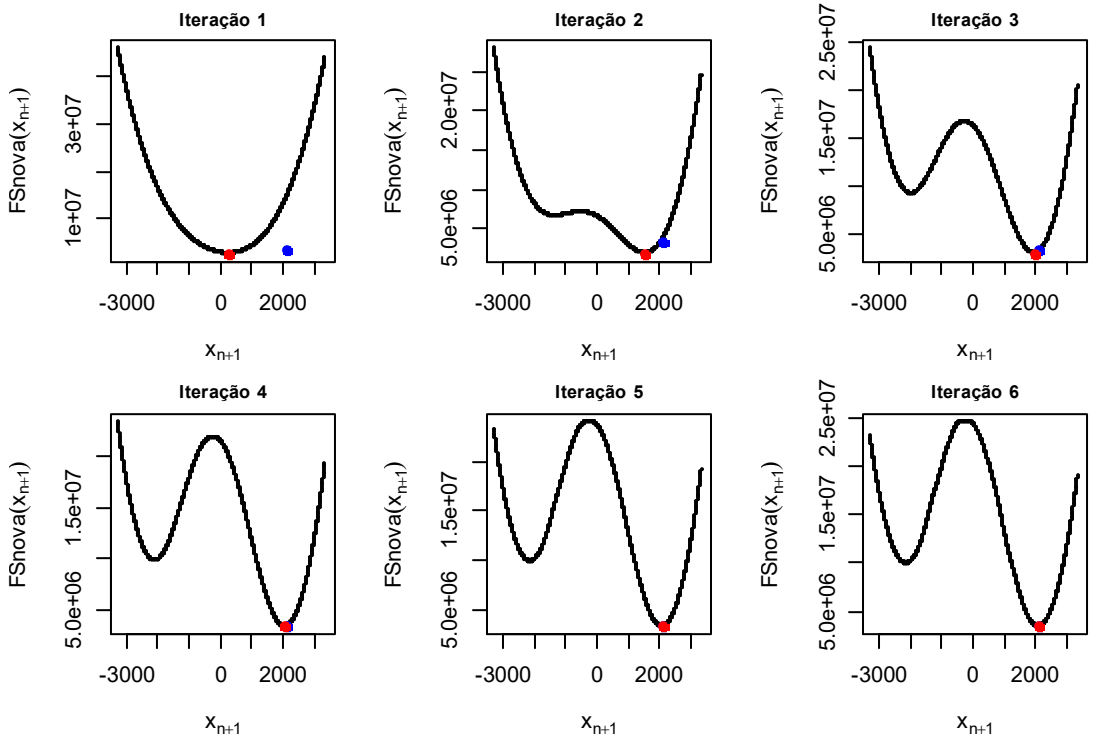

Figura 7.2: Processo de suavização $FSnova(\mathbf{x}_{n+1}, \gamma)$

Na primeira iteração, o valor do parâmetro γ alto produz uma função convexa. Na medida em que o parâmetro γ diminui, a função suavizada se aproxima da função não suavizada. A solução obtida em cada iteração é utilizada como ponto inicial para a iteração seguinte. Neste exemplo para qualquer escolha de ponto inicial, pelo fato da suavização ter produzido uma função convexa, a solução final converge para o valor de mínimo global.

A propriedade de convexificação da função $FSnova(\mathbf{x}_{n+1}, \gamma)$ será a seguir apresentada analiticamente. Cada componente do vetor gradiente da função $FSnova(\mathbf{x}_{n+1}, \gamma)$ é definida por:

$$\frac{\partial FSnova(\mathbf{x}_{n+1})}{\partial \mathbf{x}_{n+1}^{i}} = 2\sum_{p=1}^{k}\left(\frac{\overline{\mathbf{x}}_{p}^{i} - \mathbf{x}_{n+1}^{i}}{\theta(\overline{\mathbf{x}}_{p}, \mathbf{x}_{n+1}, \gamma)} + \frac{\mathbf{x}_{n+1}^{i} - \overline{\mathbf{x}}_{p}^{i}}{DZ_{n+1,p}} \right) \qquad (22)$$

A matriz Hessiana tem cada componente da sua diagonal principal definida por:

$$\frac{\partial^2 FSnova(\mathbf{x}_{n+1})}{\partial \mathbf{x}_{n+1}^{i} \partial \mathbf{x}_{n+1}^{i}} = 2\sum_{p=1}^{k}\left(\frac{(\overline{\mathbf{x}}_{p}^{i} - \mathbf{x}_{n+1}^{i})^2}{\theta(\overline{\mathbf{x}}_{p}, \mathbf{x}_{n+1}, \gamma)^3} - \frac{1}{\theta(\overline{\mathbf{x}}_{p}, \mathbf{x}_{n+1}, \gamma)} + \frac{1}{DZ_{n+1,p}} \right) \qquad (23)$$

A matriz Hessiana tem cada componente fora da diagonal principal, $i \neq j$, definida por:

$$\frac{\partial^2 FSnova(\mathbf{x}_{n+1})}{\partial \mathbf{x}_{n+1}^{i} \partial \mathbf{x}_{n+1}^{j}} = 2\sum_{a=1}^{k}\left(\frac{(\overline{\mathbf{x}}_{a}^{i} - \mathbf{x}_{n+1}^{i})(\overline{\mathbf{x}}_{a}^{j} - \mathbf{x}_{n+1}^{j})}{\theta(\overline{\mathbf{x}}_{a}, \mathbf{x}_{n+1}, \gamma)^2} \right) \qquad (24)$$

Na matriz Hessiana, na medida em que o parâmetro γ tende a ∞, todas as parcelas dos somatórios que possuem a função θ no denominador tenderão a zero. Sendo assim na equação (23), na medida em que o parâmetro γ cresce, os dois primeiros termos do somatório tenderão a zero e o terceiro termo prevalecerá sobre os demais, termo esse positivo por definição. No que concerne à equação (24), na medida em que o parâmetro γ cresce, todos os termos do somatório tenderão a zero. Pode-se assim concluir que para um valor do parâmetro γ suficientemente grande teremos uma matriz com uma diagonal estritamente dominante, com todos os componentes da diagonal principal positivos, desta forma a matriz Hessiana será estritamente definida positiva, sendo assim, a função $FSnova(\mathbf{x}_{n+1}, \gamma)$ será estritamente convexa para valores do parâmetro γ suficientemente grandes.

De modo análogo, a demonstração acima para a função $FSnova(\mathbf{x}_{n+1}, \gamma)$ também é válida para a função $FLSMDSnova(\mathbf{x}_{n+1}, \gamma)$, dada pela expressão (20). A função $FSnova(\mathbf{x}_{n+1}, \gamma)$ possui no denominador de cada parcela do somatório o termo $DZ_{n+1,p}$. Esse termo no denominados não consta na formulação $FLSMDSnova(\mathbf{x}_{n+1}, \gamma)$. Desta forma, as duas funções se diferenciam pela presença do termo $DZ_{n+1,p}$, termo esse que não depende de $\mathbf{x}_{n+1}$.

7.6 Resultados computacionais

Para ilustrar o funcionamento do método proposto, serão apresentados resultados computacionais obtidos usando duas instâncias. A primeira considera uma aplicação em redes de relacionamento, contendo dados sobre ataques terroristas. A segunda ilustra a aplicação em uma grande base de dados de imagens de dígitos.

A metodologia proposta foi implementada utilizando a linguagem estatística **R**. As tarefas de otimização foram realizadas por meio do método de Gradiente Conjugado, através da rotina *optim* da biblioteca *stats*. O processo de redução de dimensionalidade das observações do conjunto de treinamento e do conjunto de teste foi implementado utilizando processamento paralelo por meio da biblioteca doSNOW, com o processador Intel Core i7 3632QM.

Para mostrar o desempenho da metodologia proposta em redes de relacionamento, é apresentado um primeiro experimento computacional aplicando o método proposto no conjunto de dados *Terrorist Attacks*, disponível em http://linqs.umiacs.umd.edu/projects//projects/lbc/, no qual atentados terroristas são representados em um Grafo.

Este conjunto de dados consiste de 1293 ataques terroristas, classificados em seis tipos de ataques, $k = 6$, com os seguintes rótulos de classes e seus respectivos percentuais de observações em cada classe: *Arson* (2.4%), B*ombing* (43.5%), K*idnapping* (13.8%), *Nuclear, Biology, Chemistry and Radiology - NBCR* (0.6%), *Other attack* (1%) e *Weapon attack* (38.5%). Cada ataque é representado por um vetor com 106 atributos. Cada atributo assume um valor binário, indicando a ausência ou presença de uma particular característica. Assim, desde que o problema é definido no espaço dos reais, temos $\mathbf{z}_i \in \Re^{106}$, $i = 1, \cdots, 1293$. Maiores detalhes sobre esse cojunto de dados podem ser encontrados no artigo: *Event Classification and Relationship Labeling in Affiliation Networks* (ZHAO, SEN, GETOOR, 2006).

Utilizando a média de cada classe como protótipo, as observações foram reduzidas para um espaço de dimensão três. Essa redução viabiliza, em geral, uma boa visualização da rede. A Figura 7.3 mostra o posicionamento das observações da instância considerada após a redução segundo duas diferentes rotações no espaço. Na

figura, a classe *Arson* é representada em azul, *Bombing* em verde, *Kidnapping* em vermelho, *NBCR* em preto, *Other_attack* em marrom e *Weapon* em amarelo.

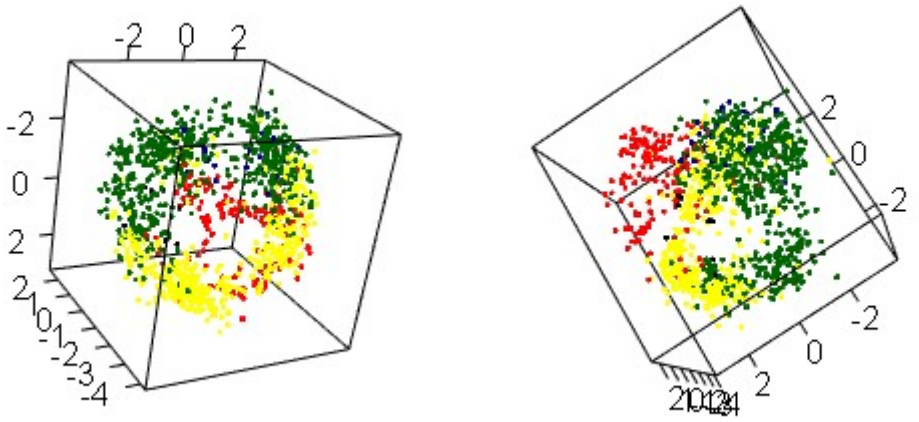

Figura 7.3: Redução de Dimensionalidade, $\mathbf{R}^{106} \rightarrow \mathbf{R}^{3}$, para Terrorist Attacks

Como mostrado pela Figura 7.3, a aplicação da metodologia proposta, de modo geral, produziu uma clara separação das observações segundo suas classes no espaço reduzido. As classes mais densas Bombing (verde), Kidnapping (vermelho), e Weapon_Attack (amarelo) podem ser visualizadas de uma forma bem separadas, com algumas sobreposições mostrando as intersecções entre as mesmas. Nas demais três classes não é possível identificar o mesmo comportamento, pois estão de forma difusa e totalizam um percentual muito pequeno dos dados (4,18%). A Figura 2 pode ser visualizada de forma interativa permitindo rotações, no site: https://dl.dropboxusercontent.com/u/110322274/TerrorAttack/index.html

O segundo experimento computacional foi realizado com a base de dados MNIST, base esta comumente utilizada na literatura em diversos artigos sobre métodos de reconhecimento de padrões e redução de dimensionalidade. Esta base de dados é formada por imagens de dígitos escritos à mão, contendo 60000 imagens no conjunto de treinamento e 10000 no conjunto de teste. Cada imagem é representada por um vetor com 784 atributos, $\mathbf{z}_i \in \Re^{784}$, e possui uma classe correspondente a um dígito de 0 até 9.

A metodologia proposta foi aplicada à base de dados completa, conjunto de treinamento e conjunto de teste, utilizando como o protótipo a média, calculada

utilizando o conjunto de treinamento, e reduzindo a dimensão para o $\Re^3$. Entretanto, no contexto de redução de dimensionalidade e visualização, por motivos computacionais, amostras são adotadas ao invés de se utilizar o conjunto completo da instância MNIST. MAATEN e HINTON (2008) e LEE *et al.* (2013) utilizam amostras dessa base de dados com somente 6000 observações.

A propriedade de escalabilidade da metodologia proposta é a seguir ilustrada com esta base de dados. Considerando somente o conjunto de treinamento, a utilização do método de Sammon para essa mesma instância compreenderia a resolução de um problema não linear com 3 X 60000 = 180000 variáveis. Assumido, para efeito de um exercício, a hipótese otimista de o problema não linear ter a mesma complexidade da resolução de um sistema de equações lineares, em que o número de operações aritméticas é proporcional ao número de variáveis ao cubo. Assim, para esse exemplo, teríamos o número de opererações aritméticas é da ordem de $180000^3 = 18^3\ 10^{12} = 5832\ 10^{12}$ operaçoes. Em contrapartida, o uso da metodologia proposta compreende a resolução de um primeiro problema não linear com 10 (número de dígitos) X 3 = 30 variáveis e subsequentes 60000 problemas não lineares com 3 variáveis. Para esse exemplo, considerando a metodologia inovadora, o número de operações aritméticas é da ordem de $30^3 + 60000\ X\ 3^3 = 1647000$ operações. Em suma, esse exercício mostra que o número de operações aritméticas da proposta inovadora é da ordem de $3{,}5\ 10^9$ menor do que a aplicação do método de Sammon no conjunto de treinamento da base de dados MNIST. Isso sem tomar em consideração as vantagens oferecidas pela resolução de um problema de otimização completamente diferenciável em comparação ao difícil problema não diferenciável de Sammon.

A Figura 7.4 mostra a projeção das 70000 observações da instância no espaço de dimensão três. É possível notar conjuntos densos de observações de uma mesma classe indicando uma separação adequada das classes. A Figura 7.4, pode ser visualizada de forma interativa permitindo rotações, no site: https://dl.dropboxusercontent.com/u/110322274/MNIST/index.html

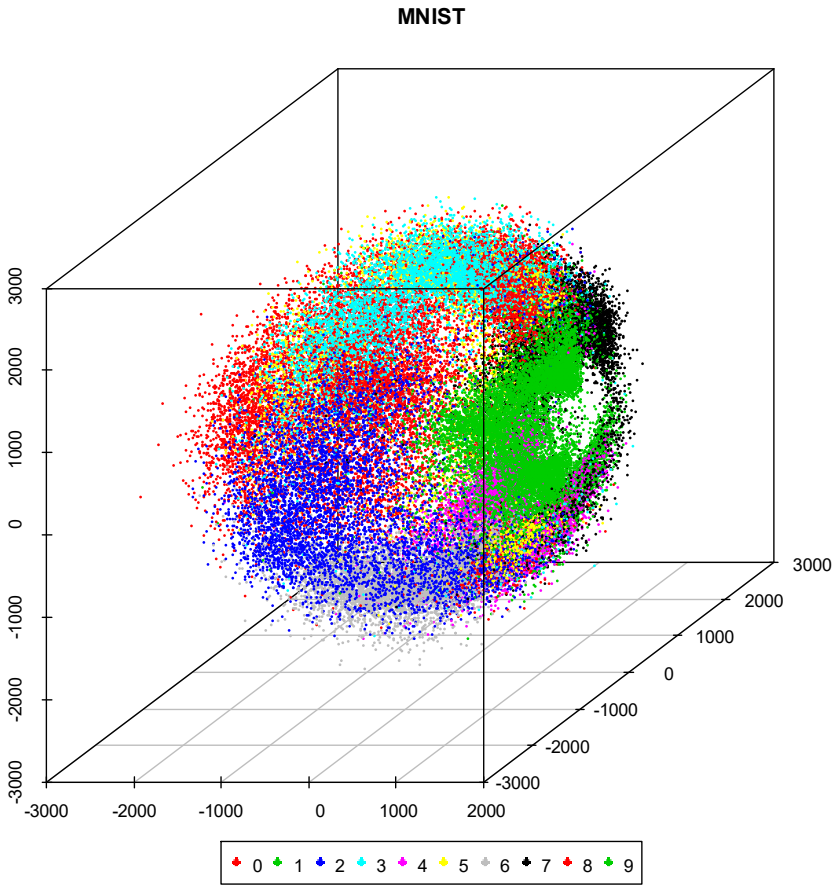

Figura 7.4: Redução de Dimensionalidade, $\mathbf{R}^{784} \rightarrow \mathbf{R}^{3}$, para MNIST

A posição de uma nova observação, sem a informação de classe, produzida pela aplicação da redução de dimensionalidade pode ser usada como indicador de sua classe de pertinência. Vários tipos de indicações podem ocorrer: Pertinência a uma classe, quando a observação está localizada em uma região do espaço que contenha somente uma classe sendo assim uma região pura; pertinência a uma região conflituosa entre classes quando a observação está localizada entre duas classes ou mais classes; não pertinência a uma específica classe, quando a observação está localizada em uma região do espaço que não contenha nenhuma observação da classe específica. A visualização pode ser utilizada em conjunto com um algoritmo de classificação supervisionada, permitindo uma avaliação visual do resultado da classificação feita pelo algoritmo.

7.7 Conclusões

A minimização das distorções entre a projeção no espaço reduzido e no espaço original dos dados não é uma tarefa trivial. A metodologia proposta tem a característica de gerar um conjunto de $n+T$ problemas de otimização separáveis, onde n faz referência ao número de observações do conjunto de treinamento e T faz referência ao número de observações do conjunto de teste. Cada problema possui dimensão muito baixa, com somente d variáveis, onde d faz referencia a dimensão do espaço reduzido. A propriedade de separabilidade, por sua vez, permite a utilização de computação paralela. Considerando somente o conjunto de treinamento, a aplicação direta dos métodos de redução de dimensionalidade que foram abordados no capítulo 2 o número de variáveis é extremamente maior, ns, inviabilizando sua aplicação prática em problemas de grande porte.

Ademais, como os problemas de otimização são não diferenciáveis, com a utilização da SH é proposta a substituição por alternativas aproximadas completamente diferenciáveis. Devido às características de baixa dimensão, de paralelismo no tratamento de cada observação e de suavização, torna-se possível resolver com eficiência e precisão problemas intratáveis por outras alternativas.

A resolução do problema teste NMIST com 70000 observações é um evento inaudito na literatura, que mostra de forma experimental uma medida do alcance da metodologia inovadora proposta nesta tese.

Capítulo 8

Conclusões

Na primeira parte desse trabalho é apresentada uma abordagem eficiente para métodos não lineares de redução de dimensionalidade da classe dos métodos de escalonamento multidimensional. Os métodos de Sammon, *Local MDS*, *Least Absolute Residuals* e *Supervised MDS* possuem a característica de serem não diferenciáveis. Com o emprego da suavização hiperbólica, uma formulação suavizada destes métodos foi proposta. Com a finalidade de validar a metodologia proposta, experimentos computacionais foram feitos com o algoritmo proposto e com outros algoritmos. Para uma comparação justa, utilizou-se sempre os mesmos pontos iniciais em todos os algoritmos testados. Além disso, experimentos computacionais foram feitos para comparar com resultados disponíveis na literatura. Para o problema de redução de dimensionalidade, em vista dos resultados obtidos, a proposta apresentada nesta tese pode ser uma boa alternativa frente a outros algoritmos tradicionais da literatura.

Na segunda parte deste trabalho, com o método *Supervised MDS* além do problema de redução de dimensionalidade foram abordados os problemas de Classificação Supervisionada e *Ranking* Bipartido, problemas estes muito relevantes e atuais. O método *Supervised MDS*, conforme proposto por WITTEN e TIBSHIRANI (2011), produz bons resultados e em alguns casos até melhores do que métodos tradicionais para os problemas de classificação supervisionada e *ranking* bipartido, como por exemplo, os métodos SVM, K-NN e regressão logística L_1. No processo de otimização dos problemas de classificação supervisionada e *ranking* bipartido nesta tese é proposta além da suavização das formulações o emprego de processamento paralelo. Novamente a metodologia proposta obteve na grande maioria dos experimentos um resultado superior quando comparado com o algoritmo *Supervised MDS*. Com base nos resultados obtidos, a metodologia proposta pode ser utilizada em muitas aplicações

práticas inclusive substituindo métodos tradicionais de classificação supervisionada e *ranking* bipartido.

Na terceira parte deste trabalho é proposta uma nova metodologia baseada em protótipos tendo como base a classe de métodos não lineares de preservação de distâncias. Os métodos de Sammon, *Least Squares MDS* e *Least Absolute Residuals* foram utilizados para a redução de dimensionalidade dos protótipos. Considerando as coordenadas dos protótipos no espaço de baixa dimensão, propõe-se a generalização dos método de Sammon, *Least Squares MDS* e *Least Absolute Residuals* para redução de dimensionalidade de uma observação nova. O mesmo critério utilizado na redução de dimensionalidade dos protótipos é empregado para as observações dos conjuntos de treinamento e do conjunto de teste. Assim, a metodologia proposta tem a importante característica de possuir unicidade de critério no tratamento de todas as observações.

Essa metodologia tem a vantajosa característica de gerar problemas de otimização independentes e de dimensão muito baixa, de modo que o número de observações não influencia na dimensão de cada problema. Como os problemas de otimização são não diferenciáveis, é proposta a substituição por alternativas aproximadas completamente diferenciáveis. Na generalização dos métodos de Sammon e *Least Squares MDS* é demonstrada a propriedade de covexificação e por conseguinte redução de mínimos locais.

Devido as características de diferenciabilidade e por serem problemas de dimensão muito baixa, torna-se viável a resolução de problemas de grande porte. A metodologia foi implementada utilizando processamento paralelo. Resultados computacionais foram apresentados, incluindo toda a instância NMIST, instância essa muito utilizada como exemplo em outros métodos de redução de dimensionalidade congêneres, entretanto os outros métodos trabalham com amostras dos dados e não processam o conjunto como um todo.

Referências Bibliográficas

AGARWAL, S., 2005, A *study of the bipartite ranking problem in machine learning*, PhD dissertation, University of Illinois at Urbana-Champaign.

AGRAFIOTIS, D.K., RASSOKHIN, D.N. AND LOBANOV, V.S., 2001, "Multidimensional scaling and visualization of large molecular similarity tables", *Journal of Computational Chemistry*, *22*(5), pp. 488-500.

BAGIROV, A. M., 2008, "Modified Global k-means Algorithm for Minimum Sum-of-Squares Clustering Problems", *Pattern Recognition*, v. 41 Issue 10, pp. 3192-3199.

BAGIROV, A. M., ORDIN, B., OZTURK, G. AND XAVIER, A. E., 2012, "An incremental clustering algorithm based on smoothing techniques", *Computational Optimization and Applications*, published Online Oct 2014. DOI 10.1007/s10589-014-9711-7

BASALAJ, W., 1999, "Incremental multidimensional scaling method for database visualization", *Electronic Imaging'99* , pp. 149-158, International Society for Optics and Photonics.

BÉCAVIN, C., TCHITCHEK, N., MINTSA-EYA, C., LESNE, A., BENECKE, A., 2011, "Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition", *Bioinformatics*,*27*(10), pp.1413-1421.

BORG, I., GROENEN, P.J. AND MAIR, P., 2012, *Applied multidimensional scaling*, Springer Science & Business Media, ISSN 2191-544X ISSN 2191-5458 (electronic), ISBN 978-3-642-31847-4 ISBN 978-3-642-31848-1 (eBook) DOI 10.1007/978-3-642-31848-1

BORG, I., GROENEN, P.J., 2005, *Modern multidimensional scaling: Theory and applications*, Springer Science & Business Media

CHANG, C.L., LEE, R.C.T., 1973, "A heuristic relaxation method for nonlinear mapping in cluster analysis", *IEEE Transactions on Systems, Man, and Cybernetics*, *2*(SMC-3), pp. 197-200.

CHEN, L., BUJA, A., 2009, "Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis", *Journal of the American Statistical Association*, *104*(485), pp. 209-219.

CHEN, L., BUJA, A., 2006, Local multidimensional scaling for nonlinear dimension reduction, graph layout and proximity analysis. *Unpublished thesis, University of Pennsylvania, http://www-stat. wharton. upenn. edu/buja/PAPERS/lmds-chen-buja. pdf.*

COX, T.F., FERRY, G., 1993, "Discriminant analysis using non-metric multidimensional scaling". *Pattern Recognition*, *26*(1), pp.145-153.

COX, T. F., COX, M. A. A., 2000, *"Multidimensional Scaling",* Chapman & Hall.

DE LEEUW, J., 1977, ″Applications of convex analysis to multidimensional scaling*″,* In: J. R. Barra, F. Brodeau, G. Romier, & B. van Cutsem (Eds.), *Recent developments in statistics* (pp. 133–145). Amsterdam, The Netherlands: North-Holland.

DEM'YANOV, V.F., MALOZEMOV, V.N., 1974. *Introduction to minimax*. Courier Corporation.

DU, D.Z., PARDALOS, P.M., 2013. *Minimax and applications,* v. 4, Springer Science & Business Media.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R., 2001, *The elements of statistical learning,* v. 1, Springer, Berlin: Springer series in statistics.

GISBRECHT, A., SCHULZ, A. HAMMER, B., 2015, "Parametric nonlinear dimensionality reduction using kernel t-SNE", *Neurocomputing*, *147*, pp.71-82

GOWER, J. C., 1966, "Some distance properties of latent root and vector methods used in multivariate analysis", *Biometrika* 53, pp. 325–328.

GROENEN, P.J.F., HEISER, W.J., MEULMAN, J.J., 1999, ″Global Optimization in Least-Squares Multidimensional Scaling by Distance Smoothing"*, Journal of Classification*, v. 16, pp. 225-254.

GUTTMAN, L., 1968, *"*A general nonmetric technique for finding the smallest coordinate space for a configuration of points*"*, *Psychometrika*, *33*, pp. 469–506.
HEISER, W.J., 1988, "Multidimensional scaling with least absolute residuals", *Classification and related methods*, pp. 455-462.

HILTON, G., ROWEIS, S., 2003, *"*Stochastic Neighbor Embedding*",* *Advances in Neural Information Processing Systems*, 15, MIT Press, Cambridge, MA.

HOTELLING, H., 1933, *"*Analysis of a complex of statistical variables into principal components"*,* *Journal of Educational Psychology*, *24*, pp. 417–441.
JACOBS, D.W., WEINSHALL, D., GDALYAHU, Y., 2000, "Classification with nonmetric distances: Image retrieval and class representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(6), pp. 583-600.

KEARSLEY, A.J., TAPIA, R.A. AND TROSSET, M.W., 1998, *The solution of the metric STRESS and SSTRESS problems in multidimensional scaling using Newton's method* (No. TR94-44). Rice Univ Houston Tx Dept of Computational and Applied Mathematics.

KRUSKAL, J. B., 1964a, *"*Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*", Psychometrika*, 29, pp. 1–27

KRUSKAL, J. B. 1964b, *"*Nonmetric multidimensional scaling: A numerical method*",* *Psychometrika*, *29*, pp. 115–129.
LEE, J.A., VERLEYSEN, M., 2007, *Nonlinear dimensionality reduction*. Springer Science & Business Media.
LEE, R.C.T., SLAGLE, J.R., BLUM, H., 1977, "A triangulation method for the sequential mapping of points from N-space to two-space", *IEEE Transactions on* Computers, 100(3), pp. 288-292.
LEE J, VERLEYSEN M., 2009, "Quality assessment of dimensionality reduction: Rank-based criteria" *Neurocomputing* 72: pp. 1431–1443. doi: 10.1016/j.neucom.2008.12.017
LEE, J.A, VERLEYSEN, M., 2005, " Nonlinear dimensionality reduction of data manifolds with essential loops", *Neurocomputing*, *67*, pp.29-53.

LEE, J.A., RENARD, E., BERNARD, G., DUPONT, P., VERLEYSEN, M., 2013, "Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation", *Neurocomputing*, *112*, pp. 92-108.

LEEUW, J.D., MAIR, P., 2009, "Multidimensional Scaling Using Majorization: SMACOF in R", *Journal of Statistical Software*, 31(3), pp. 1-30. URL http://www.jstatsoft.org/v31/i03/.

LERNER, B., GUTERMAN, H., ALADJEM, M., DINSTEIN, I., ROMEN, Y., 1998, *"*On Pattern Classification with Sammon's Nonliear Mapping - An Experimental Study*",* *Pattern Recognition*, Vol 31, pp. 371-381. doi:10.1016/S0031-3203(97)00064-

MAATEN, L.V.D., HINTON, G., 2008, "Visualizing data using t-SNE", *Journal of Machine Learning Research*, *9*(Nov), pp. 2579-2605.

MOODY, J.,DARKEN, C.J., 1989, "Fast learning in networks of locally-tuned processing units", *Neural computation*, *1*(2), pp.281-294.

MORÉ J.J., WU Z. 1995, "Epsilon-optimal solutions to distance geometry problems via global continuation", Mathematics and Computer Science Division, Argonne Lab., Preprint MCS-P520-0595.

NAGARAJAN, M.B., COAN, P., HUBER, M.B., DIEMOZ, P.C., WISMÜLLER, A., 2015, "Integrating dimension reduction and out-of-sample extension in automated classification of ex vivo human patellar cartilage on phase contrast X-ray computed tomography". *PloS one*, *10*(2), p.e0117157.

NAJIM, S.A., LIM, I.S., 2014 *"*Trustworthy Dimension Reduction for Visualization Different Data Sets *"*, *Information Sciences*, v. 278, pp. 206-220.

NAUD, A. AND DUCH, W., 2000, June. Interactive data exploration using MDS mapping. In: *Proceedings of the Fifth Conference: Neural Networks and Soft Computing,* pp. 255-260.

PARDALOS, P. M., BATSYN, M., 2014, *Constructive nonsmooth analysis and related topics*. V. F. Demyanov (Ed.). New York: Springer.

PARDALOS, P.M., SHALLOWAY, D., XUE, G., 1996, *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding: DIMACS* Workshop, March 20-21, 1995, v, 23, American Mathematical Soc.

PEKALSKA, E., DE RIDDER, D., DUIN, R. P., & KRAAIJVELD, M. A., 1999, A new method of generalizing Sammon mapping with application to algorithm speed-up. In: *ASCI,* v. 99, pp. 221-228).

PLINER, V., 1996, ″Metric Unidimensional Scaling and Global Optimization*″, Journal of Classification*, v. 13, pp. 3-18.

R Core Team, 2014, R: A language and environmentfor statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

ROWEIS, S. T., SAUL, L. K., 2000, ‘‘Nonlinear Dimensionality reduction by Local Linear Embedding,’’ *Science* 290, pp. 2323–2326.

RUBINOV A. M., 2006, ″Methods for Global Optimization of Nonsmooth Functions with Applications*″, Applied and Computational Mathematics*, 5, pp. 3-15.

SAMMON, J.W., 1969, "A nonlinear mapping for data structure analysis", *IEEE Transactions on computers*, *18*(5), pp.401-409.

SCHACHTER, B., 1978, "A nonlinear mapping algorithm for large data sets", *Computer Graphics and Image Processing*, *8*(2), pp.271-276.

SORZANO, C.O.S., VARGAS, J., MONTANO, A.P., 2014, A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

SOUZA, M., XAVIER, A.E., LAVOR, C., MACULAN, N., 2011. "Hyperbolic smoothing and penalty techniques applied to molecular structure determination", *Operations Research Letters*, *39*(6), pp.461-465.
SHEPARD, D., 1968, "A two-dimensional interpolation function for irregularly-spaced data", In: *Proceedings of the 1968 23rd ACM national conference*, pp. 517-524.

SUN, J., CROWE, M., FYFE, C., 2011, ″Extending Metric Multidimensional Scalling with Bregman Divergence"*, Pattern Recognition*, v. 44, pp. 1137-1154.

TENENBAUM, J.B., DE SILVA, V., LANGFORD, J.C., 2000, "A global geometric framework for nonlinear dimensionality reduction", *Science* 290, n. 5500, pp 2319–2323.

TORGERSON, W.S., 1958, *Theory and methods of scaling*.

VAN DER MLJP PE, VAN DEN HH J., 2009, Dimensionality reduction: A comparative review. Tilburg, Netherlands: Tilburg Centre for Creative Computing, Tilburg University, Technical Report: 2009-005

VAN DER MAATEN, L.J.P., 2007, An introduction to dimensionality reduction using matlab. *Report*, *1201*(07-07), p.62.

VENABLES, W.N., RIPLEY, B.D., 2013, *Modern applied statistics with S-PLUS*, Springer Science & Business Media.

VENCESLAU, H. M., LÜBKE, D. C., XAVIER A. E., 2014, ″Optimal Covering of Solid Bodies by Spheres Via Hyperbolic Smoothing Technique″, *Optimization Methods & Software*, pp 1-13, Published online: 23 Jul 2014 DOI 10.1080/10556788.2014.934686.

VENNA, J., KASKI, S., 2006, ″Local Multidimensional Scalling″, *Neural Networks*, v. 19, pp. 889-899.

WANG, F. AND SUN, J., 2015, "Survey on distance metric learning and dimensionality reduction in data mining.", *Data Mining and Knowledge Discovery*, *29*(2), pp. 534-564.

WITTEN, D.M., TIBSHIRANI, R., 2011, "Supervised multidimensional scaling for visualization, classification, and bipartite ranking", *Computational Statistics & Data Analysis*, *55*(1), pp. 789-801.

XAVIER, A.E., 1982, *Penalização hiperbólica: Um novo método para resolução de problemas de otimização,* M. Sc. Thesis, COPPE - UFRJ, Rio de Janeiro.

XAVIER, A.E., 2001, "Hyperbolic Penalty: A New Method for Nonlinear Programming with Inequalities", *International Transactions in Operational Research*, v. 8, Issue 6, pp:659–671. doi: 10.1111/1475-3995.t01-1-00330.

XAVIER, A.E., 2003, *"Convexificaçãodo problema de distância geométrica através da técnica de suavização hiperbólica",* Workshop em Biociências , COPPE UFRJ, Rio de Janeiro.

XAVIER, A.E., DE OLIVEIRA, A.A.F., 2005, "Optimal covering of plane domains by circles via hyperbolic smoothing", *Journal of Global Optimization*,*31*(3), pp.493-504.

XAVIER, A.E., 2010, "The hyperbolic smoothing clustering method", Pattern Recognition, Vol. 43, pp. 731-737.

XAVIER, A.E., XAVIER, V.L., 2011, "Solving the minimum sum-of-squares clustering problem by hyperbolic smoothing and partition into boundary and gravitational regions"*, Pattern Recognition*, v. 44, pp. 70-77.

XAVIER, A.E., GESTEIRA, C.M., XAVIER, V.L., 2015, "Solving the continuous multiple allocation p-hub median problem by the hyperbolic smoothing approach", *Optimization*, *64*(12), pp.2631-2647.

XAVIER, V.L. 2012, *Resolução do Problema de Agrupamento segundo o Critério de Minimização da Soma das Distâncias,* M.Sc. Thesis, COPPE - UFRJ, Rio de Janeiro.

XAVIER, V.L., FRANÇA, F.M., XAVIER, A.E., LIMA, P.M., 2014a. "A hyperbolic smoothing approach to the Multisource Weber problem", *Journal of Global Optimization*, *60*(1), pp.49-58.

XAVIER, A.E. , XAVIER, V.L., 2014, "Flying elephants: a general method for solving non-differentiable problems", *Journal of Heuristics*, pp.1-16.

ZHAO, B., SEN, P., GETOOR, L., 2006, "Event classification and relationship labeling in affiliation networks", In: *Proceedings of the Workshop on Statistical Network Analysis (SNA) at the 23rd International Conference on Machine Learning (ICML)*.

ZHANG, S., XIE, C., FAN, C., ZHANG, Q., 2007, ″An Alternate Method of Hierarchical Classification for E-nose: Combined Fisher Discriminant Analysis and Modified Sammon Mapping*″, Sensors and Actuators B*, v. 127, pp. 399-405.