Paola de Oliveira Prado

Aplicação de técnicas multivariadas para visualização de dígitos manuscritos

Niterói - RJ, Brasil

14 de Dezembro de 2020

Universidade Federal Fluminense

Paola de Oliveira Prado

Aplicação de técnicas multivariadas para visualização de dígitos manuscritos

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Prof. Dr. Hugo Henrique Kegler dos Santos

Co-Orientador: Prof. Dr. Valentin Sisko

Niterói - RJ, Brasil

14 de Dezembro de 2020

Universidade Federal Fluminense

Paola de Oliveira Prado

Aplicação de técnicas multivariadas para visualização de dígitos manuscritos

Monografia de Projeto Final de Graduação sob o título *\Aplicacao de tecnicas multivariadas para visualizacao de d gitos manuscritos"*, defendida por Paola de Oliveira Prado e aprovada em 14 de Dezembro de 2020, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Hugo Henrique Kegler dos Santos
Departamento de Estatística – UFF

Prof. Dr. Valentin Sisko
Departamento de Estatística – UFF

Profa. Dra. Estelina Serrano de Marins Capistrano
Departamento de Estatística – UFF

Profa. Dra. Jessica Quintanilha Kubrusly
Departamento de Estatística – UFF

Niterói, 14 de Dezembro de 2020

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

P896a Prado, Paola de Oliveira
Aplicação de técnicas multivariadas para visualização de dígitos manuscritos / Paola de Oliveira Prado ; Hugo Henrique Kegler dos Santos, orientador ; Valentin Sisko, coorientador. Niterói, 2020.
52 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2020.

1. Análise Multivariada. 2. Escalonamento Multidimensional. 3. T-SNE. 4. MNIST. 5. Produção intelectual. I. Santos, Hugo Henrique Kegler dos, orientador. II. Sisko, Valentin, coorientador. III. Universidade Federal Fluminense. Instituto de Matemática e Estatística. IV. Título.

CDD -

Bibliotecário responsável: Sandra Lopes Coelho - CRB7/3389

Resumo

Atualmente, bases de imagens são bastante utilizadas para a classificação de objetos na área de aprendizado de máquina ou, como mais conhecido em inglês, *Machine Learning*. Dentro dessa esfera, as bases de dígitos manuscritos vem sendo muito empregadas, principalmente, para um estudo inicial dessa área e testar o desempenho dos algoritmos. Este trabalho aplica as técnicas de Escalonamento Multidimensional, Análise de Agrupamento e *t-SNE* a fim de verificar seus desempenhos na visualização dos dígitos manuscritos. Para tal, foi utilizada a base de dígitos manuscritos, *MNIST*, com 10.000 observações. Devido a problemas de processamento computacional, realizou-se uma amostragem com 2.000 e 4.000 observações. O resultado para a técnica não linear, *t-SNE*, apresentou melhores visualizações comparado as outras técnicas analisadas.

Palavras-chave: Análise Multivariada. Análise de Agrupamento. Aprendizado de Máquina. Escalonamento Multidimensional. *MNIST*. *t-SNE*.

Agradecimentos

Primeiro eu agradeço a Deus por ter me sustentado para chegar a este momento e por permitir que eu conclua uma etapa tão especial na minha vida.

Aos meus pais por todo suporte que me dão. Aos amigos que foram um amparo para mim durante toda essa caminhada. Ao meu namorado Renan por todo suporte nessa etapa final.

Aos professores que passaram na minha vida até o presente momento, e em especial agradeço ao meu orientador Professor Hugo por todo o apoio dado no desenvolvimento deste trabalho e ao meu co-orientador Professor Valentin por sempre ter me ajudado quando preciso.

Por último, sou muito grata a todos que passaram pela minha vida em algum momento e fizeram parte da minha jornada.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Nesse capítulo será apresentada a motivação para o presente trabalho, a organização dos capítulos, assim como o objetivo do trabalho.

1.1 Motivação

Nos último anos, juntamente com os avanços tecnológicos, ocorreu um aumento no volume de dados produzidos; entretanto, analisá-los nem sempre é uma tarefa fácil. Quando se estuda um conjunto de dados em que os elementos possuem uma dimensão muito grande, perde-se a intuição do que está acontecendo, o que torna preferível uma redução da dimensionalidade dos dados.

Uma das primeiras técnicas apresentadas para redução da dimensionalidade foi o escalonamento multidimensional, desenvolvido principalmente para permitir a visualização dos dados em uma dimensão menor que a original. O conceito básico dessa técnica foi introduzido por Richardson (1938), sendo popularizado por Torgerson (1958), que incorporou o termo escalonamento multidimensional. A técnica baseia-se em relações lineares das matrizes de distâncias entre os objetos, sendo por isso muito limitada. Outra técnica bastante difundida é a análise de componentes principais Hotelling (1933), que tem como objetivo sintetizar as informações de várias variáveis em um conjunto menor de variáveis não correlacionadas. Essa técnica também baseia-se em relações lineares, ou seja, tem o foco em preservar as dissimilaridades na nova dimensão.

Atualmente, com o avanço computacional, outras técnicas mais elaboradas têm sido desenvolvidas, como por exemplo *Sammon mapping*, *Isomap*, *Locally Linear Embedding* e, mais recentemente, a técnica *t-Distributed Stochastic Neighbor Embedding*, *t-SNE*, proposta por Maaten e Hinton (2008), que tem apresentado melhores resultados na visualização dos dados, principalmente em dados mais complexos. Essas técnicas são não lineares, ou seja, tem o foco em preservar as similaridades na nova dimensão, ao contrário

das técnicas lineares, que focam em preservar as dissimilaridades.

Dentro dessa nova era de grande volumes de dados e também da necessidade de, em alguns casos, analisar dados mais complexos, o termo aprendizado de máquinas tem se tornado cada vez mais abordado. Aprendizado de máquinas é um processo pelo qual são criados algoritmos para os computadores desenvolverem o reconhecimento de padrões ou a capacidade de aprender continuamente com os dados. Uma das aplicações em aprendizado de máquina é, por exemplo, em segurança, com reconhecimento facial. Isto tem sido usado tanto em cidades, o que tem causado discussões sobre privacidade Presse (2019), como em celulares para desbloqueá-los. Outra aplicação que tem sido bastante útil é a criação de *softwares* que auxiliam pessoas com algum tipo de deficiência a uma maior inserção na sociedade, minimizando os percalços da deficiência Estadao (2019).

Neste trabalho é aplicada a técnica de redução de dimensionalidade linear, escalonamento multidimensional, juntamente com a análise de agrupamento; e aplicada a técnica de redução não linear *t-SNE*. Para tal, analisa-se uma base de dados composta de dígitos manuscritos, muito utilizada na área de aprendizado de máquinas.

O trabalho está organizado em quatro capítulos, sendo separado da seguinte forma: o Capítulo 1 apresenta a motivação para a elaboração deste trabalho, e também a descrição do objetivo geral e específico. O Capítulo 2 apresenta a base de dados utilizada, bem como uma explicação da sua estrutura, e a metodologia utilizada. O Capítulo 3 apresenta uma aplicação dos métodos apresentados na base de dígitos manuscritos, ou seja, escalonamento multidimensional, *k-means* e *t-SNE*. Por fim, o Capítulo 4 conclui o trabalho realizado. O Apêndice 1 contém os códigos computacionais para a visualização dos dígitos e leitura da base de dados, o Apêndice 2 visualizações adicionais do *t-SNE*. Já o Anexo 1 possui apresenta a demonstração do escalonamento multidimensional métrico.

1.2 Objetivos

Objetivo Geral

O objetivo geral deste trabalho é aplicar técnicas multivariadas com o foco em técnicas de redução de dimensionalidade e posteriormente comparar essas técnicas a fim de verificar a existência de diferença no desempenho.

Objetivo Específico

O objetivo específico é estudar a visualização de dígitos manuscritos de acordo com

as técnicas de redução de dimensionalidade.

2 Materiais e Métodos

Nesse capítulo serão descritos os dados utilizados assim como os métodos aplicados para o desenvolvimento do trabalho.

2.1 Materiais

O material utilizado será a base de dados *Modi ed National Institute of Standards and Technology* conhecida como *MNIST*[1] que é um subconjunto do conjunto do *National Institute of Standards and Technology*, *NIST*. A base é composta pelo arquivo de imagens, que são pequenas imagens de dígitos de 0 a 9 escritos à mão, no qual os dígitos foram normalizados em tamanho e centralizados em uma imagem de tamanho fixo de forma que cada imagem ocupe 28 28 pixels, ou seja, 784 dimensões e pelo arquivo de rótulos, que contém os dígitos de 0 a 9 associados na ordem das imagens do arquivo de imagens. A base de dados vem divida no conjunto de treinamento com 60.000 observações e nos dados de teste com 10.000 observações.

Para retratar como é um arquivo de imagem, a Figura 1 mostra a visão dos dígitos escritos enquanto a Figura 2 mostra os dígitos 2 e 9, destacados na Figura 1, com os valores dos pixels. Pode-se notar que os dígitos na Figura 2 são formados pelos números de 0 a 255, indicando a escala de cinza, onde 0 indica o fundo na cor branca e quanto maior é valor, maior a tonalidade, sendo 255 a cor preta.

[1] A base de dados *MNIST* esta dispon vel no site: *h*http://yann.lecun.com/exdb/mnist/*i*

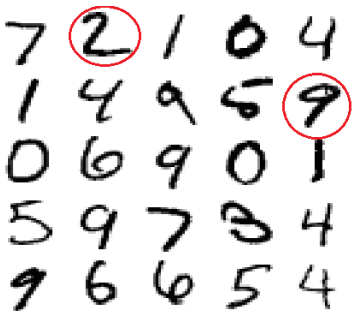


Figura 1: Dígitos de um arquivo de imagem

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	[,28]
[1,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0	116	125	171	255	255	150	93	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0	169	253	253	253	253	253	253	218	30	0	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	0	0	0	169	253	253	253	213	142	176	253	253	122	0	0	0	0	0	0	0	0	0	0
[7,]	0	0	0	0	0	0	0	52	250	253	210	32	12	0	6	206	253	140	0	0	0	0	0	0	0	0	0	0
[8,]	0	0	0	0	0	0	0	77	251	210	25	0	0	0	122	248	253	65	0	0	0	0	0	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	0	31	18	0	0	0	0	209	253	253	65	0	0	0	0	0	0	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	0	0	0	0	0	117	247	253	198	10	0	0	0	0	0	0	0	0	0	0
[11,]	0	0	0	0	0	0	0	0	0	0	0	0	76	247	253	231	63	0	0	0	0	0	0	0	0	0	0	0
[12,]	0	0	0	0	0	0	0	0	0	0	0	0	128	253	253	144	0	0	0	0	0	0	0	0	0	0	0	0
[13,]	0	0	0	0	0	0	0	0	0	0	0	176	246	253	159	12	0	0	0	0	0	0	0	0	0	0	0	0
[14,]	0	0	0	0	0	0	0	0	0	0	25	234	253	233	35	0	0	0	0	0	0	0	0	0	0	0	0	0
[15,]	0	0	0	0	0	0	0	0	0	0	198	253	253	141	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[16,]	0	0	0	0	0	0	0	0	0	78	248	253	189	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[17,]	0	0	0	0	0	0	0	0	19	200	253	253	141	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[18,]	0	0	0	0	0	0	0	0	134	253	253	173	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[19,]	0	0	0	0	0	0	0	0	248	253	253	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[20,]	0	0	0	0	0	0	0	0	248	253	253	43	20	20	20	20	5	0	5	20	20	37	150	150	150	147	10	0
[21,]	0	0	0	0	0	0	0	0	248	253	253	253	253	253	253	253	168	143	166	253	253	253	253	253	253	253	123	0
[22,]	0	0	0	0	0	0	0	0	174	253	253	253	253	253	253	253	253	253	253	253	249	247	247	169	117	117	57	0
[23,]	0	0	0	0	0	0	0	0	0	118	123	123	123	166	253	253	253	155	123	123	41	0	0	0	0	0	0	0
[24,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[25,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[26,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[27,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[28,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	[,28]
[1,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[7,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[8,]	0	0	0	0	0	0	0	0	0	0	0	0	0	36	56	137	201	199	95	37	0	0	0	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	0	0	0	0	45	152	234	254	254	254	254	254	250	211	151	6	0	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	0	46	153	240	254	254	227	166	133	251	200	254	229	225	104	0	0	0	0	0
[11,]	0	0	0	0	0	0	0	0	153	234	254	254	187	142	8	0	0	191	40	198	246	223	253	21	0	0	0	0
[12,]	0	0	0	0	0	0	8	126	253	254	233	128	11	0	0	0	0	210	43	70	254	254	254	21	0	0	0	0
[13,]	0	0	0	0	0	0	72	243	254	228	54	0	0	0	0	3	32	116	225	242	254	255	162	5	0	0	0	0
[14,]	0	0	0	0	0	0	75	240	254	223	109	138	178	178	169	210	251	231	254	254	254	232	38	0	0	0	0	0
[15,]	0	0	0	0	0	0	9	175	244	253	255	254	254	251	254	254	254	254	254	252	171	25	0	0	0	0	0	0
[16,]	0	0	0	0	0	0	0	0	16	136	195	176	146	153	200	254	254	254	254	150	16	0	0	0	0	0	0	0
[17,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	162	254	254	241	99	3	0	0	0	0	0	0	0	0
[18,]	0	0	0	0	0	0	0	0	0	0	0	0	0	118	250	254	254	90	0	0	0	0	0	0	0	0	0	0
[19,]	0	0	0	0	0	0	0	0	0	0	0	0	100	242	254	254	211	7	0	0	0	0	0	0	0	0	0	0
[20,]	0	0	0	0	0	0	0	0	0	0	0	54	241	254	254	242	59	0	0	0	0	0	0	0	0	0	0	0
[21,]	0	0	0	0	0	0	0	0	0	0	0	131	254	254	244	64	0	0	0	0	0	0	0	0	0	0	0	0
[22,]	0	0	0	0	0	0	0	0	0	0	13	249	254	254	152	0	0	0	0	0	0	0	0	0	0	0	0	0
[23,]	0	0	0	0	0	0	0	0	0	12	228	254	254	208	8	0	0	0	0	0	0	0	0	0	0	0	0	0
[24,]	0	0	0	0	0	0	0	0	0	78	255	254	254	66	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[25,]	0	0	0	0	0	0	0	0	0	209	254	254	137	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[26,]	0	0	0	0	0	0	0	0	0	227	255	233	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[27,]	0	0	0	0	0	0	0	0	0	113	255	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[28,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 2: Matriz do arquivo de imagem

Visto que já possui um pré processamento e formatação, o *MNIST* é um bom banco de dados para quem busca experimentar técnicas de aprendizado e método de reconhecimento de padrões em dados do mundo real LeCun, Cortes e Burges (2019).

Para a leitura da base de dados e a realização das análises estatísticas será utilizado o *software* R (R Core Team, 2019). Dentro do *software* será utilizado o pacote *dplyr*, criado por Wickham et al. (2019), para manipulação de dados; o pacote *ggplot2*, criado por Wickham (2016), para ilustração gráfica; e o pacote *Rtsne*, criado por Krijthe (2015), para implementar o algoritmo do t-SNE.

2.2 Medidas de Proximidade

As técnicas utilizadas nesse trabalho demandam que alguma medida de proximidade seja previamente calculada. As medidas de proximidade são as medidas de similaridade e dissimilaridade. As medidas de dissimilaridades são as distâncias, e devem satisfazer às condições descritas na Definição 2.1, enquanto as medidas de similaridades são complementares às distâncias.

Definição 2.1. *Dados 3 vetores* $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ 2 $\mathbb{R}^p$, *a distância entre eles é uma função* $d : \mathbb{R}^p \times \mathbb{R}^p \to [0; +\infty)$ *que satisfaz as seguintes condições:*

1. *Simetria:* $d(\mathbf{x}; \mathbf{y}) = d(\mathbf{y}; \mathbf{x})$
2. *Indiferenciabilidade de idênticos:* $d(\mathbf{x}; \mathbf{y}) = 0$ *se, e somente se,* $\mathbf{x} = \mathbf{y}$
3. *Desigualdade triangular:* $d(\mathbf{x}; \mathbf{y}) \leq d(\mathbf{x}; \mathbf{z}) + d(\mathbf{z}; \mathbf{y})$

O trabalho trata as medidas de dissimilaridades. Sendo assim, é necessário calcular uma medida de distância entre os dados, de forma que os menores valores dessa distância vão representar maior similaridade entre os objetos. Existem diversas maneiras de calcular distâncias entre pares de objetos, a distância Euclidiana é a mais comumente utilizada e será a única tratada para as aplicações desse trabalho. Essa distância é definida da seguinte forma:

Definição 2.2. *Sejam dois vetores* $\mathbf{x}_i$ *e* $\mathbf{x}_j$ *observações de* $\mathbf{x} = (x_1; x_2; \ldots; x_p)$ 2 $\mathbb{R}^p$, *então*

$$\begin{aligned} d(\mathbf{x}_i; \mathbf{x}_j) &= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \\ &= \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \\ &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j)} \end{aligned} \tag{2.1}$$

sendo x_{ik} e x_{jk}, $k = 1; 2; \ldots; n$ os valores para as variáveis $\mathbf{x}_i$ e $\mathbf{x}_j$, respectivamente.

2.3 Escalonamento Multidimensional

O escalonamento multidimensional é uma técnica da análise multivariada que tem como objetivo a redução da dimensionalidade dos dados de modo que qualquer distorção

causada nessa redução seja minimizada, isto é, que as similaridades ou dissimilaridades originais sejam preservadas. Entretanto, nem sempre é possível corresponder exatamente essas similaridades ou dissimilaridades, fazendo com que o escalonamento multidimensional busque encontrar configurações em uma dimensão menor, mas de modo que a correspondência seja a mais próxima possível da original. Em geral, essa redução é feita para a dimensão igual a 2 ou 3.

Dentro do escalonamento multidimensional existem dois tipos de métodos associados: o escalonamento não métrico, usado quando os dados de entrada são oriundos de classificação entre pares de elementos, ou seja, são variáveis qualitativas, e portanto, não assumem distância matemática, e o escalonamento métrico, usado quando os dados de entrada são variáveis quantitativas oriundas da similaridade ou dissimilaridade entre pares de elementos.

No presente trabalho será tratado apenas o escalonamento multidimensional métrico.

2.3.1 Escalonamento Multidimensional Métrico

O escalonamento multidimensional do tipo métrico é aplicado no cenário no qual se tem um conjunto de vetores $\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_n \in \mathbb{R}^p$, e deseja-se reduzir para dimensão q, $q < p$, obtendo-se um conjunto de vetores $\mathbf{y}_1; \mathbf{y}_2; \ldots; \mathbf{y}_n \in \mathbb{R}^q$. O método é executado com base nas medidas de proximidade, sendo no presente trabalho utilizada a medida de dissimilaridade.

O primeiro passo para realização deste método é calcular uma distância entre os vetores na dimensão p, denotada por $d(\mathbf{x}_i; \mathbf{x}_j)$, e então encontrar um conjunto de vetores $\mathbf{y}$ com a distância entre as observações, $d^q(\mathbf{y}_i; \mathbf{y}_j)$, de forma que $d(\mathbf{x}_i; \mathbf{x}_j) \approx d^q(\mathbf{y}_i; \mathbf{y}_j)$. A distância calculada em $\mathbb{R}^p$ não precisa ser do tipo Euclidiana, mas caso seja, existe uma configuração q dimensional de forma que para alguns q, $d(\mathbf{x}_i; \mathbf{x}_j) = d^q(\mathbf{y}_i; \mathbf{y}_j)$, ou seja, a distância entre os vetores $\mathbf{x}_i$ e $\mathbf{x}_j$ na dimensão p é igual a distância na dimensão reduzida q.

Seja $\mathbf{D}_{n\ n}$ a matriz de distância Euclidiana derivada da matriz original dos dados. Para a realização deste método, após obter a distância Euclidiana, o segundo passo é obter uma matriz $\mathbf{B}_{n\ n}$, de acordo com o Teorema 2.1, na qual serão extraidos os valores das novas coordenadas, $\mathbf{y} \in \mathbb{R}^q$.

Teorema 2.1. [2] *Seja a matrix* $\mathbf{A}_{n\ n}$*, sendo cada elemento representado por* $a_{ij} = -\frac{1}{2}d_{ij}^2$*. De ne-se*

$$\begin{aligned} b_{ij} &= a_{ij} - \bar{a}_{i:} - \bar{a}_{:j} + \bar{a}_{::} \\ &= -\frac{1}{2}[d_{ij}^2 - d_{i:}^2 - d_{:j}^2 + d_{::}^2] \end{aligned}$$

sendo,

$$d_{i:}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 \quad ; \quad d_{:j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 \quad e \quad d_{::}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 \ .$$

tal que $\mathbf{B}_{n\ n}$ *pode ser expressa da seguinte forma:*

$$\begin{aligned} \mathbf{B}_{n\ n} &= \mathbf{I}_{n\ n} - \frac{1}{n}\mathbf{1}_{n\ 1}\mathbf{1}_{1\ n}^{>}\ \mathbf{A}_{n\ n}\ \mathbf{I}_{n\ n} - \frac{1}{n}\mathbf{1}_{n\ 1}\mathbf{1}_{1\ n}^{>} \\ &= -\frac{1}{2}\ \mathbf{I}_{n\ n} - \frac{1}{n}\mathbf{1}_{n\ 1}\mathbf{1}_{1\ n}^{>}\ \mathbf{D}_{n\ n}^2\ \mathbf{I}_{n\ n} - \frac{1}{n}\mathbf{1}_{n\ 1}\mathbf{1}_{1\ n}^{>} \ : \end{aligned} \tag{2.2}$$

Entao, $\mathbf{D}_{n\ n}$ *e Euclidiana se, e somente se,* $\mathbf{B}$ *e positiva semide nida.*

sendo, $\mathbf{I}_{n\ n}$ é a matriz identidade, $\mathbf{1}_{n\ 1}$ e $\mathbf{1}_{1\ n}^{>}$ é uma matriz de números 1 e $D_{n\ n}^2$ é a matriz de distâncias dos dados na dimensão p, $D_{n\ n}$, ao quadrado.

Após obter a matriz $\mathbf{B}_{n\ n}$, são extraídos seus autovalores e depois os autovetores correspondentes. Assim, as novas coordenadas, $\mathbf{y}\ 2\ \mathbb{R}^q$, são obtidas da seguinte forma

$$\mathbf{y}_i = v_i \sqrt{\lambda_i}$$

sendo λ_i os autovalores não nulos e v_i os autovetores associados aos respectivos autovalores.

2.3.2 Qualidade do Ajuste

Após a redução de dimensionalidade, para determinados valores de q, pode não ser possível ter uma uma perfeita concordância entre as distâncias originais com as obtidas na nova dimensão sendo necessário verificar a qualidade do ajuste. Utiliza-se duas medidas para tal: *Stress* (*Standardized Residual Sum of Squares*) e coeficiente de mensuração.

[2] A demonstracao do teorema e a formacao das novas coordenadas mais detalhadas podem ser encontradas no Anexo 1 desta monogra a ou no livro Mingoti (2017) pagina 204.

2.3.2.1 *Stress*

O *Stress* foi proposto por Kruskal (1964) e indica o quanto a ordenação das distâncias na dimensão q está de acordo com a ordenação das distâncias na dimensão p original. A medida *Stress* é definida como:

$$Stress(q) = \left[\frac{\sum_{i<j} (d_{ij} - d_{ij}^q)^2}{\sum_{i<j} (d_{ij})^2} \right]^{\frac{1}{2}}; \tag{2.3}$$

sendo d_{ij} as distâncias originais entre os elementos i e j , e d_{ij}^q as distâncias na nova dimensão q.

Para a verificação da qualidade do ajuste, Kruskal (1964) sugere o critério descrito na Tabela 1. Quanto maior for o valor do *Stress*, pior será o ajuste.

Tabela 1: Critério Stress

Stress(%)	Qualidade do Ajuste
20	Ruim
10	Razoável
5	Bom
2,5	Excelente
0	Perfeito

2.3.2.2 Coeficiente de Mensuração

A melhor representação na dimensão reduzida q é dada pelos autovalores, $_i$, de B correspondentes aos maiores autovetores do mesmo. Para verificar a adequação da representação na dimensão q pode ser utilizada a seguinte equação:

$$P_q = \frac{\sum_{i=1}^{} \ _i}{\sum_{i=1}^{} \ _i}: \tag{2.4}$$

O resultado da Equação (2.4) acima de 0,8 é considerado que as observações estão bem representadas na nova dimensão.

2.4 Análise de Agrupamento

A análise de agrupamento é um conjunto de técnicas multivariadas cujo principal objetivo é classificar e agrupar objetos de acordo com as suas características, de forma que dentro de cada grupo exista homogeneidade, e que entre os grupos exista heterogeneidade. O propósito é minimizar a heterogeneidade dentro dos grupos e, consequentemente, maximiza-lá entre os grupos.

O foco da análise de agrupamento está na comparação de objetos com base no conjunto variáveis escolhidas pelo pesquisador. Segundo Hair et al. (2009), é a única técnica que não estima a variação empiricamente, mas usa as variáveis conforme especificado. Isso faz com que a etapa de decisão das variáveis pelo pesquisador seja uma etapa crítica. Um exemplo de aplicação da análise de agrupamento é na área geográfica para classificação de cidades e regiões de acordo com suas características físicas, econômicas e/ou demográficas.

A realização dessa técnica é feita com base nas medidas de proximidades, ou seja, medidas de similaridade ou dissimilaridade. O presente trabalho adota as medidas de dissimilaridade utilizando a distância Euclidiana. A análise de agrupamento é dividida em dois tipos: métodos hierárquicos e métodos não hierárquicos, sendo cada um explicado a seguir.

2.4.1 Métodos Hierárquicos

Os métodos hierárquicos têm como principal característica fornecer várias possibilidades no número de agrupamentos, não sendo necessário fornecer este número inicialmente, e sim determiná-lo a partir do próprio processo de agrupamento. Outra característica desse método é que os objetos não podem ser alterados de um grupo para outro. Dentro do método hierárquico, os objetos podem ser agrupados de duas maneiras, são elas:

Método Aglomerativo: Nesse método cada objeto é considerado inicialmente como um único grupo, que sofre sucessivamente uma série de fusões com outros grupos, até que no final todos os objetos estejam em um único agrupamento. Dentro desse método existem algumas formas para realizá-lo são elas: encadeamento simples, encadeamento completo, encadeamento médio, método de Ward. Esses se diferenciam pela forma de se calcular a distância entre os grupos.

Método Divisivo: Nesse método inicialmente é considerado um único grupo com todos os objetos e, em seguida ele é divido sucessivamente até que no final tenha

um grupo com apenas um objeto.

2.4.2 Métodos Não Hierárquicos

O método não hierárquico tem como característica a necessidade do pesquisador fornecer um número inicial de agrupamento, possuindo maior flexibilidade visto que os objetos podem ser transferidos de um grupo para outro. Uma outra caractéristica desse método é ser usualmente mais eficiente em conjunto de dados com um número maior de observações.

Para a aplicação do método não hierárquico existem algumas técnicas associadas, sendo o método *k-means* o principal deles e o único tratado nesse trabalho. Segundo Hair et al. (2009), é tão comumente usado que alguns utilizam o termo para se referir aos métodos não hierárquicos, de modo geral.

Para a aplicação do método, dado um conjunto de objetos $x_1; x_2; :::; x_n \ 2 \ R^p$ e o número k de grupos, o método atua particionando os n objetos nos k grupos, de acordo com os seus centróides, de forma que a soma dos quadrados dentro dos grupos seja minimizada, ou seja, deseja-se minimizar:

$$WSS = \sum_{l=1}^{k} \sum_{x_i 2 G_l} (x_i \quad {}^{(l)})^2$$

sendo ${}^{(l)}$ a média dos objetos pertencentes ao l-ésimo grupo.

Os passos do algoritmo são:

1. Especifique o número k e escolha os centróides iniciais;
2. Cada objeto é atribuído a um centróide mais próximo;
3. Após cada objeto ter sido atribuído, k centróides são recalculados;
4. Os passos 2 e 3 são repetidos até que os grupos não sejam mais modificados.

2.4.3 Número de Grupos

Como nos métodos não hierarquicos é necessário informar um número inicial de agrupamento, no momento de fazer a análise, pode ser feita algumas simulações com número diferentes de agrupamentos. Para definir qual é o número k de grupos, Calinski e Harabasz (1974) propuseram o Pseudo F. O Pseudo F considera a soma dos quadrados dentro

dos grupos, WSS, e a soma dos quadrados entre os grupos, BSS, quanto maior o valor do Pseudo F, melhor será a partição dos dados dentro do agrupamento k. Sua fórmula é expressa por:

$$Pseudo\,F = \frac{\frac{BSS}{(k \quad 1)}}{\frac{WSS}{(n \quad k)}}; \tag{2.5}$$

sendo n o número de objetos dentro de todos grupos.

2.5 t-SNE

O t-SNE é uma técnica não linear de visualização de dados derivada do SNE, Hinton e Roweis (2002). O SNE, $Stochastic\ Neighbor\ Embedding$, é aplicado, primeiramente, calculando a distância Euclidiana entre os pares de dados na alta dimensão, e depois essas distâncias são convertidas em probabilidades condicionais representando as similaridades, ou seja, quanto mais próximos forem x_i e x_j $8\ i; j = 1; : : : ; n$, maior será a probabilidade condicional, $p_{j|i}$, associada a eles. Dessa forma, dado dois pontos vizinhos, x_i e x_j, eles são associados a distribuição de probabilidade Normal centrada em x_i. A probabilidade condicional $p_{j|i}$ é representada por

$$p_{j|i} = \frac{\exp(\ -jjx_i \quad x_jjj^2{=}2\ _i^2)}{\sum_{k \neq i} \exp(\ -jjx_i \quad x_kjj^2{=}2\ _i^2)}; \tag{2.6}$$

sendo $_i^2$ é a variancia da distribuição Normal que está centrada em x_i. Quando a dimensão dos dados é reduzida, obtém-se um novo conjunto de dados, e assim, temos y_i representando x_i na dimensão reduzida, e analogamente à alta dimensão também é possível representar a probabilidade condicional na dimensão reduzida. A probabilidade condicional é representada por $q_{j|i}$ e com variância igual a $\frac{1}{2}$. Logo, $q_{j|i}$ é representada por

$$q_{j|i} = \frac{\exp(\ -jjy_i \quad y_jjj^2)}{\sum_{k \neq i} \exp(\ -jjy_i \quad y_kjj^2)}: \tag{2.7}$$

O SNE minimiza a soma de divergência de KullBack-Leiber usando o método gradiente descendente para sua otimização, ou seja, o intuito é diminuir a perda de informação quando Q_i, a distribuição probabilidade condicional de todos os pontos de dados dado y_i, é usada para se aproximar do valor de P_i, a distribuição de probabilidade condicional de

todos os pontos de dados dado x_i. A sua função de custo é expressa por

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}; \tag{2.8}$$

sendo $p_{j|i}$ a probabilidade condicional da similidade entre dois pontos x_i e x_j e $q_{j|i}$ a probabilidade condicional da similidade entre dois pontos y_i e y_j na dimensão reduzida.

Como a divergência de KullBack-Leiber não é simétrica, há um grande custo para modelar pontos que são bastante separados na alta dimensão para representar pontos de dados que são próximos na dimensão mais baixa. O método *SNE* acaba não sendo muito bom pois sua função de custo é difícil de otimizar.

Por consequência, o *t-SNE* tem como intenção reduzir os problemas do *SNE*, empregando uma função de custo mais simples e a distribuição *t-Student* com 1 grau de liberdade para calcular a similaridade entre os pontos no espaço de baixa dimensão, melhorando o problema da otimização. A distribuição *t-Student* é utilizada porque possui caudas mais pesadas, sendo adequada para atenuar os problemas de otimização do SNE. A função de custo utilizada é a versão simétrica da usada no *SNE*, ou seja, existe uma propriedade que garante que $p_{ij} = p_{ji}$ e $q_{ij} = q_{ji}$, $\forall\, i; j = 1; \ldots; n$. A probabilidade conjunta na alta dimensão é representada por

$$p_{ij} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq l} \exp(-||x_k - x_l||^2/2\sigma_i^2)}. \tag{2.9}$$

E a probabilidade conjunta na dimensão reduzida, utilizando a distruição *t-Student* com 1 grau de liberdade, é representada por

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}}. \tag{2.10}$$

Como a função de custo empregada é simétrica, ela fica expressa da seguinte forma

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{2.11}$$

Dessa forma, o gradiente utilizado é representado da seguinte forma

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}. \tag{2.12}$$

Em Maaten e Hinton (2008) [3], o método é implementado usando, primeiramente, a análise de componentes principais e reduzindo a dimensão dos dados para 30. Após, é implementando efetivamente o algoritmo do *t-SNE* no qual converte a dimensão de 30 para 2. Ao implementar o algoritmo é definido um valor de perplexidade que caracteriza a quantidade de vizinhos mais próximos que serão considerados, o artigo original utiliza a perplexidade igual a 40. Outro parâmetro que pode ser incluído é o fator de exagero, *exaggeration factor*, que é geralmente usado na fase inicial para ampliar o espaço entre os pontos e assim encontrar vizinhos mais próximos. O artigo referenciado acima utiliza o fator de exagero igual a 4 para as 50 primeiras interações.

O presente trabalho implementou a técnica de dois modos, um conforme o artigo, Maaten e Hinton (2008), e o outro modo utilizando o escalonamento multidimensional no lugar da análise de componentes principais afim de verificar se o resultado será parecido.

2.6 Exemplo sobre o desenvolvimento dos países

Para demonstração prática das técnicas expostas nas seções anteriores, foi adaptado um exemplo de Mingoti (2017) que apresenta quatro indicadores de desenvolvimento de 21 países conforme mostra a Tabela 2.

[3] O artigo pode ser encontrado no site: *h*https://lvdmaaten.github.io/tsne/*i*

Tabela 2: Índices de desenvolvimento dos países

Países	Expectativa de vida	Educação	PIB	Estabilidade política
Reino Unido	0,88	0,99	0,91	1,10
Austrália	0,90	0,99	0,93	1,26
Canadá	0,90	0,98	0,94	1,24
Estados Unidos	0,87	0,98	0,97	1,18
Japão	0,93	0,93	0,93	1,20
França	0,89	0,97	0,92	1,04
Cingapura	0,88	0,87	0,91	1,41
Argentina	0,81	0,92	0,80	0,55
Uruguai	0,82	0,92	0,75	1,05
Cuba	0,85	0,90	0,64	0,07
Colômbia	0,77	0,85	0,69	-1,36
Brasil	0,71	0,83	0,72	0,47
Paraguai	0,75	0,83	0,63	-0,87
Egito	0,70	0,62	0,60	0,21
Nigéria	0,44	0,58	0,37	-1,36
Senegal	0,47	0,37	0,45	-0,68
Serra Leoa	0,23	0,33	0,27	-1,26
Angola	0,34	0,36	0,51	-1,98
Etiópia	0,31	0,35	0,32	-0,55
Moçambique	0,24	0,37	0,36	0,20
China	0,76	0,80	0,61	0,39

O método do Escalonamento Multidimensional foi aplicado a fim de reduzir a dimensão $p = 4$ para a $q = 2$. Após aplicação do método, foi construído um gráfico de percepção na nova dimensão, conforme mostra a Figura 3. Pode-se notar que países com índices de desenvolvimento semelhantes ficaram mais próximos, indicando, visualmente, que as distâncias originais foram preservadas. Alguns países como Austrália, Canadá, Estados Unidos, França, Japão e Reino Unido ficaram tão próximos que fica difícil identificar seus nomes na Figura 3. Para verificar a qualidade do ajuste foram obtidos os seguintes valores: $Stress = 0{,}0069$ e $P_q = 0{,}9943$, indicando um ajuste perfeito.

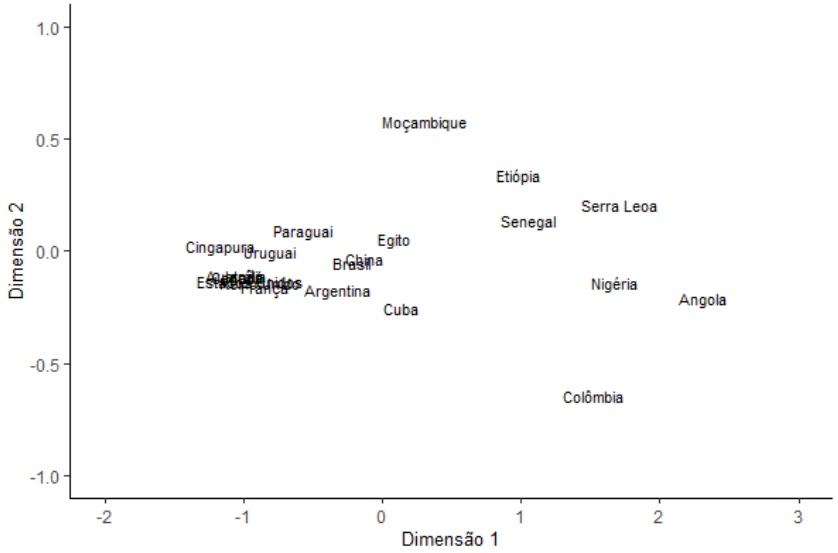
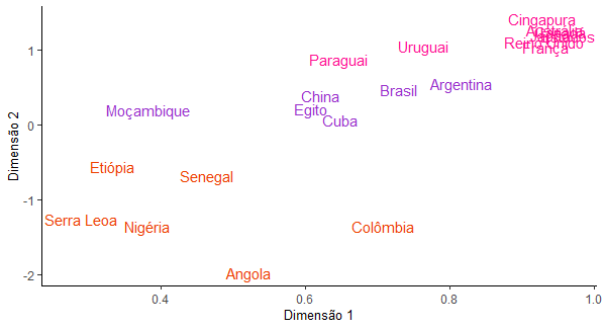
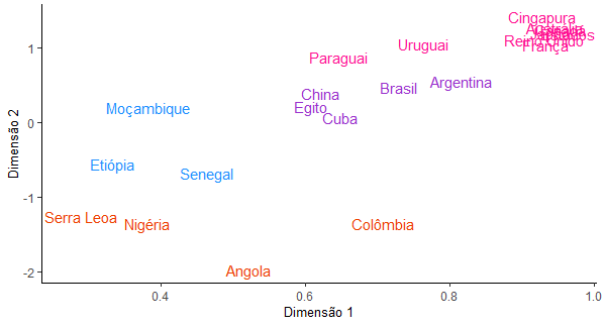


Figura 3: Gráfico do Escalonamento Multidimensional - exemplo países

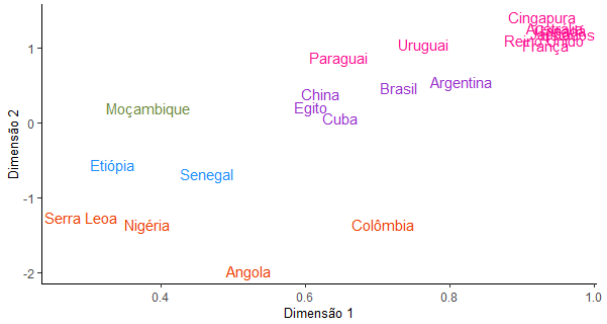
Foi aplicado também o método não hierárquico, *k-means*, na base de dados com as 4 variáveis, afim de verificar o comportamento no agrupamentos dos países, e também se vai ratificar a análise feita com base na Figura 3. Para tal, foi realizada a análise de agrupamento para três diferentes tamanhos de grupos, ou seja, com 3, 4 ou 5 grupos, conforme o resultado pode ser visto na Figura 4. Novamente, observe que Austrália, Canadá, Estados Unidos, França, Japão e Reino Unido ficaram muito próximos e nos três agrupamentos propostos permaneceram no mesmo grupo.



(a) k=3



(b) k=4



(c) k=5

Figura 4: Gráfico da Análise de Agrupamento - exemplo países

A Tabela 3 apresenta os valores do Pseudo F referente aos agrupamentos da Figura 4. Pode-se notar que agrupamento com 4 grupos foi o que obteve o maior valor, sendo então o mais adequado para os dados.

Tabela 3: Pseudo *F*

Número de grupos	BSS	WSS	Pseudo *F*
3	22,12	2,96	67,27
4	23,22	1,85	71,12
5	23,68	1,39	68,14

Por último, foi aplicada a técnica *t-SNE* aos dados para verificar o comportamento dos países e averiguar se o resultado é semelhante ao escalonamento multidimensional e *k-means*. A Figura 5 apresenta esse resultado, pode-se notar que a disposição dos países é semelhante a Figura 4, dessa forma, os países com características semelhantes ficaram próximos. A perplexidade utilizada foi igual a 6.

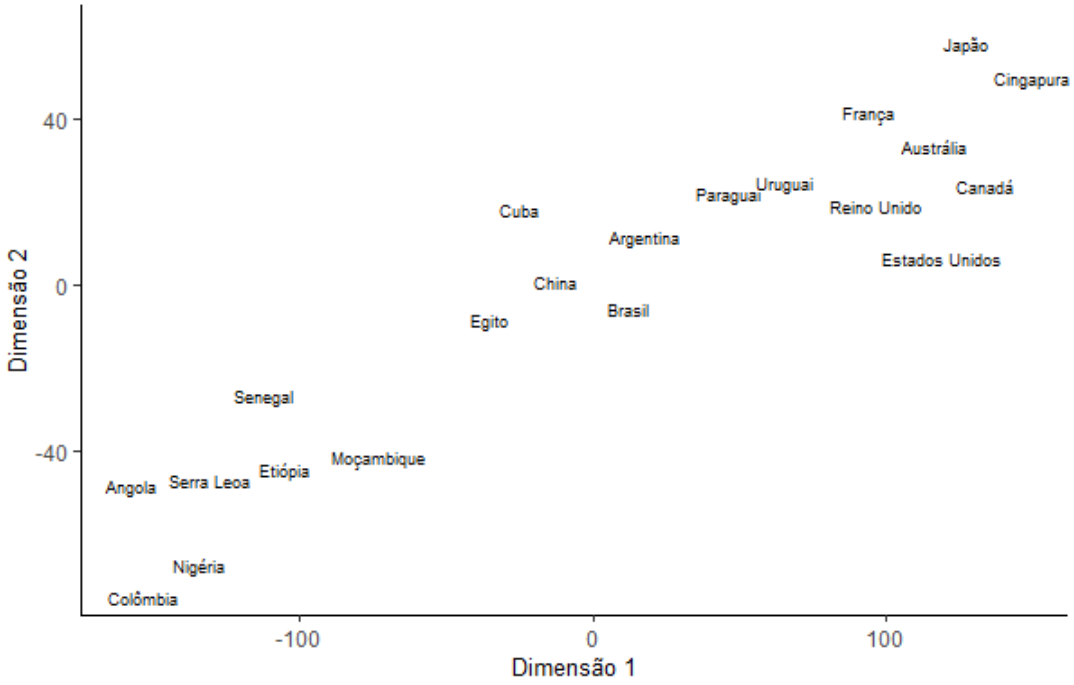


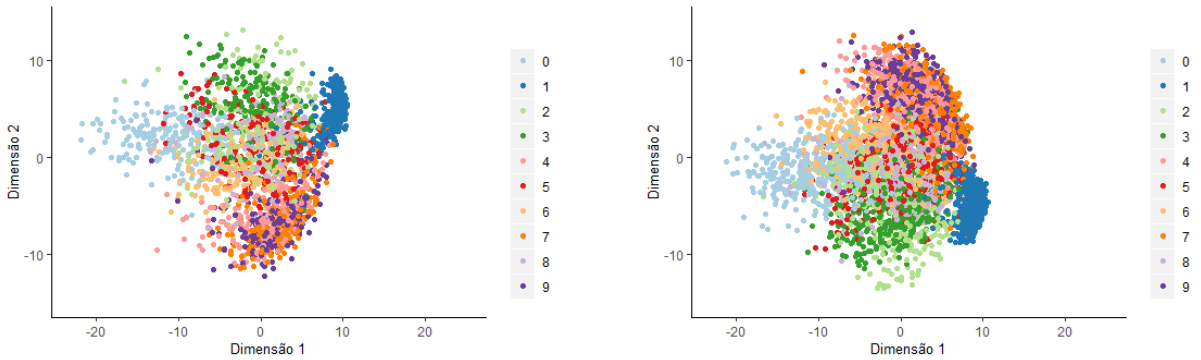
Figura 5: Gráfico do *t-SNE* - exemplo países

3 Análise dos Resultados

Para seguir com o objetivo deste trabalho, foi aplicado os métodos Escalonamento Multidimensional, *K-means* e *t-SNE* ao conjunto de imagens *MNIST* na base de imagens e rótulos de teste. Devido à problemas de processamento computacional, não foi possível utilizar a base de teste com as 10.000 observações. Para contornar essa adversidade, foi realizada uma amostragem com 2.000 e 4.000 observações.

3.1 Escalonamento Multidimensional

O método do Escalonamento Multimensional métrico foi aplicado ao conjunto de imagens *MNIST* com o intuito de ser efetuada uma redução para duas dimensões, ou seja, $q = 2$. A Figura 6 apresenta o resultado gráfico do Escalonamento Multidimensional. Analisando visualmente, pode-se observar que apesar de alguns dígitos ficarem mais proximos como, principalmente, o dígito 1, não existe um padrão de forma que todos os outros dígitos estejam mais próximos dos seus semelhantes sem se sobrepor a outros dígitos.



(a) n=2.000 (b) n=4.000

Figura 6: Gráfico do Escalonamento Multidimensional

A Tabela 4 mostra a qualidade deste ajuste de acordo com o tamanho da amostra. Pode-se observar que o *Stress* ficou acima de 20% indicando uma qualidade ruim, assim

como os coeficientes de mensuração ficaram abaixo de 0,8 reforçando que o ajuste não ficou bom, corroborando com a análise gráfica.

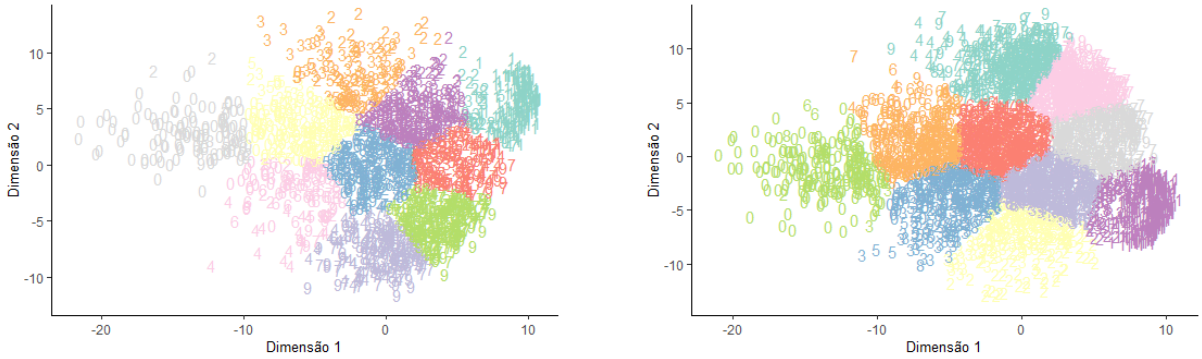
Tabela 4: Medidas de qualidade do ajuste *MNIST*

n	Stress	P_q
2.000	0,2815	0,1776
4.000	0,2843	0,1744

3.2 Escalonamento Multidimensional e *K-means*

Após executar no passo anterior o método de redução de dimensionalidade, aplicou-se, na base com duas dimensões, a Análise de Agrupamentos a fim de verificar se os dígitos iguais são alocados no mesmo grupo. Como a base de dados é formada por dígitos de 0 a 9, o método escolhido para foi o não hierárquico no algoritmo *k-means*, visto que o número k de grupos é conhecido, ou seja, $k = 10$.

A Figura 7 mostra o agrupamento dos dígitos de acordo com os 10 grupos. Analisando visualmente, parece que os grupos não ficaram tão homogêneos, com exceção dos grupos onde estão os dígitos 0 e 1 nos quais parecem terem tido uma classificação mais assertiva. Porém, como os dígitos se sobrepõem, dificulta uma interpretação mais efetiva. Para melhorar a interpretação, foi plotado um outro gráfico, Figura 8, com o intuito de mostrar a quantidade de dígitos por grupo.

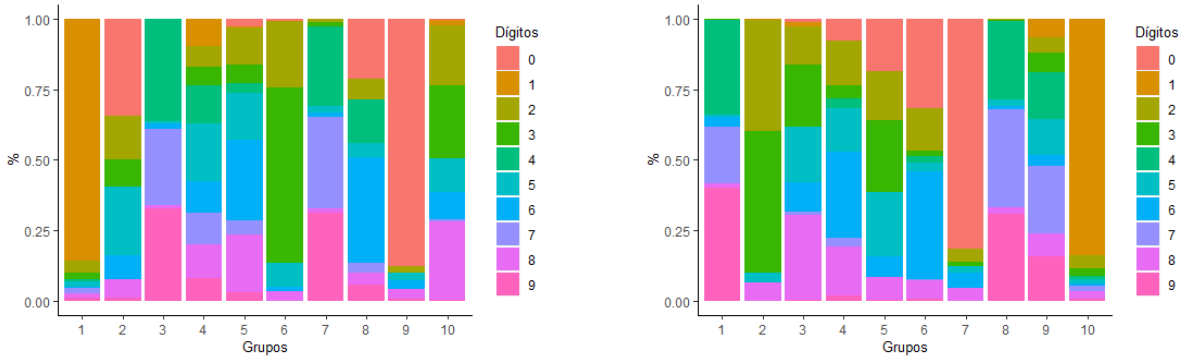


(a) n=2.000 (b) n=4.000

Figura 7: Gráfico da Análise de Agrupamento

Analisando a Figura 8, percebe-se que na Subfigura 8(a) o primeiro grupo tem uma incidência maior do dígito 1 e o nono grupo do dígito 0, enquanto que na Subfigura 8(b) o mesmo ocorre com o décimo e sétimo grupo, respectivamente. Os demais grupos ficam com a quantidade de diferentes dígitos numa proporção mais semelhante, tendo

um destaque para as Subfiguras 8(a) e 8(b) nos sexto e segundo grupo, respectivamente, apresentando uma ocorrência maior do dígito 3. Dessa forma, a maioria dos grupos ficaram heterogêneos, caracterizando que o método não foi eficaz para a classificação de dígitos.



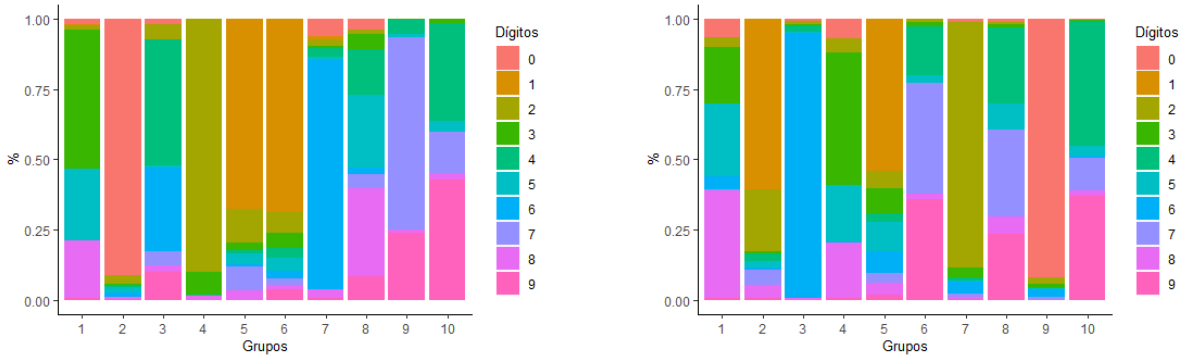
(a) n=2.000 (b) n=4.000

Figura 8: Percentual de dígitos por grupo - escalonamento multidimensional e *k-means*

3.3 *K-means*

O método de Análise de Agrupamentos também foi aplicado diretamente na base *MNIST*, sem a prévia redução da dimensionalidade, a fim de verificar seu desempenho na classificação de dígitos em uma base de alta dimensão. Assim como foi feito anteriormente, o método escolhido foi o não hierárquico no algoritmo *k-means*, visto que o número k de grupos é conhecido, ou seja, $k = 10$. A Figura 9 foi plotada para melhorar a visualização e interpretação da quantidade de dígitios por grupo.

Análisando a Figura 9, pode-se notar que na Subfigura 9(a) alguns grupos possuem uma ocorrência maior de dígitos específicos como, por exemplo, o segundo grupo com o dígito 0, o quarto grupo com o dígito 2, o quinto e sexto grupo com o dígito 1 e o sétimo grupo com o dígito 6. Enquanto a Subfigura 9(b) possui o terceiro grupo com uma ocorrência maior do dígito 6, sétimo grupo com o dígito 2 e o nono grupo com o dígito 0. Apesar disso, os demais grupos ainda continuam heterogêneos, caracterizando que mesmo sendo aplicada sozinha a técnica não é eficiente para classificação dos dígitos.



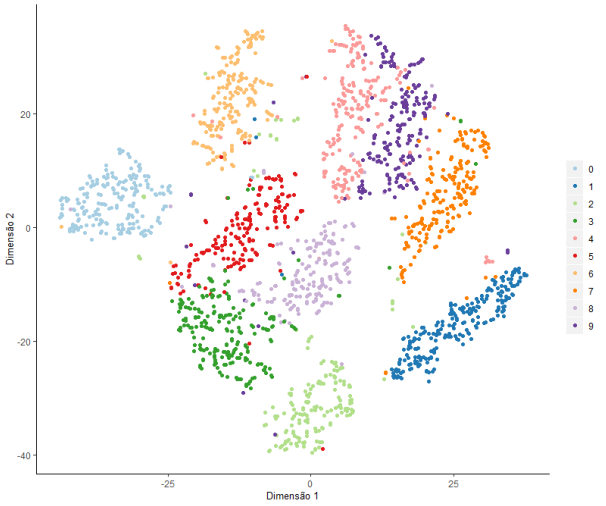
(a) n=2.000 (b) n=4.000

Figura 9: Percentual de dígitos por grupo - *k-means*

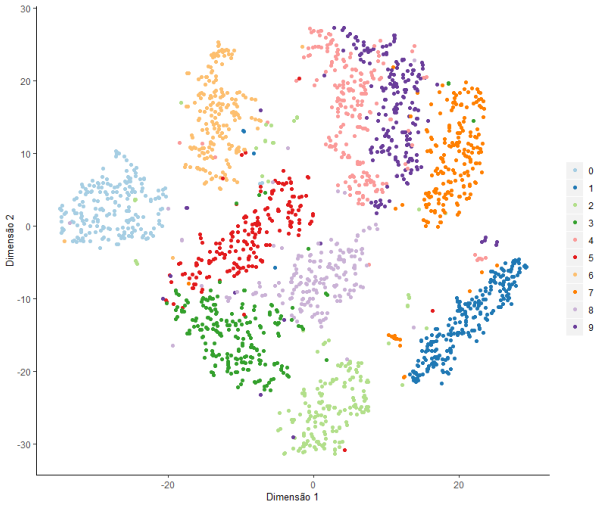
3.4 *t-SNE* com Análise de Componentes Principais

Nessa subseção será apresentada a visualização do *t-SNE* com análise de componentes principais. Foram feitas algumas simulações alterando o valor da perplexidade para encontrar uma melhor visualização. Será apresentada a visualização com três valores de perplexidade, sendo a Figura 10 representando as perplexidades referente a amostra de 2.000 observações e a Figura 11 representando as perplexidades referente a amostra de 4.000 observações.

Pode-se observar na Figura 10 que as Subfiguras 10(a) e 10(b) apresentam uma visualização semelhante em relação a visualização dos dígitos, visto que os dígitos iguais estão posicionados mais próximos uns dos outros formando uma espécie de agrupamento natural. A Subfigura 10(c) apresenta uma pior visualização em relação as anteriores por ter mais dígitos afastados dos seus dígitos semelhantes. Na Figura 11 referente a amostra com 4.000 observações, observa-se que a Subfigura 11(a), com a perplexidade 25, apresenta uma melhor visualização dos dados, a Subfigura 11(b) tem uma visualização um pouco pior em relação ao dígito 5, enquanto a Subfigura 11(c) apresenta mais dígitos flutuantes entre os grupos dos outros dígitos e o dígito 2 não ficou com uma boa visualização já que possui dois grupos dele distante um do outro.



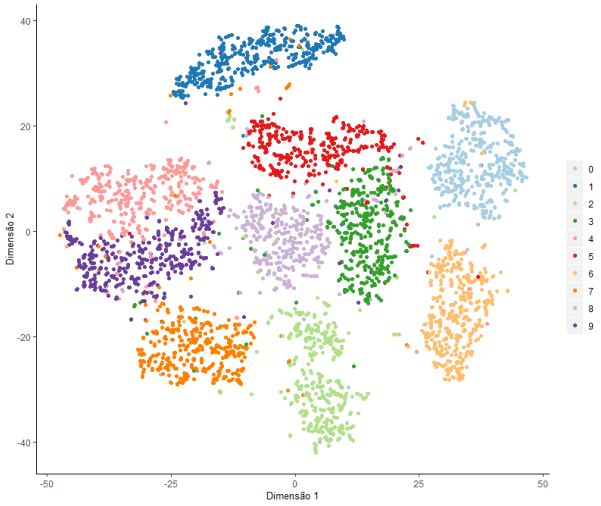
(a) Perp=25



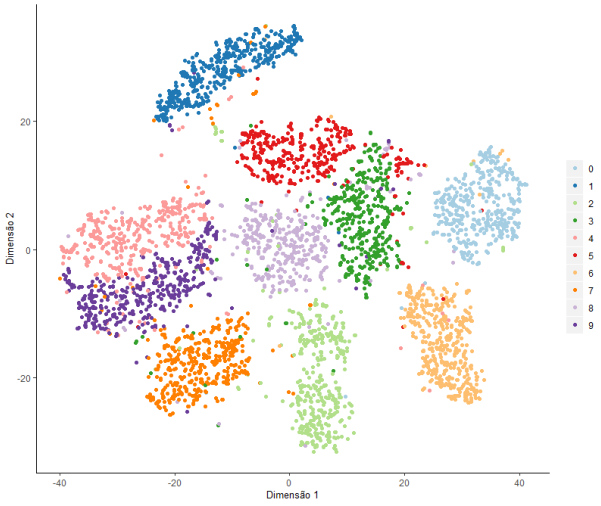
(b) Perp=40

(c) Perp=60

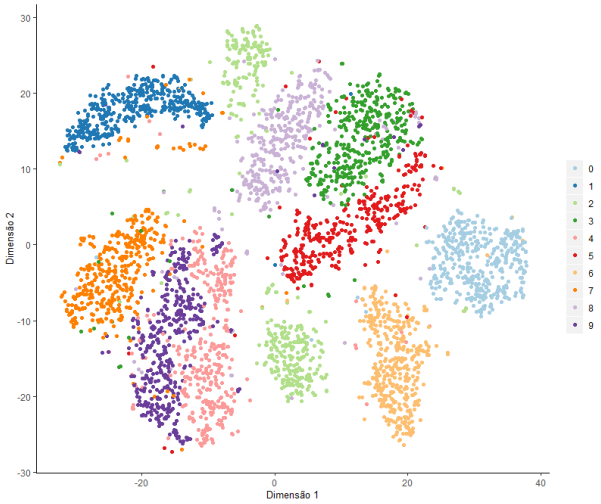
Figura 10: Gráfico do *t-SNE* com análise de componentes principais - 2.000 observações



(a) Perp=25



(b) Perp=40



(c) Perp=60

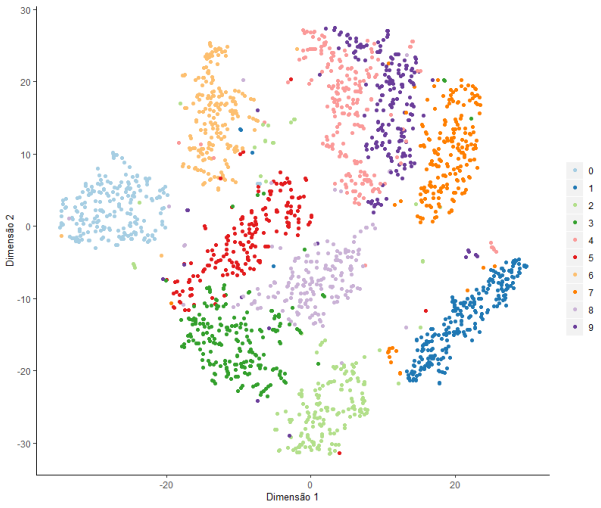
Figura 11: Gráfico do *t-SNE* com análise de componentes principais - 4.000 observações

3.5 *t-SNE* com Escalonamento Multidimensional

Nessa subseção será apresentada a visualização do *t-SNE* com escalonamento multidimensional. Para fim de comparação será apresentado o mesmo valor de perplexidade da Seção 3.4, dessa forma, a Figura 12 representa as perplexidades referente a amostra de 2.000 observações e a Figura 13 representa as perplexidades referente a amostra de 4.000 observações

Analisando a Figura 12 pode-se notar que as Subfiguras 12(a) e 12(b) apresentam uma visualização semelhante, e a Subfigura 12(c) apresenta uma visualização com dígitos mais espalhados. Na Figura 13, a Subfigura 13(a) apresenta uma visualização melhor dos dígitos com os grupos, e a Subfigura 13(c), com perplexidade igual a 60, não apresenta uma boa visualização.

(a) Perp=25

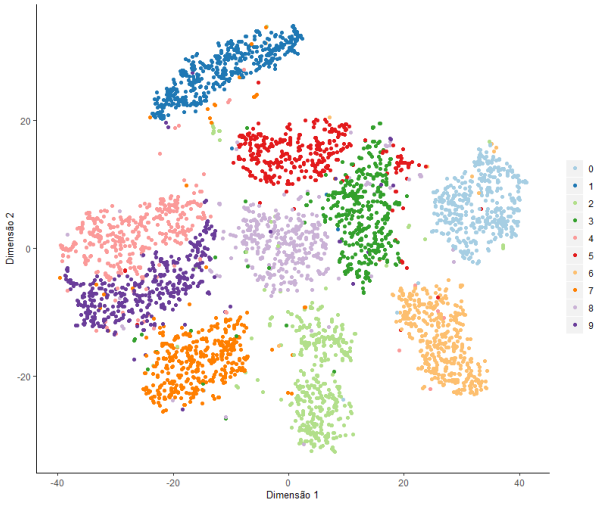


(b) Perp=40

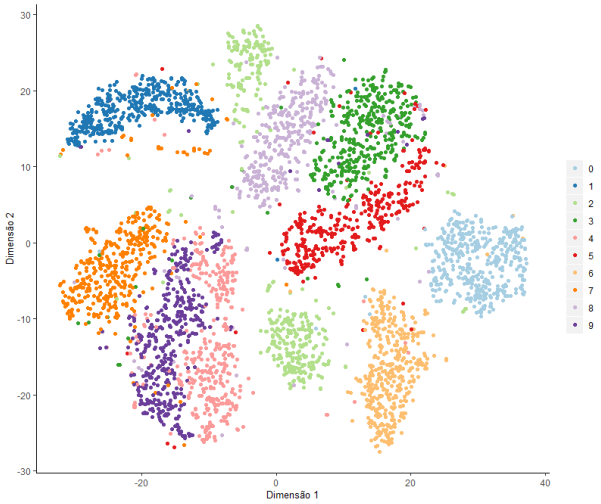
(c) Perp=60

Figura 12: Gráfico do *t-SNE* com escalonamento multidimensional - 2.000 observações

(a) Perp=25



(b) Perp=40



c

(c) Perp=60

Figura 13: Gráfico do *t-SNE* com escalonamento multidimensional - 4.000 observações

3.5.1 Comparação entre t-SNE com Análise de Componentes Principais e Escalonamento Multidimensional

Nessa Subseção será feita uma comparação entre as melhores visualizações obtidas nas Seções 3.4 e 3.5, a m de conseguir uma melhor observação se há alguma diferença signi cativa entre usar o t-SNE com análise de componentes principais ou escalonamento multidimensional. Serão comparadas as visualizações com perplexidade igual a 25 nas amostras de 2.000 e 4.000 observações.

Analisando as Figuras 14 e 15 pode-se notar que não há uma diferença relevante entre primeiro reduzir a dimensão para 30 com análise de componentes principais ou com escalonamento multidimensional, pois o resultado nal é extremamente semelhante.

(a) Análise de Componentes Principais (b) Escalonamento Multidimensional

Figura 14: Comparação entre Análise de Componentes Principais e Escalonamento Multidimensional com 2.000 observações

(a) Análise de Componentes Principais (b) Escalonamento Multidimensional

Figura 15: Comparação entre Análise de Componentes Principais e Escalonamento Multidimensional com 4.000 observações

3.6 Análise extra

Para poder analisar o comportamento do t-SNE em um outro cenário, foi escolhido o conjunto de dados Extended MNIST, EMNIST, que consiste em uma extensão da base MNIST mas que agora além de possuir dígitos manuscritos, possui letras manuscritas. O conjunto de dados foi elaborado por Cohen et al. (2017) e consiste numa coleção de 5 conjuntos de dados, variando entre conjunto de dados com dígitos e letras manuscritas balanceados e não balanceados, dígitos e letras manuscritas sem juntar nenhuma classe, somente de letras manuscritas e somente dígitos manuscritos.

O presente trabalho usa o conjunto de dados com somente letras manuscritas separadas em 26 classes, dessa forma, as letras maiúsculas e minúsculas estão alocadas dentro da mesma classe. O conjunto consiste em 28.800 observações e sua visualização é similar a do MNIST visto na Seção 2.1, ou seja, cada letra está em uma imagem de tamanho fixo 28x28 pixels em escala de cinza. Para mostrar como é um arquivo de imagem a Figura 16 apresenta as primeiras 25 letras do arquivo de imagem, e a Figura 17 mostra a letra destacada na Figura 16. Observa-se que a letra na Figura 17 é formada pelos números de 0 a 255, indicando a escala de cinza.

Figura 16: Letras de um arquivo de imagem

Figura 17: Matriz do arquivo de imagem

Devido a problemas computacionais, para analisar a base de dados completa, foi feita uma amostragem com 4.000 observac-oes. As Sub guras 18(a) e 18(b) apresentam a visualizac-ao dot-SNE. Observa-se que as duas sub guras apresentam visualizac-oes quase idênticas, corroborando com as analises feitas nas Sec-oes 3.4 e 3.5. Pode-se observar que algumas classes caram mais concentradas como por exemplo, Mm, Oo, Pp, Ss, Vv, Ww, Xx, Zz, por suas letras maiusculas e minusculas serem escritas de forma muito semelhante. Chama atenc-ao tambem as classes Ii e Ll por ter uma concentrac-ao das suas letras se sobrepondo, muito por suas letras serem escritas de forma parecidas. Classes como por exemplo, Dd e Jj que a letra maiuscula e escrita diferente da letra minuscula podemos notar a presenca de dois grupos mais separados enquanto as classes Aa e Yy possuem suas letras muito espalhadas sendo difcil encontrar uma concentrac-ao dessas classes.

(a) t-SNE com Analise de Componentes Principais

(b) t-SNE com Escalonamento Multidimensional

Figura 18: Visualizac-aoEMNIST

4 Conclusões

O presente trabalho teve como objetivo iniciar um estudo na área de aprendizado de máquina, analisando uma base de dados de dígitos manuscritos. Para isso, foi feita uma análise com técnicas que são comumente difundadas na estatística, como análise de agrupamanento e escalonamento multidimensional, e uma outra análise com uma técnica mais recente chamada t-SNE.

O trabalho iniciou fazendo uma análise com o método escalonamento multidimensional que tem como objetivo fazer a redução dos dados para uma melhor visualização do comportamento desses dados. Como é uma técnica de visualização, a visualização acaba criando um agrupamento natural de forma que os dados próximos na nova dimensão são mais semelhantes. Para ratificar esse agrupamento natural formado na visualização, foi feita também uma análise de agrupamento que tem como foco agrupar objetos. Dessa forma, realizou-se uma análise das técnicas separadas e depois a junção delas para verificar se ocorria alguma diferença. Em seguida, foi feita a análise do t-SNE que originalmente é usado juntamente com a análise de componentes principais, porém, no presente trabalho, ele também foi usado com o escalonamento multidimensional, visto que são técnicas que possuem um objetivo em comum que é redução da dimensionalidade dos dados. A ideia de comparar o comportamente das técnicas na base de dados de dígitos manuscritos surge de que o t-SNE é uma técnica não linear, ou seja, tem o foco em preservar as similaridades, ao contrário do escalonamento multidimensional que tem o foco em preservar as dissimilaridades.

Após a realização de todas as análises podemos notar que t-SNE produz melhores resultados na base MNIST do que o escalonamento multidimensional e a análise de agrupamento. Com o t-SNE podemos notar de forma mais clara a formação natural de grupos com dígitos do mesmo rótulo. Dentro da esfera do t-SNE, reduzir a dimensão primeiro com análise de componentes principais, como em Maaten e Hinton (2008), ou reduzir com escalonamento multidimensional não fez diferença no resultado final, visto que as visualizações foram muito semelhantes quando comparadas com a mesmo amostra

e o mesmo valor de perplexidade.

Para rati car o comportamento da tecnica t-SNE, foi feita uma outra analise em uma base com letras manuscritas, EMNIST, que e uma base mais complexa que a base MNIST, pois possui 26 classes com as letras maiusculas e minusculas agrupadas na mesma classe. Sendo assim, apos a aplicacao do t-SNE pode-se notar que os resultados produzidos sao bons com classes espec cas, ou seja, em que a letra maiuscula e escrita de forma parecida com a minuscula. Dessa forma, para a visualizacao de dgitos manuscritos a tecnica t-SNE e uma boa alternativa e para a visualizacao de letras manuscritas a tecnica e boa para algumas classes.

Logo, como o t-SNE e voltado para a visualizacao, ele possui uma boa usabilidade para uma visualizacao inicial dos objetos, seja para identi car a estrutura dessas imagens ou veri car se elas sao separaveis, sendo entao, indicado para um estudo inicial na analise de imagens. Para uma analise mais aprimorada, em que o algoritmo vai poder ser treinado a m de prever resultados, existe a rede neural convulocional que e um tipo de rede neural utilizada para a classi cacao de imagem, sendo proposta para um trabalho futuro.

Referências

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods, v. 3:1, p. 1{27, 1974.

COHEN, G. et al. EMNIST: an extension of mnist to handwritten letters. arXiv, v. 2, 2017. Disponvel em: hhttps://arxiv.org/abs/1702.05373i. Acesso em: 25 nov. 2020.

ESTADAO. Conectada com a inclusao: saiba como a tecnologia pode abrir portas para alunos com de ciência 2019. Disponvel em: hhttp://patrocinados.estadao.com.br/dedalus/conectada-com-a-inclusao-saiba-como-a-tecnologia-pode-abrir-portas-para-alunos-com-de cien i. Acesso em: 25 nov. 2020.

HAIR, J. F. et al. Multivariate Data Analysis. 7. ed. [S.l.]: Prentice Hall, 2009.

HINTON, G.; ROWEIS, S. Stochastic neighbor embedding. In Advances in Neural Information, The MIT Press, Cambridge, MA, USA, v. 15, p. 833{840, 2002.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, v. 24, p. 417{441, 1933.

KRIJTHE, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. [S.l.], 2015. R package version 0.15. Disponvel em: hhttps://github.com/jkrijthe/Rtsne i. Acesso em: 25 nov. 2020.

KRUSKAL, J. B. Multidimensional scaling by Optimizing goodness of t to a nonmetric hypothesis. Psycometrika, v. 29, p. 1{27, 1964.

LECUN, Y.; CORTES, C.; BURGES, C. J. The MNIST database of handwritten digits 2019. Disponvel em: hhttp://yann.lecun.com/exdb/mnist i. Acesso em: 25 nov. 2020.

MAATEN, L. van der; HINTON, G. Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research, v. 9, p. 2579{2605, 2008. Disponvel em: hhttp://www.jmlr.org/papers/v9/vandermaaten08a.html i. Acesso em: 25 nov. 2020.

MINGOTI, S. A. Analise de Dados Atraves de Metodos de Estatstica Multivariada. Uma Abordagem Aplicada. 1. ed. [S.l.]: Editora UFMG, 2017.

PRESSE, F. S. ao Francisco probe a polcia de usar reconhecimento facil 2019. Disponvel em: hhttps://g1.globo.com/pop-arte/noticia/2019/05/16/san-francisco-proibe-a-policia-de-usar-reconhecimento-facial.ghtm Acesso em: 25 nov. 2020.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2019. Disponvel em: hhttp://www.R-project.org/ i. Acesso em: 25 nov. 2020.

RICHARDSON, M. W. Multidimensional psychophysics. Psychological Bulletin, v. 35, p. 659–660, 1938.

SEBER, G. A. F. Multivariate Observations [S.l.]: Wiley, 1984. ISBN 0-471-691216.

TORGERSON, W. S. Theory & Methods of Scaling. New York: Wiley, 1958. ISBN 978-0-89874-722-5.

WICKHAM, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: https://ggplot2.tidyverse.org. Acesso em: 25 nov. 2020.

WICKHAM, H. et al. dplyr: A Grammar of Data Manipulation. [S.l.], 2019. R package version 0.8.3. Disponível em: https://CRAN.R-project.org/package=dplyr. Acesso em: 25 nov. 2020.

APÊNDICE 1 – Códigos Computacionais

Os códigos a seguir mostram como visualizar a imagem e a matriz associada aos dígitos manuscritos conforme as imagens apresentadas no Capítulo 2.

Visualização das primeiras 25 imagens

```
to.read = gzfile("t10k images idx3 ubyte.gz", "rb")
readBin(to.read, integer(), size=4, n=4, endian="big")
par(mfrow=c(5,5))
par(mar=c(0,0,0,0))
for(i in 1:25)f
  m <  matrix(readBin(to.read,integer(), size=1, n=28 28,
              endian="big", signed = F), nrow=28, byrow=T)
  image(t(m)[,28:1],col = gray(255:0 / 255), axes = F)
g
close(to.read)
```

Visualização da imagem na segunda posição

```
to.read = gzfile("t10k images idx3 ubyte.gz", "rb")
readBin(to.read, integer(), size=4, n=4, endian="big")
par(mfrow=c(1,1))
for(i in 1:2)f
  m <  matrix(readBin(to.read,integer(), size=1, n=28 28,
               endian="big", signed = F), nrow=28, byrow=T)
  image(t(m)[,28:1],col = gray(255:0 / 255), axes = F)
g
close(to.read)
m
```

Visualização da imagem na décima posição

```
to.read = gzfile("t10k images idx3 ubyte.gz", "rb")
readBin(to.read, integer(), size=4, n=4, endian="big")
par(mfrow=c(1,1))
```

```
for(i in 1:10){
  m <- matrix(readBin(to.read,integer(), size=1, n=28*28,
              endian="big", signed = F), nrow=28, byrow=T)
  image(t(m)[,28:1],col = gray(255:0 / 255), axes = F)
}
close(to.read)
m
```

Leitura da base de dados e o arquivo de rótulos

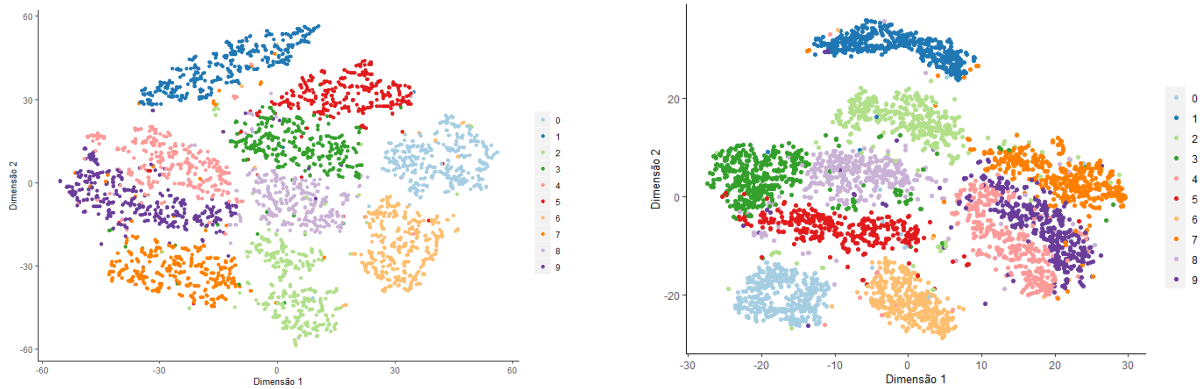
```
# Arquivo de imagens
to.read = gzfile("t10k-images-idx3-ubyte.gz", "rb")
readBin(to.read, integer(), size=4, n=4, endian="big")
m = matrix(0,ncol=28*28, nrow=10000)
for(i in 1:10000){
  m[i,] <- readBin(to.read,integer(), size=1, n=28*28,
                   endian="big", signed = F)
}
close(to.read)

# Arquivo de rotulos
to.read = gzfile("t10k-labels-idx1-ubyte.gz", "rb")
readBin(to.read, integer(), size=4, n=2, endian="big")
a = readBin(to.read,integer(), size=1, n=10000, endian="big")
close(to.read)
```

APÊNDICE 2 – Visualizações adicionais do *t-SNE*

Nesse apêndice são apresentadas outras visualizações do *t-SNE* com análise de componentes principais e escalonamento multidimensional para amostra com 4.000 observações. Os valores de perplexidade apresentados aqui são 10 e 70. A Figura 19 apresenta a visualização do *t-SNE* com análise de componentes principais e a Figura 20 com escalonamento multidimensional.

Na Figura 19 pode-se notar que a visualização com a perplexidade 10 ficou parecida com a perplexidade 25 apresentada na Seção 3.4, enquanto que com a perplexidade 70 os dígitos ficaram mais concentrados.

2.0.1 *t-SNE* com Análise de Componentes Principais

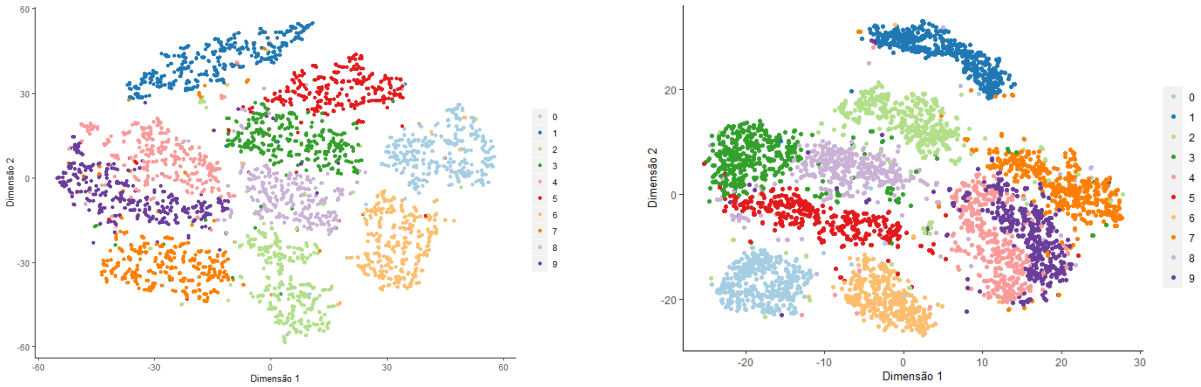


(a) Perp=10 (b) Perp=70

Figura 19: Gráfico do *t-SNE* com análise de componentes principais

A Figura 20 apresenta uma visualização muito semelhante a Figura 19, novamente, a visualização com a perplexidade 10 apresentou bons resultados, e com a perplexidade 70 os dígitos ficaram novamente concentrados.

2.0.2 *t-SNE* com Escalonamento Multidimensional



(a) Perp=10 (b) Perp=70

Figura 20: Gráfico do *t-SNE* com escalonamento multidimensional

ANEXO 1 – Demonstração Escalonamento Multidimensional Métrico

Nesse anexo será apresentado a demonstração do teorema 2.1. A demonstração foi retirada do livro *Multivariate Observations*, Seber (1984), nas páginas 236 e 237. Note que a notação do livro δ^2_{rs} equivale a d^2_{ij} desta monografia e $\mathbf{D} = [(\delta_{rs})]$.

Teorema 1.1. *Seja* $\mathbf{A} = [(a_{ij})]$, *sendo* $a_{ij} = -\frac{1}{2}\delta^2_{ij}$. *De ne-se* $b_{ij} = a^2_{ij} - a^2_{i:} - a^2_{:j} + a^2_{::}$ *tal que*

$$\mathbf{B} = [(b_{rs})] = (\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^{>})\,\mathbf{A}\,(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^{>}) : \tag{1.1}$$

Entao, $\mathbf{D}$ *e Euclidiana se, e somente se,* $\mathbf{B}$ *e positiva semide nida.*

Demonstracao. Dado que $\mathbf{D}$ é Euclidiana, existe uma composição $x_1; x_2; \ldots; x_n$ de modo que

$$2a_{rs} = -\delta^2_{rs} = \|\mathbf{x}_r - \mathbf{x}_s\|^2 = \mathbf{x}_r^{>}\mathbf{x}_r + \mathbf{x}_s^{>}\mathbf{x}_s - 2\mathbf{x}_r^{>}\mathbf{x}_s.$$

Além disso,

$$\begin{aligned}
2\bar{a}_{r:} &= \mathbf{x}_r^{>}\mathbf{x}_r + \frac{1}{n}\sum_i \mathbf{x}_i^{>}\mathbf{x}_i - 2\mathbf{x}_r^{>}\bar{\mathbf{x}};\\
2\bar{a}_{:s} &= \mathbf{x}_s^{>}\mathbf{x}_s + \frac{1}{n}\sum_i \mathbf{x}_i^{>}\mathbf{x}_i - 2\bar{\mathbf{x}}^{>}x_s;\\
2\bar{a}_{::} &= 2\frac{1}{n}\sum_i \mathbf{x}_i^{>}\mathbf{x}_i - 2\bar{\mathbf{x}}^{>}\bar{\mathbf{x}}:
\end{aligned}$$

Fazendo manipulações matemáticas, obtém-se

$$\begin{aligned}
b_{rs} &= \mathbf{x}_r^{>}\mathbf{x}_s - \mathbf{x}_r^{>}\bar{\mathbf{x}} - \bar{\mathbf{x}}^{>}\mathbf{x}_s + \bar{\mathbf{x}}^{>}\bar{\mathbf{x}}\\
&= (\mathbf{x}_r - \bar{\mathbf{x}})^{>}(\mathbf{x}_s - \bar{\mathbf{x}})
\end{aligned} \tag{1.2}$$

e,

$$\mathbf{B} = \bar{\mathbf{X}}\bar{\mathbf{X}}^{>} \quad 0: \tag{1.3}$$

sendo,

$$\bar{\mathbf{X}}^{>} = (\mathbf{x}_1 \quad \bar{\mathbf{x}}; \mathbf{x}_2 \quad \bar{\mathbf{x}}; :::; \mathbf{x}_n \quad \bar{\mathbf{x}}): \tag{1.4}$$

Suponha que B $\quad$ 0 possui o posto p. Então existe uma matriz ortogonal $\mathbf{V} = (\mathbf{v}_1; \mathbf{v}_2; :::; \mathbf{v}_n)$ tal que

$$\mathbf{V}^{>}\mathbf{B}\mathbf{V} = \begin{pmatrix} \Gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}; \tag{1.5}$$

sendo $\Gamma = diag(\ _1; \ _2; :::; \ _p)$ e $\ _1 \quad \ _2 \quad ::: \quad \ _p$ sendo eles os autovalores positivos associados a matriz $\mathbf{B}$. Seja $\mathbf{V}_1 = (\mathbf{v}_1; \mathbf{v}_2; :::; \mathbf{v}_p)$. A partir da equação (1.5), temos que

$$\begin{aligned} \mathbf{B} &= \mathbf{V} \begin{pmatrix} \Gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{V}^{>} \\ &= \mathbf{V}_1 \Gamma \mathbf{V}_1^{>} \\ &= (\mathbf{V}_1 \Gamma^{(1=2)})(\mathbf{V}_1 \Gamma^{(1=2)})^{>} \\ &= \mathbf{Y}\mathbf{Y}^{>}; \end{aligned}$$

sendo,

$$\begin{aligned} \mathbf{Y} &= (\sqrt{\ }_1\mathbf{v}_1; \sqrt{\ }_2\mathbf{v}_2; :::; \sqrt{\ }_p\mathbf{v}_p) \\ &= (\mathbf{y}^1; \mathbf{y}^2; :::; \mathbf{y}^p) \\ &= \begin{pmatrix} \mathbf{y}_1^{>} \\ \mathbf{y}_2^{>} \\ \vdots \\ \mathbf{y}_n^{>} \end{pmatrix}. \end{aligned} \tag{1.6}$$

Pode-se notar que $||\mathbf{y}^{(j)}||^2 = \ _j||\mathbf{v}_j||^2 = \ _j$. Desde que $b_{rs} = \mathbf{y}_r^{>}\mathbf{y}_s$,

$$\begin{aligned} ||\mathbf{y}_r \quad \mathbf{y}_s||^2 &= \mathbf{y}_r^{>}\mathbf{y}_r + \mathbf{y}_s^{>}\mathbf{y}_s \quad 2\mathbf{y}_r^{>}\mathbf{y}_s \\ &= b_{rr} + b_{ss} \quad b_{rs} \quad b_{sr} \\ &= a_{rr} + a_{ss} \quad 2a_{rs} \\ &= \quad 2a_{rs} \quad \text{(dado que } a_{rr} = a_{ss} = 0) \\ &= \ ^2_{rs}. \end{aligned} \tag{1.7}$$

Como $\mathbf{y_i}$ fornece a configuração necessária, consequentemente D é Euclidiana. □

O teorema 1.1 fornece o método para a construção de uma configuração $\{\mathbf{y}_i\}$ conhecido como o "clássico" método do escalonamento multidimensional.