Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Eduardo Kazuo Nakao

Abordagem contextual paramétrica na análise de componentes principais

São Carlos
2024

Eduardo Kazuo Nakao

Abordagem contextual paramétrica na análise de componentes principais

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Área de concentração: Metodologias e Técnicas de Computação

Orientador: Alexandre Luís Magalhães Levada

São Carlos
2024

Resumo

Em análise não-supervisionada, a qualidade de uma dispersão pode ser mensurada pela relação entre, a proximidade das amostras de uma classe, e a separação entre classes. Uma tarefa relevante nesse contexto é caracterizada pelo mapeamento de uma dispersão X em uma nova X' de melhor qualidade, que preserve as relações de vizinhança. O PCA é um método de mapeamento que possui como propriedade a manutenção do máximo espalhamento global das amostras. Tal propriedade pode gerar tanto uma maior separação entre classes quanto maiores espalhamentos intra-classe. O espalhamento intra-classe pode ser diminuído pela suavização da influência de amostras distantes de seus grupos. Tal suavização pode ser obtida pela substituição do cálculo de proximidade entre amostras individuais por uma medida de similaridade entre contextos. A partir dessa ideia, uma modificação no método é proposta, onde a informação contextual é extraída do grafo de vizinhança, e o conjunto de valores de uma característica é mapeado para uma distribuição estatística paramétrica, permitindo assim que uma divergência entre distribuições seja utilizada como medida de similaridade. Experimentos prévios a este trabalho envolvendo um conjunto limitado de divergências, indicaram que a abordagem é capaz de produzir resultados superiores a diversos métodos de mapeamento existentes. A primeira linha de investigação deste trabalho emprega um conjunto maior de divergências a fim de verificar a robustez da abordagem. Os resultados obtidos mostram que, para uma ampla gama de divergências, a proposta apresenta um desempenho superior a todos os outros algoritmos comparados, tanto na média quanto na maioria dos conjuntos de amostras testados. A segunda linha de investigação é definida pela verificação da existência de relação entre o desempenho de uma divergência e as propriedades de um conjunto de amostras. Do estudo realizado é possível inferir que divergências de maiores valores tendem a gerar melhores dispersões em conjuntos de maiores quantidades de classes.

Palavras-chave: Reconhecimento de padrões, Análise não-supervisionada, Aprendizagem de métricas, Teoria da informação.

Lista de ilustrações

Lista de tabelas

Lista de siglas

CSPCA Cauchy-Schwarz PCA

CH Calinski-Harabasz

CS Cauchy-Schwarz

DB Davies-Bouldin

ISOMAP Isometric Feature Mapping

JSPCA Joint Sparse PCA

KL Kullback-Leibler

KPCA Kernel PCA

KNN K-Nearest Neighbors

LAP Laplacian Eigenmaps

LDA Linear Discriminant Analysis

LLE Locally Linear Embbeding

L1PCA L1-norm PCA

MAD Median Absolute Deviation

MDS Multidimensional Scaling

MSE Mean Squared Error

PCA Principal Component Analysis

PDF Probability Density Function

RS R-Squared

RENPCA Renyi PCA

RPCA Robust PCA

SC Silhouette Coefficient

SFD Symmetric Fisher Divergence

SMPCA Sharma-Mittal PCA

SNE Stochastic Neighbor Embedding

TPCA Tsallis PCA

t-SNE t-distributed Stochastic Neighbor Embedding

TV Total Variation

UMAP Uniform Manifold Approximation and Projection

Sumário

Capítulo 1

Introdução

1.1 Contexto

Uma das principais tarefas em reconhecimento de padrões é caracterizada pelo particionamento de um conjunto X de n amostras descritas por m características em c classes. De maneira particular, quando X é representado por duas características ($m = 2$), o conjunto pode ser visualizado em um gráfico de dispersão bidimensional, onde cada amostra é um ponto x_i de coordenadas (x_{i1}, x_{i2}) onde x_{ij} é o valor da caraterística j na amostra i. Tal representação possibilita uma interpretação visual da dispersão. As Figuras 1 e 2 exibem dispersões de conjuntos hipotéticos particionados, onde cada tonalidade corresponde à uma classe.

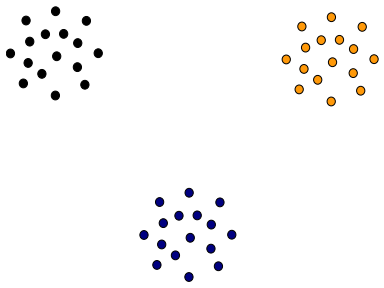


Figura 1 – Grupos individualmente coesos e pluralmente bem separados

A inspeção visual sugere que o conjunto da Figura 1 exibe uma dispersão de qualidade superior ao da Figura 2, no sentido de que o posicionamento das amostras melhor corresponde às partições, evidenciando assim o particionamento. Entretanto, conjuntos reais de amostras geralmente são descritos por mais de duas características, inviabilizando assim a representação bidimensional e a análise visual da dispersão. Dessa forma, medidas quantitativas são necessárias para mensurar a qualidade de uma dispersão tendo em vista

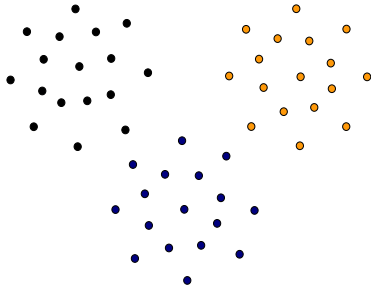


Figura 2 – Grupos de alto espalhamento individual e baixa separação plural

seu particionamento. Tais medidas utilizam as ideias de coesão de um grupo e separação entre grupos, no sentido de que, quanto mais próximas as amostras de um grupo estão umas das outras, mais coeso esse grupo é, enquanto que, quanto mais distantes os grupos se encontram entre si, melhor é a separação do conjunto. Tais critérios indicam portanto melhores particionamentos em dispersões cujos grupos são individualmente coesos e pluralmente bem separados (LIU et al., 2010).

Uma tarefa relacionada em reconhecimento de padrões é a de se transformar uma dispersão X em uma dispersão X' mapeando cada amostra para novas coordenadas na tentativa de melhorar sua qualidade sob a óptica dos critérios de coesão e separação (LI; TIAN, 2018). A Figura 3 (PEDREGOSA et al., 2011) exemplifica uma tarefa de mapeamento de forma visual apresentando uma dispersão de geometria originalmente esférica e os resultados da aplicação de alguns métodos de mapeamento, onde cada tonalidade corresponde à uma classe do particionamento.

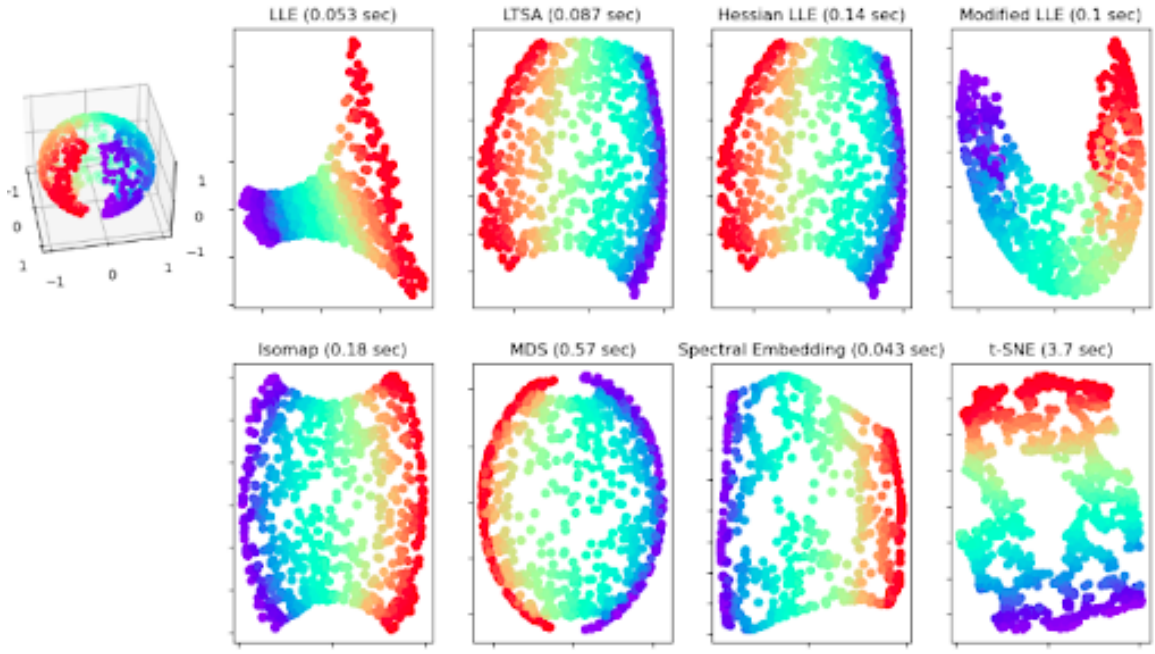


Figura 3 – Diferentes estratégias de mapeamento de um conjunto de amostras

Um fato notável nesse contexto é a heterogeneidade geométrica intrínseca dos diferentes conjuntos reais de amostras, no sentido de que, cada conjunto em si, determina uma dispersão de geometria singular. Esse fato exige a proposta de diferentes abordagens para realizar o mapeamento da maneira mais adequada para cada conjunto ou categoria geométrica de conjuntos. Trabalhos recentes fornecem uma compilação de algoritmos de mapeamento e seus contextos de adequação (LI; TIAN, 2018; SUáREZ; GARCíA; HERRERA, 2021).

Tais algoritmos tem como propósito portanto gerar representações mais adequadas para um determinado conjunto de amostras. Em problemas não-supervisionados em particular, onde nenhuma informação de classe é conhecida a priori, a existência de múltiplas estratégias de transformação é fundamental para que a melhor representação possível de uma determinada dispersão seja obtida, tornando assim o consequente reconhecimento dos padrões subjacentes mais preciso.

Um dos algoritmos de mapeamento mais populares existentes é o método da Análise de Componentes Principais (PCA) (JOLLIFFE, 2002), que possui como propriedade a manutenção do máximo espalhamento global da dispersão de entrada. Uma possível consequência de um maior espalhamento global das amostras, é uma melhor separação entre classes. Por outro lado, tal fenômeno também pode implicar em maiores espalhamentos intra-classe, o que degrada a qualidade da dispersão.

Um fator de influência natural no espalhamento intra-classe é a presença de amostras distantes de seus grupos. Dessa observação, surge a motivação de se tentar suavizar a influência de tais amostras. Uma possível ideia para isso, seria analisar a similaridade entre vizinhanças de amostras em substituição à análise usual entre amostras individuais.

A partir dessa ideia, Levada (2021) propõe uma modificação no método PCA a fim de verificar se, a incorporação de informação contextual poderia resultar em dispersões de maior qualidade. A informação contextual é obtida através do grafo de vizinhança de uma amostra e o conjunto de valores de uma característica é mapeado para uma distribuição estatística paramétrica (i.e. descrita por uma coleção de parâmetros). Tal mapeamento caracteriza uma passagem do espaço de características usual para um espaço paramétrico onde o cálculo de similaridade entre objetos é realizado por uma divergência entre distribuições estatísticas.

Experimentos iniciais (LEVADA, 2020; LEVADA, 2021) adotando as divergências de Kullback-Leibler (KL) (KULLBACK; LEIBLER, 1951), Bhattacharyya (BHATTACHARYYA, 1943) e Hellinger (PARDO, 2006) demonstraram a competitividade da abordagem, não só em comparação ao PCA original, mas também à outras modificações do PCA já propostas na literatura, assim como em comparação a diversos algoritmos baseados em grafo de vizinhança.

A divergência KL foi a primeira escolha considerada pelo fato de ser definida pela entropia relativa entre duas distribuições estatísticas. Entretanto, devido ao fato da entropia relativa ser não-simétrica (ABOU-MOUSTAFA; FERRIE, 2012), a distância de Bhattacharyya se apresenta como uma primeira alternativa a ser testada pela capacidade de definir limites superiores na probabilidade do erro de classificação Bayesiana (DUDA; HART; STORK, 2000). Adicionalmente, a distância de Hellinger se apresenta como segunda opção por ser função do coeficiente de Bhattacharyya.

1.2 Motivação e objetivos

A formulação teórica da abordagem contextual paramétrica permite a adoção de qualquer divergência estatística para o cálculo de similaridade entre dois objetos no espaço paramétrico. Desse fato, surge o interesse de se investigar os efeitos da escolha da divergência nos resultados. A primeira linha de investigação deste trabalho de doutorado tem como objetivo verificar a sensibilidade da abordagem à divergência, isto é, se independentemente da escolha da divergência, a abordagem de modo geral é capaz de gerar resultados competitivos a métodos existentes.

Uma segunda linha de investigação subsequente é a de se verificar a existência de alguma relação entre o desempenho de uma determinada divergência e as propriedades de um conjunto de amostras, no sentido de verificar se uma divergência de determinado comportamento tende a gerar uma dispersão de melhor qualidade em um conjunto de determinada propriedade.

Para conduzir tais investigações, este trabalho de doutorado tem como primeira hipótese de pesquisa a suposição de que o desempenho global da abordagem contextual paramétrica apresentado em Levada (2020) e Levada (2021), não é afetado de maneira significativa quando outras divergências são empregadas. Já a segunda hipótese é a de que existe alguma relação de adequação entre o comportamento de uma divergência e alguma propriedade de um conjunto de amostras. Para investigar as hipóteses, este trabalho é pautado pelos seguintes objetivos:

1. Verificar se a abordagem contextual paramétrica no método PCA é flexível à escolha da divergência, no sentido de não apresentar resultados significativamente inferiores quando uma divergência específica é adotada.

2. Estudar os efeitos da escolha da divergência através da identificação de possíveis propriedades de um conjunto de amostras que indicariam maior adequação de uma divergência em detrimento de outra.

1.3 Síntese da metodologia e resultados

Objetivo 1

O protocolo experimental adotado em Levada (2020) e Levada (2021) é caracterizado pela aplicação de múltiplos algoritmos de mapeamento em um dado conjunto de amostras, onde cada algoritmo gera uma nova representação para o conjunto. Subsequentemente, para cada uma dessas representações, a qualidade da dispersão gerada é avaliada, tornando possível a análise do desempenho da abordagem proposta em relação aos algoritmos comparados. Diversos conjuntos reais de amostras são utilizados na comparação.

O mesmo protocolo experimental foi utilizado para investigar o Objetivo 1 deste trabalho, onde executamos o PCA contextual paramétrico sob outras divergências ainda não testadas. Dos resultados é possível concluir que, tanto na média quanto na maioria dos conjuntos, a proposta apresentou um desempenho superior a todos os outros métodos comparados sob qualquer divergência. Dessa forma, os resultados indicam que a abordagem não é sensível à divergência. Tais análises se encontram publicadas nos artigos de **Nakao e Levada (2023)** e **Nakao e Levada (2024)**.

Objetivo 2

O Objetivo 2 é composto tanto pela identificação de possíveis tendências de comportamento das divergências sob a óptica do impacto na qualidade da dispersão gerada, quanto pela condução de experimentos para confirmar tais tendências. A metodologia de investigação do Objetivo 2 é constituída pelos seguintes passos:

1. Estudo do comportamento das divergências através de experimentos envolvendo distribuições de probabilidade hipotéticas e cenários envolvendo conjuntos sintéticos de amostras. Como resultado desse estudo, duas sub-hipóteses de pesquisa são geradas:

   - ❏ A hipótese de que uma divergência limitada é mais adequada para problemas de classificação binária.
   - ❏ A hipótese de que uma divergência não-limitada de maiores valores resultantes é mais adequada para problemas de mais de duas classes.

2. Experimentos com o objetivo de verificar se as sub-hipóteses teóricas se confirmam. Para a condução de tal verificação, os seguintes passos são adotados:

   a) aplicar em conjuntos reais de amostras a abordagem contextual paramétrica utilizando divergências de comportamentos significativamente distintos.

b) avaliar com diferentes medidas a qualidade da dispersão gerada pelo método de transformação sob cada divergência.

c) identificar os conjuntos que apresentam diferença significativa de desempenho entre divergências.

d) analisar se há correspondência entre as propriedades dos conjuntos de amostras identificados com os comportamentos teóricos de cada divergência.

Após a condução dos experimentos, os resultados em sete conjuntos de amostras se mostraram significativamente distintos do ponto de vista da qualidade da dispersão. Desses conjuntos, a divergência limitada se mostrou mais adequada em três, sendo que os três possuíam apenas duas classes. Em contrapartida, todos os outros quatro conjuntos onde a divergência não-limitada apresentou melhores resultados possuíam mais de duas classes. Tais resultados confirmam portanto as sub-hipóteses teóricas.

1.4 Organização do texto

Todos os temas e conceitos mencionados nas seções anteriores são explicados, expandidos, ilustrados e detalhados na fundamentação teórica composta pelos Capítulos 2 a 6 desta dissertação. Seguindo a mesma ordem de ideias apresentadas até o momento, primeiramente formalizamos no Capítulo 2 o tema da qualidade de uma dispersão fornecendo as formulações das medidas utilizadas neste trabalho.

A seguir, apresentamos no Capítulo 3 a derivação matemática completa do algoritmo de mapeamento modificado neste trabalho (PCA), para no Capítulo 4 discutirmos a ideia da representação paramétrica para a informação contextual, explicando sua formulação e aplicação no PCA ao se adotar o modelo Gaussiano univariado. Ao fim do capítulo é possível entender o papel das divergências estatísticas no processo.

No Capítulo 5 é discutida a ideia de divergência estatística, ferramenta esta que utiliza conceitos pertencentes ao campo da teoria da informação. Inicialmente uma breve introdução à área é feita, onde os conceitos fundamentais de informação e entropia são apresentados e desenvolvidos até os detalhes matemáticos das divergências estocásticas. No Capítulo 6 fornecemos uma intuição da ideia de divergência geodésica, uma classe de divergências que utiliza a informação de Fisher em substituição à entropia, sendo assim uma alternativa à abordagem estocástica.

Tendo sido construída a fundamentação teórica necessária, nos Capítulos 7 e 8 detalhamos a metodologia de investigação dos Objetivos 1 e 2 e apresentamos os resultados experimentais obtidos e suas conclusões. Encerramos o texto apresentando no Capítulo 9 trabalhos futuros e relacionados, onde apontamos direções de melhoria na abordagem proposta e apresentamos as ideias e referências de cinco algoritmos de mapeamento que incorporam informação contextual através do grafo vizinhança.

Capítulo 2

Medidas da qualidade de uma dispersão

Seguindo a ordem de ideias introduzidas no Capítulo 1, o primeiro assunto a ser tratado diz respeito à avaliação da qualidade de uma dispersão. Até o momento, fornecemos somente uma intuição das ideias de coesão e espalhamento como critérios de avaliação e apontamos a necessidade de se formalizar tais critérios em medidas que quantifiquem a qualidade. Este capítulo tem como função portanto, apresentar algumas formas de se mensurar a coesão de um grupo e a separação entre grupos, apresentando as formulações de algumas medidas existentes. Uma compilação de outras medidas pode ser encontrada nos trabalhos de Liu et al. (2010) e Deborah, Baskaran e Kannan (2010).

R-squared

A medida R-squared (RS) (SHARMA, 1995) é definida por uma razão da forma

$$(separação - coesão)/separação \tag{1}$$

em que a *separação* é calculada pela soma das distâncias quadráticas entre uma amostra e o centro do conjunto de amostras, e a *coesão* é calculada pela soma das distâncias quadráticas entre uma amostra pertencente a um grupo e o centro do mesmo.

Portanto, quanto maior o valor retornado por essa medida, melhor a qualidade da dispersão. Note que a razão é do tipo $(A - B)/A$, com $A \geq B$, retornando assim valores somente no intervalo $[0, 1]$. Considerando:

- x: uma amostra
- n: a quantidade de amostras do conjunto
- $centro$: a amostra central do conjunto
- G: um grupo
- c: a quantidade de classes (grupos)
- $centro(j)$: a amostra central do grupo j

$$RS = \frac{\sum_{i=1}^{n} d^2(x_i, centro) - \sum_{j=1}^{c} \sum_{x \in G_j} d^2(x, centro(j))}{\sum_{i=1}^{n} d^2(x_i, centro)} \tag{2}$$

Calinski-Harabasz

A medida de Calinski-Harabasz (CH) (CALIńSKI; JA, 1974) é definida por uma razão ponderada, da forma *separação/compacidade*. A *separação* é baseada na distância quadrática entre os centros dos grupos e o centro do conjunto de amostras, e portanto, quanto maior seu valor, mais bem separados os grupos estarão e maior será o valor final da medida.

O valor de *compacidade* é baseado na distância quadrática entre a amostra de um grupo e o centro do mesmo, e portanto, quanto menor esse valor, mais compacto é o grupo e maior será o valor final da medida. Assim, a medida CH também indica melhores dispersões em maiores valores. Sua formulação é dada por:

$$CH = \frac{\frac{\sum_{j=1}^{c} |G_j| d^2(centro(j), centro)}{(c-1)}}{\frac{\sum_{j=1}^{c} \sum_{x \in G_j} d^2(x, centro(j))}{(n-c)}} \tag{3}$$

Davies-Bouldin

A medida de Davies-Bouldin (DB) (DAVIES; BOULDIN, 1979) mensura a similaridade média entre os grupos, onde a similaridade é uma medida que compara a distância entre os grupos com seus tamanhos. O menor valor que a medida assume é zero. Valores próximos de zero indicam partições com maior separação entre grupos. Essa medida é definida como sendo a similaridade média entre cada grupo G_i (para $i = 1, ..., c$) e o seu grupo mais similar G_j. No contexto desse índice, a similaridade é definida como uma medida R_{ij} que leva em consideração:

- ❑ a distância média entre cada amostra de um grupo e o centro do mesmo (também entendida como o diâmetro do grupo)
- ❑ a distância entre os centros dos grupos i e j.

Uma maneira de se construir R_{ij} de forma não-negativa e simétrica é:

$$R_{ij} = \frac{\left[\frac{1}{|G_i|-1}\sum_{x\in G_i,} d(x, centro(i))\right] + \left[\frac{1}{|G_i|-1}\sum_{x\in G_i, j\neq i} d(x, centro(j))\right]}{d(centro(i), centro(j))} \tag{4}$$

Com isso, a medida de Davies-Bouldin é definida como:

$$DB = \frac{1}{c}\sum_{i=1}^{c} \max_{i\neq j} R_{ij} \tag{5}$$

Silhouette

Seja G_k o k-ésimo grupo, então para cada amostra $x_i \in G_k$ seja $d_m(x_i)$ a distância média entre x_i e todas as outras amostras do mesmo grupo G_k:

$$d_m(x_i) = \frac{1}{|G_k| - 1} \sum_{x_j \in G_k, j\neq i} d(x_i, x_j) \tag{6}$$

em que $d(x_i, x_j)$ é a distância entre as amostras x_i e x_j no grupo G_k. Em outras palavras, podemos interpretar $d_m(x_i)$ como uma medida de pertinência da amostra x_i ao seu grupo (quanto menor esse valor, melhor).

Então, defini-se a dissimilaridade média entre uma amostra x_i a um grupo G como a média das distâncias entre x_i a todas as amostras em G. Para cada amostra x_i, seja $md_m(x_i)$ a menor distância média entre x_i e todas as amostras em qualquer outro grupo onde x_i não é membro:

$$md_m(x_i) = \min_{k\neq i} \frac{1}{|G_k|} \sum_{x_j \in G_k} d(x_i, x_j) \tag{7}$$

O grupo com a menor dissimilaridade média é o grupo vizinho de x_i pois é o segundo grupo mais adequado à x_i. Defini-se por $s(x_i)$ o valor de silhueta da amostra x_i:

$$s(x_i) = \frac{md_m(x_i) - d_m(x_i)}{max\{d_m(x_i), md_m(x_i)\}}, \text{ se } |G_k| > 1 \tag{8}$$

$$s(x_i) = 0, \text{ se } |G_k| = 1 \tag{9}$$

Combinando todas as definições temos:

$$s(x_i) = \begin{cases} 1 - \dfrac{d_m(x_i)}{md_m(x_i)} & \text{se } d_m(x_i) < md_m(x_i) \\ 0 & \text{se } d_m(x_i) = md_m(x_i) \\ \dfrac{md_m(x_i)}{d_m(x_i)} - 1 & \text{se } d_m(i) > md_m(x_i) \end{cases} \tag{10}$$

Note que $-1 \leq s(x_i) \leq 1$. Um $s(x_i)$ próximo de 1 significa que os dados estão devidamente agrupados. Se $s(x_i)$ tende a -1, isso indica que x_i é mais similar a seu grupo vizinho. Um $s(x_i)$ próximo de 0 significa que a amostra x_i está na borda entre os dois grupos. A média $s(x_i)$ de todas as amostras de um grupo é uma medida de compacidade. Portanto, a média $s(x_i)$ de todas as amostras do conjunto de dados, conhecida como coeficiente Silhouette (SC) (ROUSSEEUW, 1987), é uma medida da qualidade da dispersão.

Considerações finais

As medidas apresentadas são utilizadas na investigação do Objetivo 1 para avaliar a qualidade das dispersões geradas pelos diferentes algoritmos de mapeamento. Já na investigação do Objetivo 2, tais medidas são utilizadas para avaliar a qualidade das dispersões geradas por diferentes divergências. A seguir, apresentamos no Capítulo 3 a formulação e derivação completa do algoritmo PCA que preserva a máxima variância do conjunto de entrada. Tal apresentação possibilita o entendimento da adaptação do método para viabilizar a abordagem contextual.

Capítulo 3

Análise de componentes principais

O PCA é uma família de técnicas que utiliza as dependências entre as amostras do conjunto para gerar uma nova representação que maximize a variabilidade intrínseca do mesmo, conforme ilustrado na Figura 4 (GUPTA; SEHGAL; ACKEN, 2025). Se trata de um dos algoritmos de mapeamento pioneiros na área, tendo sido redescoberto em diferentes ocasiões em áreas distintas. O PCA implementa a Transformação de Karhunen-Loève (ou Transformação de Hotteling) como é conhecida na literatura de reconhecimento de padrões (Jolliffe (2002) exibe na seção 1.2 um breve histórico com todas as referências relevantes).

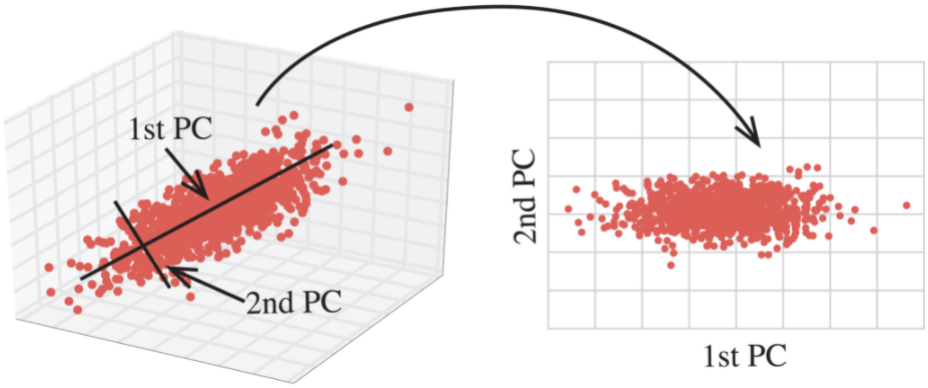


Figura 4 – Visualização da aplicação do PCA em uma dispersão bidimensional

O PCA é um método clássico de análise estatística de dados multivariados que realiza a expansão de um vetor x em termos dos d autovetores da matriz de covariância associados aos d maiores autovalores. Por essa razão, é limitado à estatística de segunda ordem onde nenhuma hipótese sobre as densidades de probabilidade é necessária, uma vez que

toda a informação pode ser estimada diretamente das amostras. Dado um conjunto de dados multivariados, o objetivo é reduzir a redundância existente para encontrar a melhor representação que minimize o critério do erro quadrático médio (*Mean Squared Error* - MSE). A redundância é medida pela correlação entre os dados.

O requisito básico no PCA é a existência de um vetor aleatório X com n elementos. Considera-se X como um vetor coluna. Devem estar disponíveis amostras $x_1, ..., x_n$ desse vetor. Nenhum modelo generativo é assumido para o vetor X, mas é necessário que os seus elementos sejam mutuamente correlacionados. Na transformação PCA, primeiramente os dados são centralizados subtraindo-lhes a média amostral. Em seguida, x é transformado linearmente em um vetor y contendo d elementos, com $d \leq m$, de maneira que a redundância introduzida pela correlação é eliminada.

Geometricamente, tal condição é obtida através de uma rotação do sistema de coordenadas ortogonal, de modo que os componentes de x no novo sistema sejam não-correlacionados. Simultaneamente, as variâncias das projeções de x nos novos eixos são maximizadas, sendo que o primeiro eixo corresponde à maior variância, o segundo à maior na direção ortogonal ao primeiro e assim por diante. Pode ser mostrado (FUKUNAGA, 1990) que, se λ_j e u_j são, respectivamente, o j-ésimo autovalor e autovetor da matriz de covariância de X , então para $j \neq k$

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

ou seja, todos os autovalores são positivos e os autovetores são mutuamente ortogonais entre si. Como consequência, para uma matriz de posto m, tem-se m autovetores ortonormais (assumindo que $\parallel u_j \parallel = 1$ para $j = 1, ..., m$) associados aos autovalores $\lambda_1, ..., \lambda_m$. Matematicamente, pode-se expressar a rotação do sistema de coordenadas, definida pela Transformação Karhunen-Loève, como uma matriz ortonormal $Z = [T^t, S^t]$ de dimensões $D \times m$, com $T^t = [w_1, ..., w_d]_{m \times d}$ representando os eixos do novo sistema de coordenadas e $S^t = [w_{d+1}, ..., w_m]_{m \times (m-d)}$ representando os eixos referentes às componentes eliminadas. A condição de ortonormalidade implica que $w_j \cdot w_k = 0$ para $j \neq k$, e $w_j \cdot w_k = 1$ para $j = k$. Pode-se escrever o vetor m-dimensional x através de sua expansão nos vetores da base:

$$x = \sum_{j=1}^{m} (x^t w_j) w_j = \sum_{j=1}^{m} c_j w_j \tag{11}$$

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

$$y^t = x^t T^t = \sum_{j=1}^{m} c_j w_j^t \, [w_1, ..., w_d] = [c_1, ..., c_d] \tag{12}$$

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

$$J_1^{PCA}(T) = E\left[\parallel y \parallel^2\right] = E\left[y^t y\right] = \sum_{j=1}^{d} E\left[c_j^2\right] \tag{13}$$

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

$$J_1^{PCA}(w_j) = \sum_{j=1}^{d} E\left[w_j^t x x^t w_j\right] = \sum_{j=1}^{d} w_j^t E\left[x x^t\right] w_j = \sum_{j=1}^{d} w_j^t \Sigma_X w_j \tag{14}$$

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

$$J_1^{PCA}(w_j, \gamma_j) = \sum_{j=1}^{d} w_j^t \Sigma_X w_j - \sum_{j=1}^{d} \gamma_j (w_j^t w_j - 1) \tag{15}$$

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

$$\Sigma_X w_j = \lambda_j w_j \tag{16}$$

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

$$\Sigma_Y = T^t \Sigma_X T = T^t T diag(\lambda_1, ..., \lambda_n) T^t T = diag(\lambda_1, ..., \lambda_n) \tag{17}$$

Uma outra abordagem para o PCA é a da minimização do MSE. Nessa abordagem, busca-se um conjunto de d vetores ortonormais de base que gerem um subespaço d-dimensional tal que, o MSE entre o vetor original x e sua projeção nesse subespaço, seja mínima. Denotando os vetores da base por $w_1, ..., w_m$, pela condição de ortonormalidade tem-se $w_i^t w_j = \delta_{ij}$ em que

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

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

A projeção de x no subespaço gerado pelos vetores w_j, com $j = 1, ..., d$, é dada pela Equação 11 e portanto o critério MSE a ser minimizado torna-se:

$$J_{MSE}^{PCA}(w_j) = E\left[\| x - \sum_{j=1}^{d} \left(x^t w_j\right) w_j \|^2\right] \tag{18}$$

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

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

$$E\left[\| x \|^2\right] - \sum_{j=1}^{d} E\left[w_j^t x x^t w_j\right] = E\left[\| x \|^2\right] - \sum_{j=1}^{d} w_j^t \Sigma_X w_j \tag{19}$$

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

$$\sum_{j=1}^{d} w_j^t \Sigma_X w_j \tag{20}$$

Porém, da Equação 14, esse mesmo problema de otimização foi resolvido através de multiplicadores de Lagrange, e o resultado obtido é que os vetores w_j devem ser os autovetores de Σ_X. Então, substituindo-se a Equação 16 em 19, tem-se:

$$J_{MSE}^{PCA}(w_j) = E\left[\| x \|^2\right] - \sum_{j=1}^{d} \gamma_j \tag{21}$$

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

$$J_{MSE}^{PCA}(w_j) = \sum_{j=d+1}^{m} \gamma_j \tag{22}$$

A seguir sumarizamos o algoritmo PCA:

Algoritmo 1 PCA

0: **function** PCA(X)
0: Calcule a media e a matriz de covariância das amostras:
$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$
$\Sigma_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)^t$
0: Calcule os autovalores e os autovetores de Σ_x
0: Defina a matriz de transformação $T = [w_1, w_2, ..., w_d]$ com os d autovetores associados aos d maiores autovalores.
0: Projete os dados X no subespaço PCA:
$y_i = Tx_i \quad \text{for} \quad i = 1, 2, ..., n$
0: **return** Y
0: **end function**=0

Capítulo 4

Abordagem contextual paramétrica

Inspirado na ideia de métodos já existentes de incorporar a informação contextual fornecida pelo grafo KNN construído do conjunto de entrada (vide Capítulo 9), Levada (2020) propõe uma abordagem contextual ao PCA que mapeia o conjunto de valores de uma característica da vizinhança de uma amostra para um espaço de distribuições estatísticas paramétricas. A seguir apresentamos a derivação matemática completa da abordagem.

Um conjunto de amostras é definido como sendo o conjunto $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$, em que $\vec{x}_i \in R^m$. Se, para todo i, $\vec{x}_i$ é conectado aos seus K vizinhos mais próximos, o grafo KNN é definido por $G = (V, E)$, em que $|V| = n$. A distância Euclidiana pode ser usada para essa conexão, assumindo que uma vizinhança é um subespaço Euclidiano em si (ROWEIS; SAUL, 2000). Entretanto, outras medidas como Jaccard, Minkowski e Coseno também poderiam ser utilizadas. Um patch P_i é definido por $\{\vec{x}_i\} \cup \{\vec{x}_j \in N(i)\}$, com $N(i)$ sendo a vizinhança de $\vec{x}_i$. Então

$$P_i = [\vec{x}_i, \vec{x}_{i1}, \vec{x}_{i2}, ..., \vec{x}_{ik}] \tag{23}$$

é a matriz $m \times (K + 1)$ que representa o i-ésimo patch.

Assume-se que cada linha da matriz P_i é uma amostragem de tamanho $K + 1$ de uma variável aleatória x, caracterizada por uma função densidade de probabilidade $p(x; \vec{\theta})$, onde $\vec{\theta} \in R^L$ é um vetor de L parâmetros. Neste trabalho considera-se o modelo Gaussiano, ou seja, $L = 2$ e $\theta_1 = \mu$ denota a média e $\theta_2 = \sigma^2$ denota a variância.

Portanto, cada variável aleatória corresponde à uma de m características de entrada. Cada P_i é mapeado para um vetor m-dimensional de tuplas 2D, onde cada tupla j para $j = 1, ..., m$ possui os estimadores de máxima-verosimilhança para os parâmetros de cada característica. Em outras palavras, calcula-se a média e variância da amostragem dada

por cada linha da matriz P_i. O vetor de características paramétrico $\vec{p}_i$ para o patch P_i é dado por:

$$\vec{p}_i = \left[\vec{\theta}_1^{(i)}, \vec{\theta}_2^{(i)}, ..., \vec{\theta}_m^{(i)}\right] \tag{24}$$

onde cada componente é uma tupla de dois parâmetros:

$$\vec{\theta}_j^{(i)} = \left(\mu_j^{(i)}, (\sigma_j^2)^{(i)}\right) \tag{25}$$

A Figura 5 ilustra esse mapeamento de um patch P_i para um vetor de características paramétrico $\vec{p}_i$.

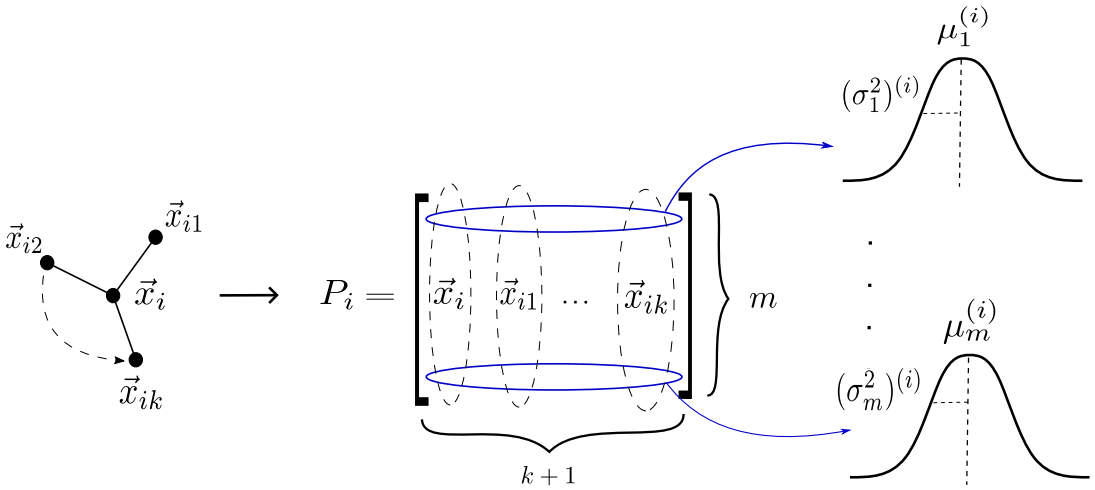


Figura 5 – Mapeamento de um patch P_i para um vetor paramétrico $\vec{p}_i$

O conjunto de todos os $\vec{p}_i$, para $i = 1, 2, ..., n$ define o espaço de características paramétrico. Pode-se associar ao espaço de características paramétrico, um centroide, que representa a distribuição média:

$$\tilde{\vec{p}} = \frac{1}{n}\sum_{i=1}^{n}\vec{p}_i \tag{26}$$

Seja a diferença paramétrica entre dois vetores $\vec{p}_i$ e $\vec{p}_j$ no espaço de características paramétrico a divergência estatística entre cada uma das tuplas nos vetores:

$$\begin{aligned}\vec{p}_i - \vec{p}_j &= \left[D(\vec{\theta}_1^{(i)}, \vec{\theta}_1^{(j)}), ..., D(\vec{\theta}_m^{(i)}, \vec{\theta}_m^{(j)})\right] \\ &= \vec{d}\,(\vec{p}_i, \vec{p}_j)\end{aligned} \tag{27}$$

onde $D(p, q)$ é a divergência estatística entre as funções densidade de probabilidade p e q.

Definimos a matriz de covariância paramétrica C como uma substituta para a matriz de covariância usual, com $\vec{d}(\vec{p}_i, \tilde{\vec{p}})$ sendo o vetor m-dimensional de divergências estatísticas:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}\vec{d}(\vec{p}_i, \tilde{\vec{p}})\vec{d}(\vec{p}_i, \tilde{\vec{p}})^t \tag{28}$$

A nova matriz pode então ser usada com transparência no algoritmo PCA original. Portanto, a matriz de projeção final do PCA (responsável pela projeção linear das coordenadas antigas para as novas coordenadas) pode ser construída normalmente com os

autovetores da nova matriz de covariância paramétrica cuja análise espectral resulta nas componentes principais dos dados observados. Note que a matriz C é real, simétrica e positiva semi-definida, o que implica que todos seus autovalores são não negativos.

A seguir apresentamos um sumário do método proposto. A entrada para o PCA contextual paramétrico é uma matriz de dados $X_{m\times n}$, em que cada coluna $\vec{x}_j \in R^m$ é uma amostra. Note que a abordagem paramétrica depende da definição de uma vizinhança pois, se $K = n$, a distribuição dos valores das coordenadas de um vetor representativo de uma característica j qualquer seria a mesma para todas as n amostras, fazendo com que o vetor diferença entre duas amostras fosse sempre nulo. Sob hipótese Gaussiana, o algoritmo PCA contextual paramétrico pode ser resumido como:

1. A partir dos dados de entrada $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n \in R^m$ construir um grafo KNN não direcionado;

2. Para cada patch $\{\vec{x}_i\} \cup \{\vec{x}_j \in N(i)\}$, onde $N(i)$ denota a vizinhança local de $\vec{x}_i$, compute os estimadores de máxima verossimilhança dos parâmetros do modelo para cada uma das características. Ao fim deste passo, gera-se para cada patch, o seguinte vetor paramétrico:

$$\vec{p}_i = \left[(\theta_{11}^{(i)}, ..., \theta_{1L}^{(i)}), (\theta_{21}^{(i)}, ..., \theta_{2L}^{(i)}), ...(\theta_{m1}^{(i)}, ..., \theta_{mL}^{(i)})\right] \tag{29}$$

3. Computar o vetor paramétrico média para todos os patches, $\tilde{\vec{p}}$, que representa a distribuição média:

$$\tilde{\vec{p}} = \left[(\tilde{\theta}_{11}, ..., \tilde{\theta}_{1L}), (\tilde{\theta}_{21}, ..., \tilde{\theta}_{2L}), ..., (\tilde{\theta}_{m1}, ..., \tilde{\theta}_{mL})\right] \tag{30}$$

4. Computar a matriz de covariâncias paramétrica C, baseada na divergência estatística entre os vetores paramétricos $\vec{p}_i$ e a distribuição média $\tilde{\vec{p}}$ como:

$$C = E\left[\vec{d}(\vec{p}_i, \tilde{\vec{p}})\ \vec{d}(\vec{p}_i, \tilde{\vec{p}})^t\right] = \frac{1}{n-1}\sum_{i=1}^{n} \vec{d}(\vec{p}_i, \tilde{\vec{p}})\ \vec{d}(\vec{p}_i, \tilde{\vec{p}})^t \tag{31}$$

   onde $\vec{d}(\vec{p}_i, \tilde{\vec{p}})$ é um vetor de divergências estatísticas.

5. Selecionar os $d < m$ autovetores associados aos d maiores autovalores da matriz C para compor a matriz de projeção W.

6. Projetar os dados no subespaço gerado por W.

A abordagem paramétrica envolve portanto uma divergência estatística em seu processo. No próximo capítulo serão apresentadas a origem e formulação das divergências adotadas neste trabalho, muitas dais quais utilizam conceitos pertencentes ao campo da teoria da informação.

Capítulo 5

Teoria da informação

A teoria da informação é um ramo da matemática que se preocupa com a quantificação da informação na transmissão, processamento, extração e compressão de dados. Reconhecida como um novo campo de pesquisa após o trabalho de Shannon (1948), a teoria da informação é particularmente relevante no estudo de processos aleatórios como forma de caracterizar e quantificar a noção de incerteza de maneira precisa e formal.

O conceito de informação é utilizado para quantificar a surpresa de um evento, no sentido de que, a realização de um evento de alta probabilidade não traz uma quantidade significativa de surpresa (não é muito informativa). Formalmente, a informação de um evento X é dada por:

$$I(X) = -log\ p(X) \tag{32}$$

A entropia de Shannon é uma medida estatística usada para quantificar a informação de uma variável aleatória, isto é, quantificar o grau de incerteza em processos aleatórios. A entropia de uma variável aleatória discreta x de n eventos X_i representa a informação obtida quando o valor da variável se torna conhecido:

$$H(x) = -\sum_{i=1}^{n} P(X_i) log\ p(X_i) \tag{33}$$

Dado que $\sum_{i=1}^{n} p(X_i) = 1$ então uma variável que possua um evento de probabilidade muito alta apresentará uma baixa entropia. A entropia de uma variável aleatória contínua x é o valor esperado da informação própria dado pela média do negativo do logaritmo da distribuição $p(x)$:

$$H(p) = -\int p(x)[log\ p(x)]dx = -E\left[log\ p(x)\right] \tag{34}$$

onde $p(x)$ é a função densidade de probabilidade (PDF) de x. Assumindo que x é normalmente distribuída por $N(\mu, \sigma^2)$, então $p(x)$ é dada por:

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{35}$$

onde μ denota a média e σ^2 denota a variância de x. Calculando o logaritmo de $p(x)$:

$$log\ p(x) = -\frac{1}{2} log\ (2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 \tag{36}$$

Substituindo a Equação (36) na Equação (34) tem-se:

$$H(p) = \frac{1}{2} log\ (2\pi\sigma^2) + \frac{1}{2\sigma^2} E[(x-\mu)^2] = \frac{1}{2}\left(1 + log\ (2\pi\sigma^2)\right) \tag{37}$$

A entropia cruzada quantifica a entropia entre duas distribuições de probabilidade no mesmo conjunto de eventos:

$$H(p, q) = -\sum_x P(x) log Q(x) \tag{38}$$

A entropia cruzada entre duas PDF's é:

$$H(p, q) = -\int p(x)[log\ q(x)]dx \tag{39}$$

A entropia relativa ou divergência de Kullback-Leibler (KL) nada mais é que a diferença entre a entropia cruzada de $p(x)$ e $q(x)$ e a entropia de $p(x)$:

$$\begin{aligned} D_{KL}(p, q) = H(p, q) - H(p) &= -\int p(x)[log\ q(x)]dx + \int p(x)[log\ p(x)]dx \\ &= \int p(x) log\left(\frac{p(x)}{q(x)}\right) dx = E_p\left[log\left(\frac{p(x)}{q(x)}\right)\right] \end{aligned} \tag{40}$$

Note que a entropia relativa é sempre não-negativa, ou seja, $D_{KL}(p, q) \geq 0$, atingindo o valor zero, se e somente se, $p(x) = q(x)$. Para verificar isso, primeiramente note que $log(a) \leq a - 1$ para $a > 0$. Então:

$$\begin{aligned} -D_{KL}(p, q) &= -\int p(x) log\left(\frac{p(x)}{q(x)}\right) dx \\ &= \int p(x) log\left(\frac{q(x)}{p(x)}\right) dx \leq \int p(x)\left(\frac{q(x)}{p(x)} - 1\right) dx \\ &= \int p(x)dx - \int q(x)dx = 1 - 1 = 0 \end{aligned} \tag{41}$$

Assumindo que $p(x)$ e $q(x)$ são densidades Gaussianas univariadas, a divergência KL possui uma expressão fechada:

$$
\begin{aligned}
D_{KL}(p,q) &= E_p\left[-log\ \sigma_1 - \frac{1}{2\sigma_1^2}(x-\mu_1)^2 + log\ \sigma_2 - \frac{1}{2\sigma_2^2}(x-\mu_2)^2\right] \\
&= log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2\sigma_2^2}E_p[(x-\mu_2)^2] - \frac{1}{2\sigma_1^2}E_p[(x-\mu_1)^2] \qquad (42)
\end{aligned}
$$

O cálculo dos momentos estatísticos de segunda ordem é muito simples nesse modelo:

$$E_p[(x-\mu_1)^2] = \sigma_1^2 \qquad (43)$$

$$E_p[(x-\mu_2)^2] = E_p[x^2] - 2E_p[x]\mu_2 + \mu_2^2 \qquad (44)$$

$$E_p[x^2] = Var_p[x] + E_p^2[x] = \sigma_1^2 + \mu_1^2 \qquad (45)$$

o que finalmente leva a:

$$
\begin{aligned}
D_{KL}(p,q) &= log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2\sigma_2^2}(\sigma_1^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2) - \frac{1}{2} \\
&= log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \qquad (46)
\end{aligned}
$$

Note que a entropia relativa não é simétrica (ABOU-MOUSTAFA; FERRIE, 2012), ou seja, $D_{KL}(p,q) \neq D_{KL}(q,p)$. Para utilizá-la como medida de similaridade, é necessário computar sua versão simetrizada. A divergência KL simetrizada entre $p(x)$ e $q(x)$ é dada por:

$$D_{KL}^{sym}(p,q) = \frac{1}{4\sigma_1^2\sigma_2^2}\left[\left(\sigma_1^2 - \sigma_2^2\right)^2 + (\mu_1-\mu_2)^2\left(\sigma_1^2 + \sigma_2^2\right)\right] \qquad (47)$$

$$= \frac{\left(\sigma_1^2 - \sigma_2^2\right)^2}{4\sigma_1^2\sigma_2^2} + \frac{(\mu_1-\mu_2)^2\left(\sigma_1^2 + \sigma_2^2\right)}{4\sigma_1^2\sigma_2^2} \qquad (48)$$

$$= \frac{1\left(\sigma_1^2 - \sigma_2^2\right)^2}{4\sigma_1^2\sigma_2^2} + \frac{\left(\sigma_1^2 + \sigma_2^2\right)(\mu_1-\mu_2)^2}{4\sigma_1^2\sigma_2^2} \qquad (49)$$

Uma generalização da entropia de Shannon é a conhecida entropia de Renyi (van Erven; Harremos, 2014):

$$H_R^\alpha(p) = \frac{1}{1-\alpha}log\left(\int p(x)^\alpha dx\right) \qquad (50)$$

em que no caso limite de $\alpha \to 1$, tem-se a entropia de Shannon original. Da mesma forma que no caso da entropia de Shannon, a entropia de Renyi relativa é a divergência de Renyi de ordem α, uma generalização da divergência KL:

$$D_R^\alpha(p,q) = \frac{1}{\alpha-1}log\left(\int \frac{p(x)^\alpha}{q(x)^{\alpha-1}}dx\right) \qquad (51)$$

Pode-se mostrar que no caso Gaussiano, a divergência de Renyi de ordem α pode ser computada pela seguinte expressão fechada (GIL; ALAJAJI; LINDER, 2013):

$$D_R^{\alpha}(p,q) = log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2(\alpha-1)} log\left(\frac{\sigma_2^2}{\sigma_\alpha^2}\right) + \frac{1}{2}\left(\frac{\alpha(\mu_1-\mu_2)^2}{\sigma_\alpha^2}\right) \tag{52}$$

com $\sigma_\alpha^2 = \alpha\sigma_2^2 + (1-\alpha)\sigma_1^2 > 0$.

Note que quando $\alpha = 2$, a entropia de Renyi torna-se:

$$H_R^2(p) = -log\left(\int p(x)^2 dx\right) \tag{53}$$

também conhecida como entropia quadrática.

O equivalente da divergência KL para a entropia quadrática é a divergência de Cauchy-Schwarz (CS) (HOANG et al., 2015):

$$\begin{aligned} D_{CS}(p,q) &= -log\frac{\int p(x)q(x)dx}{\sqrt{\int p(x)^2dx \int q(x)^2dx}} \\ &= \frac{1}{2}log\left(\int p(x)^2dx\right) + \frac{1}{2}log\left(\int q(x)^2dx\right) - log\left(\int p(x)q(x)dx\right) \end{aligned} \tag{54}$$

No modelo Gaussiano univariado (Spurek; Palka, 2016):

$$D_{CS}(p,q) = \frac{1}{2}log\left(\frac{(\sigma_1^2+\sigma_2^2)^2}{4\sigma_1^2\sigma_2^2}\right) + \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} \tag{55}$$

A entropia de Sharma-Mittal generaliza a entropia de Renyi, adicionando um segundo parâmetro, denotado por β (NIELSEN; NOCK, 2011a):

$$H_{SM}^{\alpha,\beta}(p) = \frac{1}{1-\beta}\left[\left(\int p(x)^{\alpha}dx\right)^{\frac{1-\beta}{1-\alpha}} - 1\right] \tag{56}$$

com $\alpha, \beta > 0$, $\alpha \neq 1 \neq \beta$ e $\alpha \neq \beta$. Pode-se mostrar que no caso limite de $\beta \to 1$ a entropia de Sharma-Mittal converge para a entropia de Renyi. De forma análoga aos casos anteriores, a entropia de Sharma-Mittal relativa origina a divergência de Sharma-Mittal:

$$D_{SM}^{\alpha,\beta}(p,q) = \frac{1}{\beta-1}\left[\left(\int p(x)^{\alpha}q(x)^{1-\alpha}\right)^{\frac{1-\beta}{1-\alpha}} - 1\right] \tag{57}$$

para qualquer $\alpha > 0$, $\alpha \neq 1$ e $\beta \neq 1$. Pode-se mostrar que para $\alpha, \beta \to 1$, a divergência de Sharma-Mittal converge para a divergência KL. Pode-se mostrar que no caso Gaussiano univariado a divergência de Sharma-Mittal é dada por (NIELSEN; NOCK, 2011a):

$$D_{SM}^{\alpha,\beta}(p,q) = \frac{1}{\beta-1}\left\{\left[\alpha\left(\frac{\sigma_2^2}{\sigma_1^2}\right)^{1-\alpha} + (1-\alpha)\left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{\alpha}\right]^{-\frac{1}{2}\frac{(1-\beta)}{(1-\alpha)}} \times exp\left\{-\frac{1}{2}\frac{(1-\beta)}{(1-\alpha)}\frac{(\mu_1-\mu_2)^2}{\overline{\sigma^2}}\right\} - 1\right\} \tag{58}$$

com

$$\overline{\sigma^2} = \frac{\sigma_1^2\sigma_2^2}{\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2} > 0 \tag{59}$$

Outra generalização da entropia de Shannon é a entropia de Tsallis, proposta a partir de estudos de sistemas físicos multi-fractais e pertencente a uma família de funções entrópicas derivadas axiomaticamente por Havrda e Charvat (HAVRDA; CHARVAT, 1967), sendo definida por (TSALLIS, 1988):

$$H_{TS}^{\alpha}(p) = \frac{1}{1-\alpha}\left[\int p(x)^{\alpha}dx - 1\right] \tag{60}$$

com $\alpha \neq 1$. Utilizando a noção de entropia relativa, é possível definir a divergência de Tsallis entre duas distribuições como:

$$D_{TS}^{\alpha}(p,q) = \frac{1}{\alpha-1}\left[\int p(x)^{\alpha}q(x)^{1-\alpha}dx - 1\right] \tag{61}$$

Pode-se mostrar que no caso Gaussiano univariado, existe uma fórmula fechada para a divergência de Tsallis, dada por (NIELSEN; NOCK, 2011b):

$$D_{TS}^{\alpha}(p,q) = \frac{1}{1-\alpha}\left[exp\{-J_F^{\alpha}(\theta_1,\theta_2)\} - 1\right] \tag{62}$$

em que $J_F^{\alpha}(\theta_1,\theta_2)$ é dado por:

$$J_F^{\alpha}(\theta_1,\theta_2) = \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2) \tag{63}$$

com

$$F(\theta_1) = \frac{\mu_1^2}{2\sigma_1^2} + \frac{1}{2}log(2\pi\sigma_1^2) \tag{64}$$

$$F(\theta_2) = \frac{\mu_2^2}{2\sigma_2^2} + \frac{1}{2}log(2\pi\sigma_2^2) \tag{65}$$

$$F(\alpha\theta_1 + (1-\alpha)\theta_2) = \frac{\hat{\mu}}{2\hat{\sigma^2}} + \frac{1}{2}log(2\pi\hat{\sigma^2}) \tag{66}$$

onde

$$\hat{\mu} = \alpha\mu_1 + (1-\alpha)\mu_2 \tag{67}$$

$$\hat{\sigma^2} = \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 \tag{68}$$

As divergências de Kullback-Leibler, Renyi, Cauchy-Schwarz, Sharma-Mittal e Tsallis são derivadas a partir de generalizações da entropia de Shannon e são denominadas divergências entrópicas. Devido ao fato da entropia relativa não ser simétrica, as distâncias[1] de Bhattacharyya, Hellinger e Total Variation são alternativas às divergências entrópicas.

Denotando por $D_B(p,q)$ a distância de Bhattacharyya entre as distribuições p e q,

$$D_B(p,q) = -log\ BC(p,q) \tag{69}$$

em que $BC(p,q)$ é o coeficiente de Bhattacharyya, definido em termos das distribuições $p(x|\vec{\theta}_i)$ e $q(x|\vec{\theta}_j)$ como:

$$BC(p,q) = \int \sqrt{p(x|\vec{\theta}_i)q(x|\vec{\theta}_j)}dx \tag{70}$$

Assumindo o modelo Gaussiano univariado, tem-se:

$$BC(p,q) = \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}exp\left\{-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right\} \tag{71}$$

A distância de Bhattacharyya possui um papel importante na derivação de limites superiores para a probabilidade de erro na classificação Bayesiana. Pode-se mostrar que, em problemas binários de classificação sob hipótese Gaussiana (DUDA; HART; STORK, 2000)

$$p(erro) \leq \epsilon_{1/2} = \sqrt{p(\omega_1)p(\omega_2)}e^{-k(1/2)} \tag{72}$$

onde $k(1/2)$ é a distância de Bhattacharyya entre as distribuições.

Vale ressaltar que a distância de Hellinger também é função do coeficiente de Bhattacharyya:

$$H^2(p,q) = \frac{1}{2}\int\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2 dx = \frac{1}{2}\left(2-2BC(p,q)\right) = 1-BC(p,q) \tag{73}$$

Dadas duas distribuições de probabilidade $p(x)$ e $q(x)$, a distância Total Variation (TV) é definida por:

$$D_{TV}(p,q) = \frac{1}{2}\int|p(x)-q(x)|dx \tag{74}$$

e pode ser calculada para distribuições discretas como (Verdu, 2014):

$$D_{TV}(p,q) = \frac{1}{2}\sum_{i=1}^{n}|p_i-q_i| \tag{75}$$

[1] Divergências são definidas especificamente em distribuições de probabilidade, enquanto distâncias podem ser definidas em outros objetos. Todas as medidas de distância entre distribuições de probabilidade são divergências, mas uma divergência pode ou não ser uma distância. Uma distância deve satisfazer os axiomas que definem uma métrica (positividade, simetria, desigualdade triangular).

Pode-se mostrar que para o caso Gaussiano univariado, é possível se calcular $D_{TV}(p,q)$ pela seguinte expressão (Nielsen; Sun, 2018):

$$D_{TV}(p,q) = \frac{1}{2}\left|erf\left(\frac{x_1-\mu_1}{\sigma_1\sqrt{2}}\right) - erf\left(\frac{x_1-\mu_2}{\sigma_2\sqrt{2}}\right)\right| + \frac{1}{2}\left|erf\left(\frac{x_2-\mu_1}{\sigma_1\sqrt{2}}\right) - erf\left(\frac{x_2-\mu_2}{\sigma_2\sqrt{2}}\right)\right| \tag{76}$$

onde $erf(x)$ denota uma função de erro definida por:

$$erf(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} exp\{-t^2\}dt \tag{77}$$

e x_1 e x_2 são as duas soluções para a equação quadrática $ax^2 + bx + c = 0$, onde os coeficientes são dados por:

$$a = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \tag{78}$$

$$b = 2\left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) \tag{79}$$

$$c = \left(\frac{\mu_1}{\sigma_1}\right)^2 - \left(\frac{\mu_2}{\sigma_2}\right)^2 + 2log\left(\frac{\sigma_1}{\sigma_2}\right) \tag{80}$$

Uma observação importante é a de que, a distância Total Variation é limitada no intervalo $[0,1]$ (CANONNE, 2022).

As distâncias de Bhattacharyya, Hellinger e Total Variation são denominadas divergências paramétricas. Tanto divergências entrópicas quanto paramétricas são consideradas divergências estocásticas (i.e. não-determinísticas), pois são medidas de similaridade entre distribuições de probabilidade de variáveis aleatórias (i.e. mensuram o quão próxima uma variável está de outra).

Divergências entrópicas são baseadas portanto na entropia de Shannon, ideia esta pertencente ao campo da teoria da informação. Entretanto, existe um segundo paradigma de divergências que se baseiam na informação de Fisher e são categorizadas como divergências geodésicas. Tais divergências pertencem a um sub-campo da teoria da informação denominado geometria da informação.

Capítulo 6

Geometria da informação

O campo de pesquisa denominado geometria da informação é um ramo da teoria da informação que fornece um tratamento geométrico para diversos modelos paramétricos, sendo assim responsável por estudar as relações entre distribuições de probabilidade e propriedades geométricas. Nesse contexto, é possível investigar como duas variáveis aleatórias independentes de um modelo paramétrico se relacionam em termos geométricos (ARWINI; DODSON, 2008).

A geometria da informação é desenvolvida pela aplicação de métodos teóricos de geometria diferencial ao estudo da estatística matemática. Esse campo tem sido uma área de pesquisa relevante desde os trabalhos de Amari (1985), tendo sido expandido e explorado com sucesso por pesquisadores em uma ampla gama de áreas da ciência, desde física estatística e mecânica quântica até teoria dos jogos e aprendizado de máquina. O ponto de partida da geometria da informação é a demonstração de que o espaço paramétrico define uma variedade Riemanniana (RAO, 1945).

6.1 Variedades de Riemann e distância geodésica

O conhecido espaço Euclidiano caracteriza uma variedade sem curvatura, onde a distância entre dois pontos em duas dimensões é dada pelo comprimento de reta que perpassa os pontos (Figura 6) e cuja forma analítica é dada por

$$ds^2 = dx^2 + dy^2 \tag{81}$$

Entretanto, de maneira informal, uma variedade pode representar um espaço curvo. Nesse caso, a distância entre dois pontos é calculada pelo comprimento do arco entre os

mesmos, conforme ilustrado pela Figura 7.

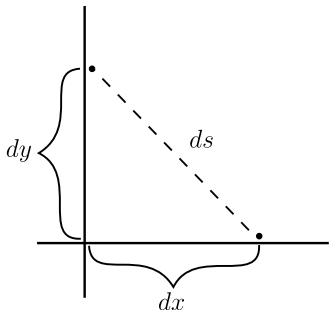


Figura 6 – Distância no espaço Euclidiano

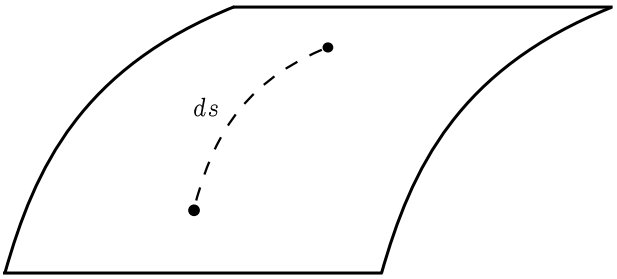


Figura 7 – Representação de distância em um espaço curvo

Tal comprimento é obtido de forma local adaptativa através de sucessivos deslocamentos infinitesimais. Para isso, incorpora-se no cálculo da distância, uma matriz de pesos responsável por caracterizar a curvatura local ao se caminhar sobre a superfície, onde tal matriz assume valores distintos em cada ponto da variedade. Essa estrutura matemática é chamada de tensor métrico do espaço.

O tensor métrico define produtos internos nos espaços tangentes locais e torna possível expressar o quadrado de um deslocamento infinitesimal na variedade em função de um deslocamento infinitesimal no espaço tangente que, no caso de uma variedade 2D, é dado por um vetor em $[du, dv]$. Assumindo uma notação matricial, temos:

$$ds^2 = \begin{bmatrix} du & dv \end{bmatrix} \begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} du \\ dv \end{bmatrix} = Adu^2 + 2Bdudv + Cdv^2 \tag{82}$$

onde a matriz simétrica de coeficientes A, B, C é o tensor métrico. Se a matriz é definida positiva, então a variedade é conhecida como Riemanniana. Note que, quando o tensor métrico é a matriz identidade, a expressão (82) se iguala à expressão (81), ou seja, o espaço Euclidiano é de fato uma variedade que não apresenta curvatura pois possui a matriz identidade como tensor.

Suponha que $p(x; \vec{\theta})$ é um modelo estatístico pertencente à família exponencial, onde $\vec{\theta}$ denota o vetor de parâmetros do modelo. Então, a coleção de todos os vetores admissíveis $\vec{\theta}$ define o espaço paramétrico Θ. O espaço paramétrico define uma variedade Riemanniana que, no caso Gaussiano univariado, é descrita por (NIELSEN, 2022):

$$\Omega = \{(\mu, \sigma^2) : \mu \in R, \sigma^2 > 0\} \tag{83}$$

A distância infinitesimal entre duas variáveis aleatórias $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$ é mapeada portanto como sendo a distância entre dois pontos na variedade:

$$ds^2 = \begin{bmatrix} d\mu & d\sigma^2 \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} d\mu \\ d\sigma^2 \end{bmatrix} = Ad\mu^2 + 2Bd\mu d\sigma^2 + Cd(\sigma^2)^2 \tag{84}$$

Amari (1985) demonstra que, para qualquer tipo de distribuição de probabilidade, o tensor métrico da variedade do espaço paramétrico é dado pela matriz de informação de Fisher.

6.2 Informação de Fisher

A informação de Fisher (FISHER, 1922) é uma medida de incerteza análoga à entropia de Shannon, com a diferença de se basear em uma função de verossimilhança em vez de se basear em probabilidades. A informação de Fisher mede a quantidade de informação que uma amostra aleatória transmite sobre um parâmetro desconhecido. A informação de Fisher é definida pelo valor esperado do produto das derivadas parciais em cada parâmetro da função log-verosimilhança. Considerando uma função de distribuição de probabilidade $p(x; \vec{\theta})$, onde $\vec{\theta} \in R^k$, a matriz de informação de Fisher é definida como (CASELLA; BERGER, 2001):

$$I(\vec{\theta})_{ij} = E\left[\left(\frac{\partial}{\partial\theta_i} log\ p(x;\vec{\theta})\right)\left(\frac{\partial}{\partial\theta_j} log\ p(x;\vec{\theta})\right)\right] = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} log\ p(x;\vec{\theta})\right] \tag{85}$$

Mostra-se que a igualdade acima é válida sob certas condições de regularidade. Observe que:

$$E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} log\ p(x;\vec{\theta})\right] = E\left[\frac{\partial}{\partial\theta_j}\left(\frac{\partial}{\partial\theta_i} log\ p(x;\vec{\theta})\right)\right] = E\left[\frac{\partial}{\partial\theta_j}\left(\frac{1}{p(x;\vec{\theta})}\frac{\partial}{\partial\theta_i} p(x;\vec{\theta})\right)\right] \tag{86}$$

Pela regra do produto, temos:

$$\begin{aligned} &E\left[\frac{\partial}{\partial\theta_j}\left(\frac{1}{p(x;\vec{\theta})}\frac{\partial}{\partial\theta_i}p(x;\vec{\theta})\right)\right] = E\left[-\frac{1}{p(x;\vec{\theta})^2}\frac{\partial}{\partial\theta_i}p(x;\vec{\theta})\frac{\partial}{\partial\theta_j}p(x;\vec{\theta}) + \frac{1}{p(x;\vec{\theta})}\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\vec{\theta})\right] \\ &= -E\left[\left(\frac{1}{p(x;\vec{\theta})}\frac{\partial}{\partial\theta_i}p(x;\vec{\theta})\right)\left(\frac{1}{p(x;\vec{\theta})}\frac{\partial}{\partial\theta_i}p(x;\vec{\theta})\right)\right] + E\left[\frac{1}{p(x;\vec{\theta})}\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\vec{\theta})\right] \end{aligned} \tag{87}$$

Pela definição de valor esperado, o segundo termo da Equação (87), pode ser simplificado para:

$$E\left[\frac{1}{p(x;\vec{\theta})}\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\vec{\theta})\right] = \int p(x;\vec{\theta})\frac{1}{p(x;\vec{\theta})}\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\vec{\theta})dx = \int \frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\vec{\theta})dx \quad (88)$$

Sob certas condições de regularidade é possível trocar os operadores de integração e diferenciação:

$$\int \frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\vec{\theta})dx = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int p(x;\vec{\theta})dx = \frac{\partial^2}{\partial\theta_i\partial\theta_j}1 = 0 \quad (89)$$

note também que:

$$\frac{1}{p(x;\vec{\theta})}\frac{\partial}{\partial\theta_i}p(x;\vec{\theta}) = \frac{\partial}{\partial\theta_i}log\ p(x;\vec{\theta}) \quad (90)$$

o que finalmente leva à igualdade:

$$E\left[\left(\frac{\partial}{\partial\theta_i}log\ p(x;\vec{\theta})\right)\left(\frac{\partial}{\partial\theta_j}log\ p(x;\vec{\theta})\right)\right] = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}log\ p(x;\vec{\theta})\right] \quad (91)$$

Considerando o caso Gaussiano univariado, onde $\vec{\theta} = (\mu, \sigma^2)$ a log-verossimilhança é dada por:

$$log\ p(x;\vec{\theta}) = -\frac{1}{2}log\ (2\pi) - \frac{1}{2}log\ \sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2 \quad (92)$$

e a primeira e segunda derivadas em relação a μ são iguais a:

$$\frac{\partial}{\partial\mu}log\ p(x;\vec{\theta}) = \frac{1}{\sigma^2}(x-\mu) \quad (93)$$

$$\frac{\partial^2}{\partial\mu^2}log\ p(x;\vec{\theta}) = -\frac{1}{\sigma^2} \quad (94)$$

o que leva a:

$$-E\left[\frac{\partial^2}{\partial\mu^2}log\ p(x;\vec{\theta})\right] = \frac{1}{\sigma^2} \quad (95)$$

As derivadas cruzadas de segunda ordem são iguais a zero, ou seja:

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}log\ p(x;\vec{\theta}) = \frac{\partial^2}{\partial\sigma^2\partial\mu}log\ p(x;\vec{\theta}) = 0 \quad (96)$$

A primeira e a segunda derivadas em relação a σ^2 são dadas por:

$$\frac{\partial}{\partial\sigma^2}log\ p(x;\vec{\theta}) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2 \quad (97)$$

$$\frac{\partial^2}{\partial(\sigma^2)^2}log\ p(x;\vec{\theta}) = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(x-\mu)^2 \quad (98)$$

O cálculo do valor esperado da segunda derivada resulta em:

$$-E\left[\frac{\partial}{\partial\sigma^2} log\ p(x;\vec{\theta})\right] = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}E\left[(x-\mu)^2\right] = -\frac{1}{2\sigma^4} + \frac{1}{\sigma^4} = \frac{1}{2\sigma^4} \tag{99}$$

Assim, a matriz de informação de Fisher do modelo é:

$$I(\vec{\theta}) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \tag{100}$$

Para outras distribuições de probabilidade distintas da Gaussiana univariada, a matriz de informação de Fisher pode não possuir forma fechada.

6.3 Distância simétrica de Fisher

Portanto, $I(\vec{\theta})$ (100) é o tensor métrico do espaço paramétrico Gaussiano univariado e assim podemos expressar o deslocamento infinitesimal (84) na variedade em termos da matriz de informação de Fisher (100):

$$ds^2 = \begin{bmatrix} d\mu & d\sigma^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \begin{bmatrix} d\mu \\ d\sigma^2 \end{bmatrix} = \frac{1}{\sigma^2}d\mu^2 + \frac{1}{2\sigma^4}(d\sigma^2)^2 \tag{101}$$

Nesta configuração, podemos aproximar a divergência $D_F(p,q)$ entre duas Gaussianas univariadas $p(x;\vec{\theta}_1)$ e $p(x;\vec{\theta}_2)$ para $\vec{\theta}_1 = (\mu_1, \sigma_1^2)$ e $\vec{\theta}_2 = (\mu_2, \sigma_2^2)$ como:

$$\begin{aligned} D_F(p,q) &= \Delta\vec{\theta}^T I(\vec{\theta})\Delta\vec{\theta} = \left(\vec{\theta}_1 - \vec{\theta}_2\right)^t I(\vec{\theta}) \left(\vec{\theta}_1 - \vec{\theta}_2\right) \\ &= \frac{1}{\sigma_1^2}(\mu_1 - \mu_2)^2 + \frac{1}{2\sigma_1^4}(\sigma_1^2 - \sigma_2^2)^2 \end{aligned} \tag{102}$$

De modo similar, $D_F(q,p)$ é dado por:

$$\begin{aligned} D_F(q,p) &= \Delta\vec{\theta}^T I(\vec{\theta})\Delta\vec{\theta} = \left(\vec{\theta}_2 - \vec{\theta}_1\right)^t I(\vec{\theta}) \left(\vec{\theta}_2 - \vec{\theta}_1\right) \\ &= \frac{1}{\sigma_2^2}(\mu_2 - \mu_1)^2 + \frac{1}{2\sigma_2^4}(\sigma_2^2 - \sigma_1^2)^2 \end{aligned} \tag{103}$$

e podemos definir a distância simétrica de Fisher (SFD) $D_{F_s}(p,q)$ como:

$$D_{F_s}(p,q) = \frac{1}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)(\mu_1 - \mu_2)^2 + \frac{1}{4}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)(\sigma_1^2 - \sigma_2^2)^2 \tag{104}$$

$$= \frac{1}{2}\left(\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2\sigma_2^2}\right)(\mu_1 - \mu_2)^2 + \frac{1}{4}\left(\frac{\sigma_1^4 + \sigma_2^4}{\sigma_1^4\sigma_2^4}\right)(\sigma_1^2 - \sigma_2^2)^2 \tag{105}$$

$$= \frac{(\sigma_1^4 + \sigma_2^4)\left(\sigma_1^2 - \sigma_2^2\right)^2}{4\sigma_1^4\sigma_2^4} + \frac{(\sigma_1^2 + \sigma_2^2)\left(\mu_1 - \mu_2\right)^2}{2\sigma_1^2\sigma_2^2} \tag{106}$$

Capítulo 7

Flexibilidade da abordagem à divergência

Os capítulos anteriores compõem a fundamentação teórica necessária para o entendimento do intuito deste trabalho: primeiramente apresentamos medidas para se avaliar a qualidade de uma dispersão do ponto de vista da compactação intra-grupo e separação inter-grupos; posteriormente mostramos a formulação da abordagem contextual paramétrica aplicada ao algoritmo PCA; por fim deduzimos as formulações das divergências estatísticas envolvidas no processo.

Outros trabalhos da literatura (LEVADA, 2020; LEVADA, 2021) indicaram que a proposta apresenta potencial competitivo a diversos algoritmos de mapeamento existentes quando da utilização das divergências Kullback-Leibler, Bhattacharyya e Hellinger. Conforme introduzido na Seção 1.2, neste trabalho aprofundamos o estudo da proposta através da investigação de dois objetivos:

1. Verificar se a abordagem contextual paramétrica no método PCA é de fato flexível à escolha da divergência, no sentido de não apresentar resultados significativamente inferiores ao se adotar alguma divergência específica.

2. Estudar os efeitos da escolha da divergência através da identificação de possíveis propriedades de um conjunto de amostras que indicariam maior adequação de uma divergência em detrimento de outra.

Neste capítulo detalharemos a metodologia de investigação do Objetivo 1. O capítulo é dividido em duas partes: a primeira relata os experimentos sobre a sensibilidade da proposta à escolha da divergência e a segunda exibe resultados obtidos ao se aumentar o

rigor experimental da primeira. O conteúdo da primeira parte se encontra publicado em **Nakao e Levada (2024)** e as discussões da segunda parte se encontram publicadas em **Nakao e Levada (2023)**.

Em um primeiro momento, investigamos a sensibilidade do PCA contextual paramétrico à escolha da divergência. Testamos a estabilidade do método através da avaliação das divergências Total Variation, Renyi, Sharma-Mittal e Tsallis apresentadas no Capítulo 5. Comparamos o desempenho da abordagem sob tais divergências contra os seguintes algoritmos já existentes:

- ❏ Original PCA (PCA) (JOLLIFFE, 2002)
- ❏ Joint-Sparse PCA (JSPCA) (YI et al., 2017)
- ❏ Kernel PCA (KPCA) (SCHöLKOPF; SMOLA; MüLLER, 1999)
- ❏ Robust PCA (RPCA) (CANDèS et al., 2011)
- ❏ L1-based PCA (L1PCA) (MARKOPOULOS et al., 2017)
- ❏ Locally Linear Embedding (LLE) (ROWEIS; SAUL, 2000)
- ❏ Isometric Feature Mapping (ISOMAP) (TENENBAUM; SILVA; LANGFORD, 2000)
- ❏ Laplacian Eigenmaps (LAP) (BELKIN; NIYOGI, 2003)

Foram utilizados os trinta conjuntos reais de amostras descritos[1] na Tabela 1. Vale notar que tais conjuntos exibem diferenças significativas na quantidade de características, amostras e classes. Devido às diferentes escalas de valores que as características podem apresentar, o processo de normalização foi aplicado antes da execução do mapeamento.

Nesse primeiro momento por simplicidade, fixamos a dimensionalidade-alvo para o plano ($d = 2$). Na definição do parâmetro k de vizinhança, utilizamos uma busca linear guiada por desempenho. Tendo em vista a necessidade de um balanço entre o tamanho amostral e a preservação de localidade, fixamos uma janela de incremento baseada na quantidade n de amostras, fazendo com que menores valores para k fossem adotados em conjuntos de poucas amostras, na tentativa de preservar a localidade de vizinhança.

Na Tabela 2 encontram-se os valores da medida SC para cada algoritmo de mapeamento, onde as últimas quatro colunas correspondem às divergências Total Variation, Tsallis, Sharma-Mittal e Renyi. Um valor negrito significa que o método correspondente demonstrou superioridade em relação a todos os outros métodos. Um valor sublinhado significa que o método obteve o segundo melhor resultado. Ao final da tabela encontram-se os valores do SC médio, o desvio-padrão, a mediana e o desvio absoluto médio (MAD).

[1] Todos os conjuntos e suas descrições estão disponíveis em <openml.org>

Tabela 1 – conjuntos de dados: quantidade de amostras, características e classes

Conjunto	amostras	características	classes
analcatdata_happiness	60	3	3
breast-tissue	106	9	4
molecular-biology_promoters	106	57	2
ar4	107	29	2
ar1	121	29	2
fruitfly	125	4	2
mux6	128	6	2
datatrieve	130	8	2
transplant	131	3	2
hayes-roth	132	4	2
iris	150	4	3
analcatdata_wildcat	163	5	2
servo	167	4	2
KnuggetChase3	194	39	2
pwLinear	200	10	2
machine_cpu	209	6	2
heart-statlog	270	13	2
heart-h	294	13	5
vertebra-column	310	6	3
diggle_table_a2	310	8	2
visualizing_galaxy	323	4	2
Engine1	383	5	3
mw1	403	37	2
kc3	458	39	2
sa-heart	462	9	2
thoracic_surgery	470	16	2
pm10	500	7	2
rmftsa_ladata	508	10	2
threeOf9	512	9	2
arsenic-male-lung	559	4	2
arsenic-female-bladder	559	4	2
strikes	625	6	2
Australian	690	14	2
blood-transfusion-service-center	748	4	2
diabetes	768	8	2
stock	950	9	2
car	1728	6	2
pc3	1563	37	2
mfeat-fourier	2000	76	10
kc1	2109	21	2
bank-marketing	4521	16	2
page-blocks	5473	10	2
first-order-theorem-proving	6118	51	6
delta_ailerons	7129	5	2
mammography	11183	6	2

Tabela 2 – medida SC das dispersões geradas por cada algoritmo

	L1PCA	PCA	KPCA	ISO	LLE	LAP	JSPCA	RPCA	TVPCA	TPCA	SMPCA	RenPCA
iris	0.303	0.401	0.469	0.452	0.365	0.541	0.470	0.551	0.426	0.414	0.457	**0.615**
blood	0.006	0.086	0.026	0.082	0.008	0.004	0.092	0.083	**0.124**	**0.179**	**0.170**	**0.175**
kc1	0.283	0.371	0.210	0.187	0.187	-0.459	0.370	0.369	**0.442**	**0.457**	**0.462**	**0.466**
Australian	0.297	0.279	0.276	0.291	0.130	0.346	0.272	0.312	0.321	**0.490**	0.323	**0.433**
transplant	0.333	0.485	0.436	0.486	0.410	0.438	0.480	0.520	0.506	**0.521**	**0.540**	**0.541**
servo	0.029	0.121	0.105	0.114	0.104	0.085	0.120	**0.279**	0.130	0.221	0.151	0.217
analcatdata	0.051	0.151	0.081	0.125	0.149	0.028	0.170	0.107	**0.175**	**0.172**	**0.188**	**0.208**
datatrieve	0.119	0.239	0.011	0.096	0.066	0.081	0.236	0.174	**0.261**	**0.257**	**0.262**	**0.264**
machine_cpu	0.395	0.498	0.399	0.492	0.496	0.410	0.494	**0.575**	0.504	0.556	0.509	0.505
arsenic-female	0.220	0.122	0.008	0.170	0.143	0.030	0.104	0.068	0.217	**0.252**	0.212	0.215
page-blocks	0.432	0.419	0.218	0.527	0.581	0.436	0.426	0.419	**0.595**	0.564	**0.701**	**0.638**
arsenic-male	0.788	0.563	-0.182	0.674	0.697	-0.057	0.504	0.057	0.729	**0.790**	0.728	0.737
mw1	0.287	0.349	0.122	0.286	0.175	0.18	0.337	0.346	**0.423**	**0.423**	**0.423**	**0.428**
car	0.080	0.029	**0.189**	0.046	0.163	0.079	0.01	0.068	0.176	0.183	0.178	0.182
ar1	0.174	0.265	0.028	0.216	-0.004	-0.002	0.276	0.246	**0.388**	**0.363**	**0.429**	**0.441**
diggle_table	0.273	0.406	0.409	0.450	0.328	0.304	0.407	0.444	**0.470**	**0.484**	**0.467**	**0.489**
rmftsa_ladata	0.008	0.228	0.242	0.238	0.185	0.230	0.225	0.236	**0.276**	**0.269**	**0.291**	**0.299**
kc3	0.423	0.386	0.103	0.233	0.045	-0.129	0.394	0.394	**0.516**	**0.560**	**0.544**	**0.570**
diabetes	0.059	0.117	0.100	0.115	0.101	0.054	0.111	0.106	**0.140**	**0.132**	**0.145**	0.113
mammography	0.518	0.349	0.032	0.307	0.070	-0.251	0.348	0.349	**0.628**	**0.645**	**0.653**	**0.643**
bank-marketing	0.032	0.082	-0.006	-0.001	0.078	-0.257	0.082	0.082	**0.233**	**0.331**	**0.292**	**0.321**
heart-h	0.007	0.056	0.041	0.076	0.087	-0.004	0.066	0.134	**0.152**	**0.189**	**0.172**	**0.207**
molecular	-0.009	0.106	0.134	0.138	0.035	0.137	0.105	0.170	0.121	**0.252**	**0.236**	**0.254**
delta_ailerons	0.349	0.117	0.341	0.383	0.077	0.419	0.114	0.117	0.365	0.314	**0.437**	**0.469**
pc3	0.071	0.201	0.074	-0.017	-0.003	-0.341	0.201	0.188	**0.229**	**0.226**	**0.223**	**0.230**
ar4	0.143	0.357	0.176	0.318	0.203	0.131	0.361	0.356	**0.463**	**0.443**	**0.467**	**0.492**
KnuggetChase3	0.195	0.199	0.070	0.187	0.077	0.091	0.196	0.203	**0.290**	**0.363**	**0.312**	**0.324**
threeOf9	0.095	0.034	0.017	0.049	0.095	0.044	0.048	0.029	**0.191**	**0.190**	**0.192**	**0.193**
galaxy	0.125	0.179	0.255	0.193	0.235	0.270	0.177	0.219	**0.278**	0.265	**0.272**	**0.279**
thoracic_surgery	0.048	0.006	-0.002	-0.006	0.082	-0.021	0.008	-0.075	**0.213**	**0.277**	**0.247**	**0.287**
Média	0.204	0.240	0.146	0.230	0.179	0.094	0.238	0.238	**0.333**	**0.359**	**0.356**	**0.375**
Desvio-padrão	0.186	0.156	0.154	0.178	0.174	0.240	0.153	0.167	0.167	0.164	0.170	0.169
Mediana	0.159	0.215	0.104	0.190	0.117	0.080	0.213	0.211	**0.284**	**0.323**	**0.302**	**0.323**
MAD	0.150	0.134	0.124	0.143	0.127	0.181	0.132	0.138	0.142	0.136	0.146	0.148

Conclui-se dos resultados que, tanto na média quanto na maioria dos conjuntos, as quatro divergências apresentaram um SC superior a todos os outros métodos. Para testar se as diferenças obtidas entre os valores da medida em um dado par de algoritmos foram significativas, o teste de Wilcoxon (WILCOXON, 1945) com significância $\alpha = 1\%$ foi aplicado para confirmar a superioridade em relação ao segundo melhor algoritmo. Portanto os testes de sensibilidade indicam que o PCA sob abordagem paramétrica é flexível à escolha da divergência, sendo capaz de gerar resultados competitivos em diversos conjuntos de amostras para uma ampla gama de divergências.

Experimentos adicionais

Em um segundo momento, um procedimento experimental mais rigoroso foi adotado. Dois algoritmos de mapeamento mais modernos foram adicionados na comparação: *t-Distributed Stochastic Neighbor Embedding* (t-SNE) (MAATEN; HINTON, 2008) e *Uniform Manifold Approximation and Projection* (UMAP) (MCINNES; HEALY; MELVILLE, 2020). Adicionalmente, tanto a dimensionalidade-alvo quanto a vizinhança foram definidos através de uma busca exaustiva guiada por desempenho, isto é, o valor de SC representativo de cada algoritmo foi o melhor obtido na variação exaustiva tanto do parâmetro d quanto do parâmetro k. Essas modificações no protocolo experimental aumentam o grau de competitividade entre os métodos.

Utilizamos nesse segundo momento uma nova divergência ainda não testada. Devido ao aumento na complexidade experimental trazido pela adição de mais métodos na comparação e pela variação exaustiva de ambos parâmetros e, conjuntamente ao resultado de que a abordagem é flexível à divergência, utilizamos somente a divergência de Cauchy-Schwarz (CS) nessa segunda etapa. Executamos os experimentos em quinze conjuntos reais de amostras. Na Tabela 3 nota-se que em aproximadamente metade dos casos (sete), o PCA contextual paramétrico com a divergência CS (CSPCA) foi o método mais adequado. Nota-se também que CSPCA exibiu a maior média para o valor de SC. Isso mostra que o método foi individualmente superior a um outro método comparado para diversos conjuntos de amostras, e também foi superior a todos os outros métodos comparados para alguns conjuntos de amostras.

Adicionalmente, realizamos uma comparação do ponto de vista dos recursos computacionais. A Tabela 4 exibe os tempos de execução[2] dos algoritmos baseados em grafo de vizinhança para o conjunto *mfeat-fourier* (2000 amostras e 76 características). A respeito da alocação de memória de processamento, CSPCA consumiu apenas 4,8GB enquanto qualquer outro algoritmo exigiu ao menos 12GB (t-SNE representando o pior caso consumindo 24GB). Portanto, ao considerar tanto o tempo quanto o espaço de processamento exigido, é possível concluir que a nova abordagem é computacionalmente competitiva aos algoritmos baseados em grafo de vizinhança comparados.

[2] Contabilizando a busca exaustiva pelos parâmetros (d e k) ótimos

Finalizamos reiterando uma observação feita na Seção 1.1, sobre a importância da existência de múltiplas técnicas de mapeamento, tendo em vista que cada uma é mais adequada a um determinado conjunto de amostras do que a outra. Entretanto, a identificação dessa adequação em conjuntos reais só pode ser obtida experimentalmente. Assim, os experimentos e resultados apresentados corroboram essa ideia e mostram que não há um método único superior aos demais em todos os conjuntos.

Tabela 3 – valores da medida SC para a segunda série experimental

	PCA	KPCA	ISO	LLE	LAP	SPCA	RPCA	t-SNE	UMAP	CSPCA
hayes-roth	-0.015	0.056	0.042	0.106	0.087	0.031	0.030	0.067	0.101	**0.140**
pwLinear	0.182	0.199	0.272	0.339	0.301	0.195	0.482	0.334	0.350	**0.499**
pm10	0.000	0.003	0.003	0.004	0.007	0.001	0.000	0.006	0.006	**0.015**
threeOf9	0.064	0.062	0.193	0.159	0.224	0.061	0.130	0.140	0.122	**0.308**
strikes	0.030	0.018	0.033	0.034	0.029	0.029	0.019	0.033	0.029	**0.113**
analcatdata	-0.026	-0.020	-0.021	-0.008	0.001	-0.026	-0.072	-0.008	-0.023	**0.076**
breast-tissue	0.004	0.000	0.009	0.014	0.020	0.004	0.000	0.013	0.008	**0.031**
fruitfly	-0.006	-0.003	0.010	0.027	**0.086**	-0.006	0.001	0.053	0.028	-0.001
mux6	0.090	0.062	0.125	0.125	0.112	0.090	0.045	**0.140**	0.117	0.112
transplant	0.511	0.529	0.529	0.617	0.627	0.512	0.520	0.582	**0.748**	0.583
iris	0.606	0.469	0.607	0.673	**0.666**	0.603	0.577	0.564	0.657	0.639
vertebra-column	0.119	0.186	0.121	0.162	0.184	0.117	0.121	0.205	**0.269**	0.121
engine1	-0.095	**0.003**	-0.065	-0.021	-0.021	-0.102	-0.067	**0.003**	-0.023	-0.050
sa-heart	0.096	0.061	0.128	**0.162**	0.159	0.094	0.075	0.130	0.121	0.134
mfeat-fourier	0.082	0.088	0.139	0.119	0.200	0.081	0.082	0.205	**0.266**	0.136
Média	0.109	0.114	0.142	0.168	0.179	0.112	0.130	0.164	0.185	**0.190**

Tabela 4 – tempos de execução no conjunto de dados *mfeat-fourier*

	ISO	LLE	LAP	t-SNE	UMAP	CSPCA
mfeat-fourier	13h57m	42h10m	4h21m	1h02m	15h21m	4h18m

Capítulo 8

Adequação da divergência à dispersão

No capítulo anterior apresentamos experimentos que indicam a flexibilidade do PCA contextual paramétrico à escolha da divergência, resultado este que valida a primeira hipótese de pesquisa e conclui o Objetivo 1 deste trabalho. Um fato notável dos experimentos é o de que, em alguns conjuntos, resultados significativamente distintos são gerados quando diferentes divergências são empregadas, indicando assim que certas divergências são mais adequadas para certos conjuntos do que outras.

Desse fato surge a necessidade de se identificar uma relação de adequação entre divergência e dispersão. Isto é, identificar alguma propriedade de um conjunto de amostras que se relacione ao comportamento da divergência, de modo a justificar a qualidade da dispersão gerada pelo PCA contextual paramétrico quando determinada divergência é empregada. Este capítulo é dedicado portanto à investigação da segunda hipótese de pesquisa norteada pelo Objetivo 2 do trabalho.

8.1 Estudo do comportamento das divergências

Uma maneira de se estudar o comportamento de uma divergência pode ser obtida da análise de sua taxa de crescimento quando calculada sobre uma série de pares de distribuições Gaussianas univariadas hipotéticas, isto é, em uma série de permutações dos parâmetros (μ, σ^2). Tal procedimento quando repetido para cada divergência gera diferentes curvas de crescimento. Nos experimentos do capítulo anterior, a normalização prévia à aplicação do mapeamento sempre é executada. Assim, os valores para μ foram limitados no intervalo $[0, 1]$ para a geração de tais curvas.

Note que quando a série está contida nesse intervalo, o valor máximo de σ^2 é igual a 0.25. Assim, os pares de distribuições são gerados através da seguinte variação de parâmetros: médias em $[0, 1]$ com incremento de 0.1 e variâncias em $[0, 0.25]$ com incremento de 0.05. Dessa forma, cada divergência é calculada sobre cada par de distribuições, formando assim uma coleção de valores que, após ordenada, permite a geração do gráfico das curvas de crescimento exibido na Figura 8a (cada tonalidade representando uma divergência).

Nota-se que a Divergência Simétrica de Fisher (SFD) (geodésica) possui uma taxa de crescimento significativamente superior às demais (estocásticas). Dessa observação, exibimos na Figura 8b o gráfico contendo somente as divergências estocásticas. Nota-se que a divergência entrópica de Tsallis e as divergências paramétricas de Bhattacharyya, Hellinger e Total Variation são limitadas em $[0, 1]$. Já as demais divergências (entrópicas) apresentam comportamentos intermediários. O gráfico da Figura 8c exibe portanto somente as divergências limitadas.

8.2 Efeito do crescimento da divergência na abordagem

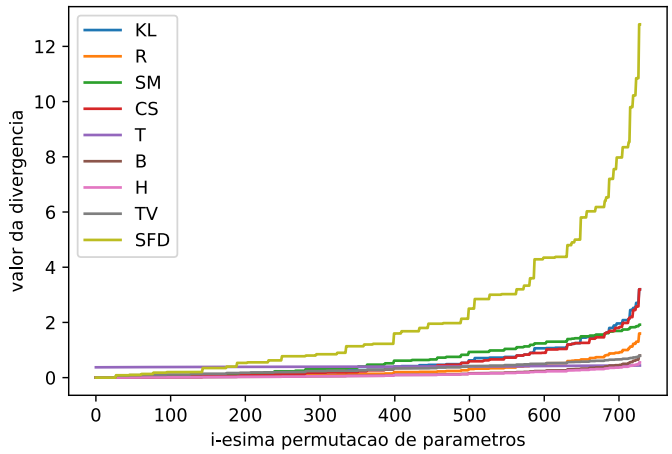
O próximo passo consiste em estudar o efeito prático dos diferentes comportamentos de crescimento das divergências no PCA contextual paramétrico. Para isso, primeiramente vamos retomar as formulações das matrizes de covariância usual (Σ) e paramétrica (C):

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)^t \tag{107}$$

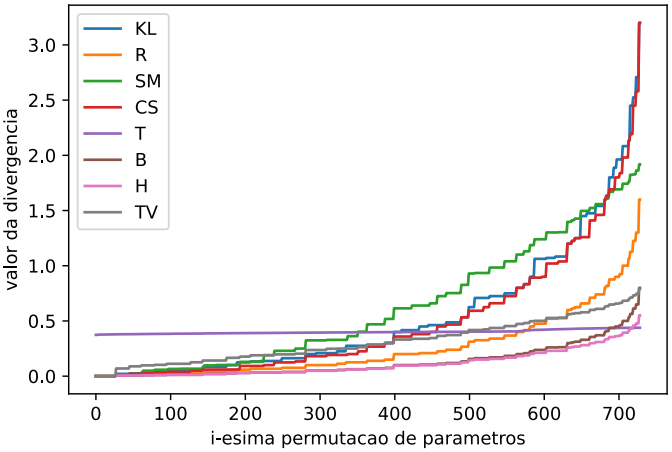
$$C = \frac{1}{n-1}\sum_{i=1}^{n}\vec{d}(\vec{p}_i, \tilde{\vec{p}})\vec{d}(\vec{p}_i, \tilde{\vec{p}})^t \tag{108}$$

Tendo em mente as ideias apresentadas nos Capítulos 3 e 4, a primeira observação a ser considerada é a de que, na prática, o PCA contextual paramétrico é implementado através do fornecimento da matriz C para o PCA original. Isso significa que o efeito das etapas intermediárias de mapeamento das amostras para o espaço paramétrico e a subsequente obtenção da matriz C, equivale a um pré-mapeamento da dispersão original no espaço usual para uma nova pseudo-dispersão no mesmo espaço usual, onde essa nova pseudo-dispersão é fornecida como entrada ao PCA original.

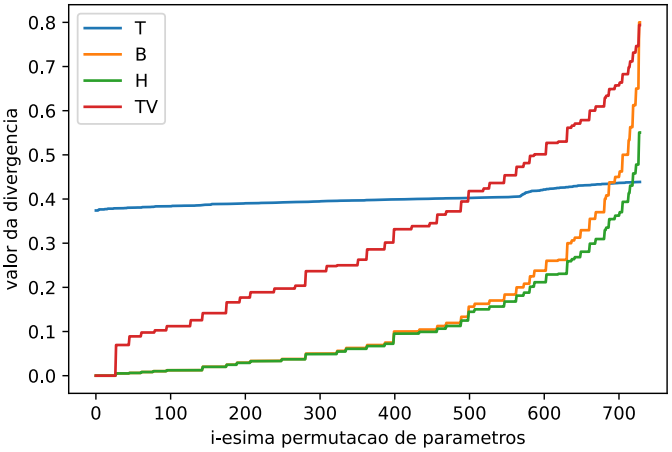
Com essa observação em mente, é possível se pensar no efeito dos diferentes comportamentos de crescimento das divergências. Note que uma divergência de menor taxa de crescimento gera portanto menores valores para as coordenadas do vetor $\vec{d}(\vec{p}_i, \tilde{\vec{p}})$ envolvido no cálculo da matriz C. Unindo esse fato à observação do parágrafo anterior de que as matrizes C e Σ são análogas, tendo em vista que a componente $(x_i - \mu_x)$ de Σ representa a variância e portanto caracteriza uma medida de espalhamento das amostras



(a) Curvas de crescimento de todas as divergências



(b) Curvas de crescimento das divergências estocásticas



(c) Curvas de crescimento das divergências limitadas

Figura 8 – Curvas de crescimento das divergências geodésica e estocásticas

em relação ao centro da dispersão (definido pela amostra média), conclui-se que quanto menor o crescimento da divergência, menor o espalhamento global das amostras na nova pseudo-dispersão.

Quando o espalhamento global da pseudo-dispersão é baixo, é natural se esperar que o espalhamento interno de cada grupo correspondente a cada classe também seja menor, o que melhoraria a qualidade da pseudo-dispersão. Entretanto, esse mesmo cenário de menor espalhamento global também pode aproximar amostras de classes distintas, o que degradaria o critério de separação inter-classes.

Para motivar essas ideias, considere o caso de uma divergência limitada em $[0, 1]$. Note que nesse caso $\vec{d}(\vec{p}_i, \tilde{\vec{p}})$ é um vetor cujas coordenadas são valores escalares em $[0, 1]$, o que caracteriza um mapeamento para uma pseudo-dispersão contida em uma hiperesfera de raio 1 centralizada na origem. Esse fenômeno implica em uma compactação dos dados que possivelmente aproximaria amostras de classes distintas, prejudicando assim o critério de separação inter-classes.

Com essas intuições em mente, podemos então enunciar uma hipótese sobre a influência da taxa de crescimento de uma divergência (e o consequente espalhamento das amostras) na qualidade da dispersão final. Denotando por r o limite do subespaço da pseudo-dispersão, Qs o valor da medida de separação inter-classes e Qe o valor da medida de espalhamento intra-classe, podemos formular a primeira sub-hipótese do Objetivo 2:

> *O crescimento do valor de r implica no crescimento tanto da separação inter-classes Qs quanto do espalhamento intra-classe Qe (e vice-versa).*

8.3 Influência da quantidade de classes na qualidade da dispersão

Uma segunda observação é a de que, além do espalhamento das amostras, um outro fator que também influencia a qualidade de uma dispersão é a quantidade de classes. Considere por exemplo o caso extremo da Figura 9 onde a quantidade de classes é a maior possível, sendo igual à quantidade de amostras ($c = n$). Nesse caso, o espalhamento intra-classe é mínimo (ótimo), pois cada classe corresponderia à somente uma amostra, e a separação inter-classes seria mínima (péssima), pois cada amostra definiria uma classe, gerando assim a máxima mistura de classes.

Figura 9 – Dispersão de máxima quantidade de classes

De forma inversa, podemos pensar no caso oposto, onde a dispersão possui somente uma classe, como ilustrado na Figura 10. Nesse caso o espalhamento intra-classe seria máximo (péssimo) e a separação inter-classes seria máxima (ótima). Com essas intuições em mente, podemos formular a segunda hipótese do Objetivo 2:

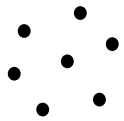


Figura 10 – Dispersão de mínima quantidade de classes

> *O crescimento do valor de c implica na diminuição tanto da separação inter-classes Qs quanto do espalhamento intra-classe Qe (e vice-versa).*

8.4 Motivações empíricas para as hipóteses

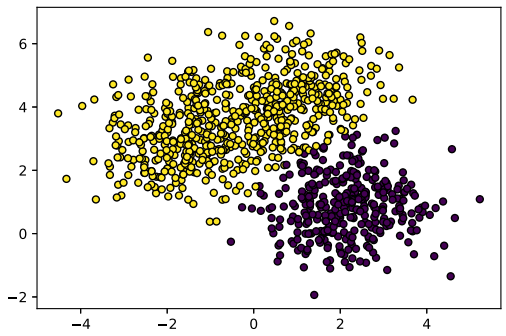
Esta seção apresentará motivações empíricas para as hipóteses construídas. Tais motivações serão fornecidas através da análise dos valores de Qs e Qe ao se variar os parâmetros r e c em uma dispersão bidimensional sintética. A Figura 11 exibe as dispersões e partições analisadas. Utilizamos a função *make_blobs* para gerar uma dispersão não rotulada de três nuvens Gaussianas isotrópicas. Em um primeiro momento fixamos o raio de geração ($r = 5$) e particionamos a dispersão gerada com o algoritmo *SpectralClustering* (SHI; MALIK, 2000) em duas, três e quatro classes, produzindo o particionamento dos itens 11a, 11b e 11c respectivamente.

Em um segundo momento, variamos o raio de geração e fixamos o particionamento em três classes, produzindo assim as dispersões dos itens 11d, 11e e 11f. Sob cada dispersão calculamos os critérios de coesão e separação de cada medida de qualidade descrita no Capítulo 2. A Tabela 5 apresenta os valores obtidos nos critérios quando da variação do raio da dispersão e a Tabela 6 apresenta os valores obtidos nos critérios quando da variação da quantidade de classes. Resumimos a seguir sem simbologia matemática os critérios das medidas.

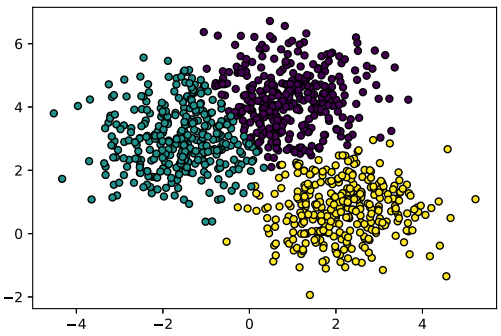
R-Squared

- espalhamento intra-classe (Qe): soma das distâncias quadráticas entre uma amostra de um grupo e o centro do grupo

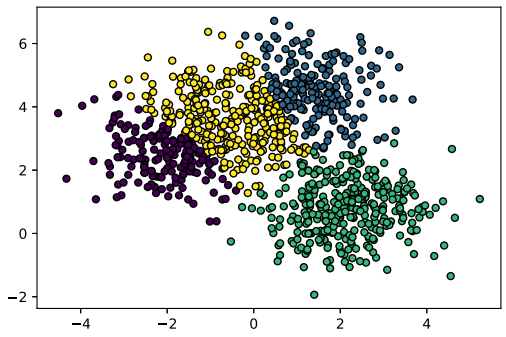
- separação inter-classes (Qs): soma das distâncias quadráticas entre uma amostra e o centro do conjunto de amostras



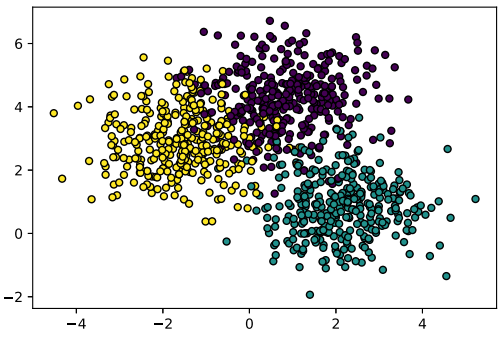
(a) $r = 5 \mid c = 2$



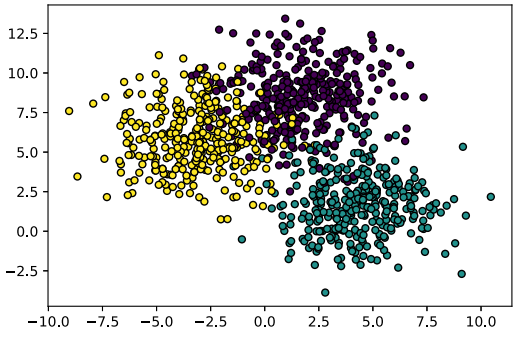
(b) $r = 5 \mid c = 3$



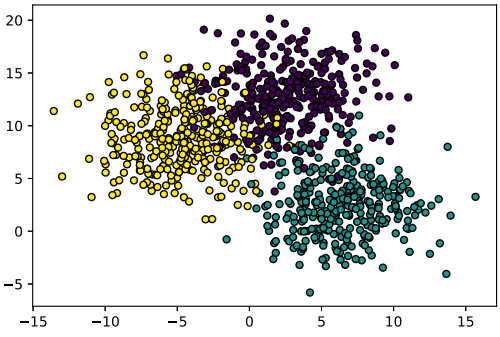
(c) $r = 5 \mid c = 4$



(d) $r = 5 \mid c = 3$



(e) $r = 10 \mid c = 3$



(f) $r = 15 \mid c = 3$

Figura 11 – Dispersões bidimensionais geradas

Tabela 5 – valores dos critérios das medidas de qualidade em função de r

r	medida	Qe	Qs	Qs'
5	RS	1898	6189	-
10	RS	7594	24759	-
15	RS	17086	55709	-
5	S	573	3.339	-
10	S	1147	6.678	-
15	S	1721	10.017	-
5	CH	1.904	2135	-
10	CH	7.617	8541	-
15	CH	17.130	19217	-
5	DB	1.235	6.673	4.231
10	DB	2.470	13.346	8.462
15	DB	3.705	20.019	12.694

Tabela 6 – valores dos critérios das medidas de qualidade em função de c

c	medida	Qe	Qs	Qs'
2	RS	3041	6189	-
3	RS	1687	6189	-
4	RS	1470	6189	-
2	S	1072	4.219	-
3	S	542	3.376	-
4	S	376	2.779	-
2	CH	3.047	3510	-
3	CH	1.692	2242	-
4	CH	1.476	1596	-
2	DB	1.743	3.871	3.747
3	DB	1.183	6.787	4.343
4	DB	1.151	11.036	4.563

Silhouette

- ❑ espalhamento intra-classe (Qe): distância média entre uma amostra e todas as amostras de seu grupo

- ❑ separação inter-classes (Qs): distância média entre uma amostra e todas as amostras do grupo mais próximo

Calinski-Harabasz

- ❏ espalhamento intra-classe (Qe): distância quadrática entre uma amostra e o centro de seu grupo
- ❏ separação inter-classes (Qs): distância quadrática entre os centros dos grupos e o centro do conjunto amostral

Davies-Bouldin

- ❏ espalhamento intra-classe (Qe): valor máximo de **intra**

  intra: distância média entre cada amostra de um grupo e o centro do mesmo
- ❏ separação inter-classes (Qs): valor máximo de **inter**

  inter: distância média entre a amostra de um grupo e o centro de outro grupo
- ❏ separação inter-classes (Qs'): valor máximo de **inter'**

  inter': distância entre os centros de dois grupos

Analisando a Tabela 5 é possível notar que, para todas as medidas, o aumento do raio implica no crescimento tanto dos valores de espalhamento intra-classe quanto dos valores de separação inter-classes, motivando assim a intuição da primeira sub-hipótese.

Já a análise da Tabela 6 nos permite concluir que, o aumento da quantidade de classes implica, na maioria dos casos, na diminuição tanto do espalhamento intra-classe, quanto da separação inter-classes, motivando assim a intuição da segunda sub-hipótese. Adicionalmente observamos o seguinte:

1. na medida RS o valor de separação não se altera com a variação de c;
2. em todas as medidas, o valor de espalhamento parece convergir;
3. na medida DB os valores de separação aumentam com o crescimento de c.

A explicação da observação (1) vem do fato de que o critério de separação RS se baseia somente no raio, não considerando portanto a quantidade de classes em seu cálculo. Com relação à observação (2) note que, ao se particionar um conjunto de amostras em uma quantidade crescente de classes, a curva de decrescimento dos valores de espalhamento intra-classe se comporta de maneira semelhante à função $f(c) = r/c$ (Figura 12) pois, ao se aumentar a quantidade de classes, chega-se no caso limite em que a quantidade de classes é igual a de amostras (i.e. quantidade de amostras por grupo igual a 1), caso este em que a distância das amostras do grupo ao seu centro é igual a 0.

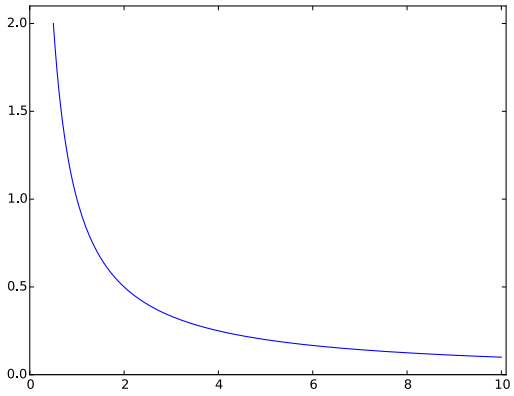


Figura 12 – Curva da função $f(c) = r/c$ para $r = 1$

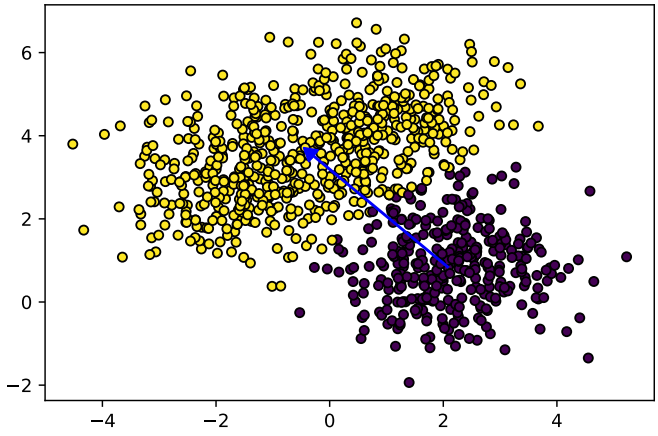
Para se justificar a observação (3), note que o critério Qs' da medida DB é dado pela maior distância entre os centros de um par de grupos. Esse critério é ilustrado na Figura 12. Na Figura 12a, a seta em azul corresponde a distância máxima entre os centros de um par de grupos (que para o caso $c = 2$ é única). Na Figura 12b a seta em vermelho corresponde a distância máxima entre os centros de um par de grupos considerando todos os pares possíveis. Na Figura 12c é possível perceber que o valor da norma da seta em vermelho é de fato superior ao valor de norma da seta em azul, mostrando portanto que o aumento da quantidade de classes implica em uma maior distância máxima entre pares de centros para essa dispersão. O crescimento do critério Qs pode ser justificado por raciocínio análogo, tendo em vista que tal critério é calculado pela distância média entre cada amostra de um grupo e o centro de cada outro grupo[1].

8.5 Relação entre divergência e quantidade de classes

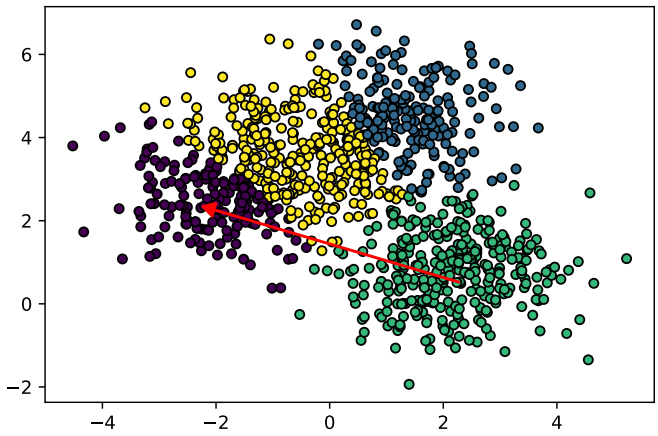
Na seção anterior mostramos que as intuições motivadoras das sub-hipóteses podem ser evidenciadas em dispersões bidimensionais, fortalecendo assim as suspeitas de que: (1) o aumento do raio da pseudo-dispersão gerada por uma divergência de maior valor resultante, pode implicar no aumento tanto do espalhamento intra-classe quanto da separação inter-classes; (2) quanto maior a quantidade de classes, maior a tendência do espalhamento intra-classe e da separação inter-classes diminuírem.

Tais hipóteses quando consideradas em conjunto geram como consequência a expectativa de que o crescimento do raio amenize os efeitos do crescimento da quantidade de classes (e vice-versa). Ou seja, representações mais compactas seriam mais adequadas para conjuntos de menores quantidades de classes (e vice-versa).

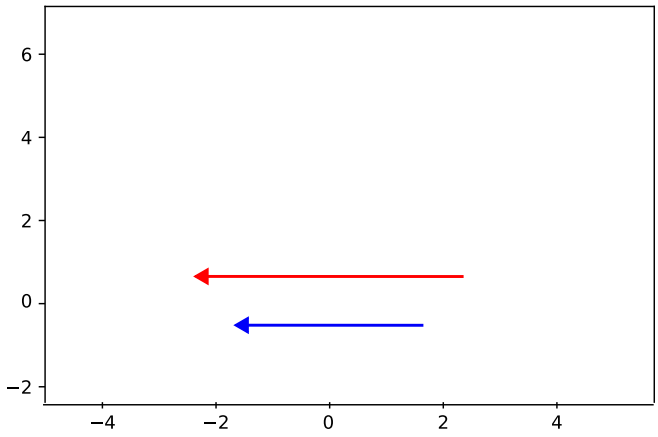
[1] Para não estender a seção para além de seu escopo, caso se deseje entender a diferença de crescimento entre Qs e Qs', basta se verificar o crescimento dos valores no mesmo conjunto experimental.



(a) Qs' para $c = 2$



(b) Qs' para $c = 4$



(c) comparação entre os casos (a) e (b)

Figura 12 – Ilustração do critério Qs' para a medida DB

Na Seção 8.2 mostramos que, quanto menor o valor resultante de uma divergência, mais compacta é a representação gerada. Pelo fato de divergências limitadas gerarem a maior compactação possível no cenário desse estudo, infere-se que as mesmas seriam mais adequadas para problemas com a menor quantidade possível de classes ($c = 2$).

Vimos na Seção 8.1 que a divergência simétrica de Fisher (SFD) apresenta maiores valores, enquanto que as divergências paramétricas apresentam os menores. Portanto, iremos comparar somente a divergência SFD contra uma divergência limitada, devido aos seus comportamentos extremos antagônicos que definem os limites superior e inferior respectivamente. Imaginamos que o estudo das demais curvas intermediárias poderia ser inconclusivo.

Dentre as quatro divergências limitadas, escolhemos a divergência Total Variation (TV) como representativa da categoria pelo fato de possuir a taxa de crescimento mais estável entre as divergências paramétricas, e também pelo fato de já possuirmos resultados experimentais nos conjuntos reais de amostras.

8.6 Investigação das hipóteses em conjuntos reais de amostras

Da seção anterior conclui-se que, é esperado que a divergência TV seja mais adequada para conjuntos de duas classes, enquanto que a divergência SFD seja mais adequada para conjuntos de mais de duas classes. Para verificar essa hipótese, aplicamos o PCA contextual paramétrico com ambas divergências em diversos conjuntos reais de amostras e avaliamos com as medidas descritas no Capítulo 2 a qualidade das dispersões geradas.

Realizada essa etapa, identificamos os conjuntos onde as divergências implicaram em dispersões de qualidade significativamente distintas de acordo com alguma medida. Como resultado dessa etapa os seguintes conjuntos foram identificados: *heart-statlog, Australian, threeOf9, iris, mfeat-fourier, cardiotocography.*

Cada item da Tabela 7 apresenta os valores da qualidade de dispersão resultantes de uma medida de avaliação nas dispersões geradas por ambas as divergências em tais conjuntos[2]. Cada tabela também exibe a quantidade de classes de cada conjunto. Em cada linha destacamos o melhor valor de qualidade apresentado[3]. É possível observar que na maioria dos pares de células, a superioridade é exibida em primeira casa decimal.

A análise das tabelas nos mostra que, para todas as medidas de avaliação, existe uma divisão bem definida onde: (a) todos (e somente) os conjuntos de amostras onde a divergência TV apresentou superioridade possuem duas classes e, (b) todos (e somente) os conjuntos onde a divergência SFD se mostrou superior possuem mais de duas classes. Tais resultados indicam que as hipóteses se confirmam em cenários reais.

[2] Adotando a melhor dispersão variando-se exaustivamente d e k.
[3] A medida DB em particular indica melhores dispersões em menores valores (vide Capítulo 2).

Tabela 7 – qualidade da dispersão gerada por cada divergência (valores por medida)

(a)

Silhouette	c	SFD	TV
heart-statlog	2	0.136	**0.315**
Australian	2	0.191	**0.330**
threeOf9	2	0.137	**0.216**
iris	3	**0.667**	0.412
mfeat-fourier	10	**0.175**	0.064
cardiotocography	10	**0.321**	0.237

(b)

R-squared	c	SFD	TV
heart-statlog	2	0.133	**0.427**
Australian	2	0.141	**0.419**
threeOf9	2	0.170	**0.303**
iris	3	**0.938**	0.766
visualizing-livestock	5	**0.288**	0.145
mfeat-fourier	10	**0.729**	0.669
cardiotocography	10	**0.641**	0.496

(c)

Calinski-Harabasz	c	SFD	TV
heart-statlog	2	0.041	**0.200**
Australian	2	0.113	**0.497**
threeOf9	2	0.104	**0.222**
iris	3	**1.123**	0.240
visualizing-livestock	5	**0.013**	0.005
mfeat-fourier	10	**0.597**	0.447
cardiotocography	10	**0.421**	0.232

(d)

Davies-Bouldin	c	SFD	TV
heart-statlog	2	0.245	**0.094**
Australian	2	0.271	**0.094**
threeOf9	2	0.183	**0.123**
iris	3	**0.045**	0.078
visualizing-livestock	5	**1.169**	7.128
mfeat-fourier	10	**0.440**	0.546
cardiotocography	10	**0.126**	0.143

8.7 Considerações finais

Nesse capítulo exibimos os comportamentos de cada divergência e escolhemos as divergências de maiores e menores valores resultantes para comparação. Mostramos que quanto maior o valor retornado pela divergência, maior o espalhamento global da pseudo-dispersão gerada pela abordagem contextual paramétrica. Levantamos a hipótese de que

um maior espalhamento global pode implicar em um maior espalhamento intra-classe e uma maior separação inter-classes. Supomos também que uma maior quantidade de classes pode implicar tanto em um menor espalhamento intra-classe quanto uma menor separação inter-classes. Exibimos experimentos em dispersões bidimensionais que motivaram tais intuições. Unificamos essas ideias através da relação de que divergências de menores valores resultantes seriam mais adequadas para problemas de menor quantidade de classes. Por fim, apresentamos resultados em conjuntos reais que validaram tal relação, concluindo assim o Objetivo 2 do trabalho.

Capítulo 9

Trabalhos futuros e relacionados

9.1 Trabalhos futuros

A abordagem atual representa o conjunto de valores das características da vizinhança de uma amostra por uma distribuição Gaussiana univariada, descrita pelos parâmetros média e variância. A substituição da média pela mediana e do desvio padrão pelo desvio absoluto mediano poderia amenizar a influência de amostras distantes de seus grupos. Outra alternativa nesse sentido seria a adoção de um modelo de cauda mais pesada e consequentemente menos sensível à amostras distantes de seus grupos, como a distribuição t-Student por instância. Note entretanto que essa alternativa exigiria o ajuste de um parâmetro adicional (graus de liberdade) que aumenta a complexidade experimental.

Ainda sobre a adequação do modelo Gaussiano univariado, a diferença semântica entre as características pode fazer com que o conjunto de valores a ser mapeado para uma distribuição exiba um comportamento multimodal. Além disso, conjuntos com uma quantidade relativamente baixa de amostras poderiam exibir vizinhanças locais limitadas. Nesse sentido, imaginamos que a Estimativa de Densidade por Kernel (ROSENBLATT, 1956; PARZEN, 1962) poderia ser mais adequada do que o modelo Gaussiano univariado. Note que essa alternativa também exige o ajuste de parâmetros adicionais, dado que múltiplas funções de kernel e larguras de banda poderiam ser testadas.

Conforme mencionado no Capítulo 4, a distância Euclidiana é empregada na construção do grafo KNN pela justificativa da assunção de vizinhança linear (ROWEIS; SAUL, 2000). Entretanto, outras métricas poderiam ser adotadas, tais como as distâncias de Jaccard, Minkowski e Cosseno. Note que a estratégia de se fixar um k e adotar os k vizinhos mais próximos, aumenta a chance de uma vizinhança incluir amostras de classes distintas. Nesse sentido, algumas alternativas se apresentam, como por exemplo a definição de um

raio de vizinhança (ϵ-*ball*), ou o ajuste do tamanho da vizinhança de maneira adaptativa através da análise da matriz Hessiana local, cujo intuito seria definir vizinhanças maiores em regiões de baixa curvatura, e definir vizinhanças menores em regiões de alta curvatura. Aqui também observamos que esse procedimento aumenta o custo computacional do método.

Uma terceira alternativa seria uma abordagem supervisionada, que incorporaria na vizinhança somente amostras da mesma classe. Note que nessa alternativa, métricas de avaliação de tarefas de classificação seriam necessárias em substituição às métricas de qualidade de dispersão. Após o processo de mapeamento, um classificador poderia ser investigado e métricas como acurácia, *precision*, *recall* e *F1-score* seriam utilizadas para avaliar a tarefa de classificação. Esse protocolo poderia ser empregado para se comparar a abordagem contextual paramétrica supervisionada no PCA contra outros algoritmos de mapeamento supervisionados já existentes, tais como *Linear Discriminant Analysis*, *Supervised PCA* e *Partial Least Squares*.

Por fim, uma observação importante: a adoção da abordagem contextual paramétrica não se restringe ao método PCA. Sua aplicação pode ser estudada em qualquer método de mapeamento, tais como ISOMAP, LLE, LAP, t-SNE, LDA. Trabalhos futuros poderiam utilizar o protocolo experimental do Objetivo 1 para comparar o desempenho da abordagem em cada método contra suas propostas originais.

9.2 Trabalhos relacionados

Na literatura de reconhecimento de padrões (LI; TIAN, 2018; WANG; SUN, 2015) encontram-se diferentes propostas de se mapear um conjunto de amostras na tentativa de capturar sua estrutura geométrica intrínseca da melhor maneira possível, preservando as relações originais de vizinhança e gerando uma dispersão de melhor qualidade, facilitando assim a análise dos padrões subjacentes. Este capítulo de maneira particular apresentará os algoritmos de mapeamento que também partem do grafo de vizinhança como estrutura inicial e que foram utilizados na investigação do Objetivo 1 deste trabalho. Apresentaremos a ideia principal de cada método seguida de seus passos algorítmicos acompanhados de um esquema ilustrativo. Fornecemos a referência ao trabalho original de cada proposta caso o leitor deseje mais detalhes.

9.2.1 Isometric Feature Mapping

A ideia geral do algoritmo *Isometric Feature Mapping* (ISOMAP) (TENENBAUM; SILVA; LANGFORD, 2000) é a de implementar um mapeamento isométrico das características, isto é, um mapeamento cujo objetivo é aprender a métrica de similaridade intrínseca à dispersão original a fim de preservá-la na dispersão-alvo. A motivação do método vem da observação de que a métrica Euclidiana pode não ser adequada como medida

de similaridade em dispersões de geometria curva, como a apresentada na Figura 13-A (TENENBAUM; SILVA; LANGFORD, 2000). O método se baseia na hipótese de que a medida de similaridade intrínseca nesse tipo de situação pode ser aproximada pelo caminho mínimo no grafo dos k vizinhos mais próximos entre o par de vértices correspondente a um par de amostras (Figura 13-B).

A ideia geral do algoritmo se baseia portanto na construção do grafo KNN a partir do conjunto de amostras, seguida do cálculo dos menores caminhos entre cada par de vértices, para com isso encontrar um mapeamento que preserve as distâncias fornecidas por esses caminhos (Figura 13-C). O método pode ser resumido pelos seguintes passos ilustrados na Figura 14 (WEINBERGER; SAUL, 2006) e detalhados no Algoritmo 2.

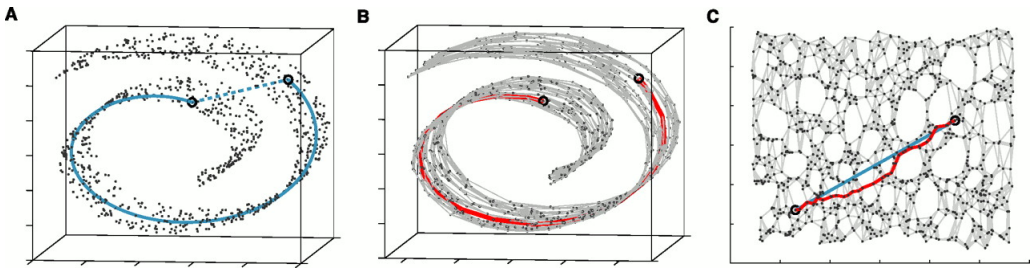


Figura 13 – Caminho mínimo entre vértices como aproximação da métrica intrínseca

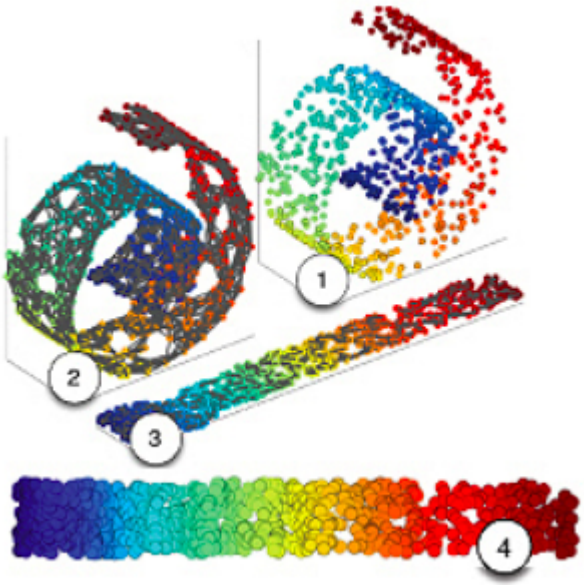


Figura 14 – Ilustração dos passos do algoritmo ISOMAP

1. Induzir o grafo KNN a partir do conjunto de amostras.

2. Construir a matriz D de distâncias correspondentes aos caminhos mínimos entre cada par de vértices utilizando o algoritmo de Dijkstra (DIJKSTRA, 1959).

3. Encontrar o mapeamento para a nova dispersão preservando as distâncias de D utilizando a técnica *Multidimensional Scaling* (MDS) (COX; COX, 2001).

Algoritmo 2 Isometric Feature Mapping

0: **function** ISOMAP(X)
0: Construir o grafo KNN a partir do conjunto de amostras $X_{m\times n}$.
0: Obter a matriz de distâncias $D_{n\times n}$ entre cada par de vértices.
0: Calcular $A = -\frac{1}{2}D$.
0: Calcular $H = I - \frac{1}{n}U$, onde U é uma matriz $n \times n$ de 1's.
0: Calcular $B = HAH$.
0: Encontrar os autovetores e autovalores de B.
0: Selecionar os $d < m$ autovetores e autovalores de B e definir:

$$\tilde{V} = \begin{bmatrix} | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \\ \vec{v}_1 & \vec{v}_2 & \dots & \dots & \vec{v}_d \\ | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \end{bmatrix}_{n\times d} \tag{109}$$

$$\tilde{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_d) \tag{110}$$

0: Calcular $\tilde{X} = \tilde{\Lambda}^{1/2}\tilde{V}^T$
0: **return** $\tilde{X}$
0: **end function** =0

9.2.2 Locally Linear Embedding

O aprendizado da métrica intrínseca de uma dada dispersão constitui a filosofia central dos métodos de mapeamento. O ISOMAP adota uma abordagem global pelo fato de utilizar em seu processo todas as amostras da dispersão em conjunto. Contudo, uma interpretação local também pode ser adotada, onde apenas a localidade em torno de cada amostra é considerada. O método da Imersão Localmente Linear (*Locally Linear Embedding* - LLE) (ROWEIS; SAUL, 2000) parte da hipótese de que a k-vizinhança de uma amostra define uma localidade linear em um plano tangente como representado na Figura 15 (LEE; VERLEYSEN, 2007).

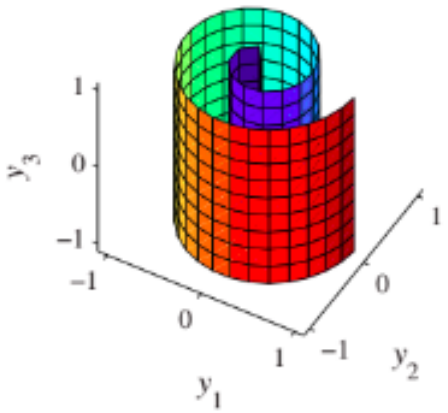


Figura 15 – Representação de vizinhanças por planos locais

Com essa assunção, a distância Euclidiana é utilizada como medida de similaridade entre amostras vizinhas. Com isso, é possível descrever uma amostra como combinação linear de suas vizinhas através de coeficientes lineares

$$x_i \approx \sum_j w_{ij} x_j \tag{111}$$

para x_j na vizinhança de x_i, conforme ilustrado na Figura 16 (ROWEIS; SAUL, 2000). Tais coeficientes são os pesos de reconstrução linear de uma amostra em função de suas vizinhas. Fixados esses pesos, o método obtém as novas coordenadas que preservam tais pesos na nova representação. O processo pode ser sumarizado pelas três etapas a seguir e a técnica completa é descrita pelo Algoritmo 3.

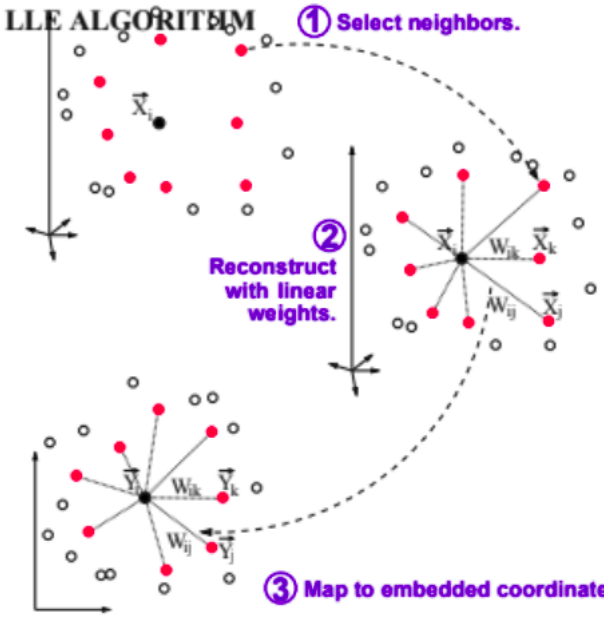


Figura 16 – Ilustração dos passos do algoritmo LLE

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

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

$$E(W) = \sum_{i=1}^{n} \| \, x_i - \sum_j w_{ij} x_j \, \|^2 \tag{112}$$

onde $w_{ij} = 0$ a menos que x_j seja um dos k vizinhos mais próximos de x_i.

3. Para cada i, $\sum_j w_{ij} = 1$ encontre as coordenadas Y que minimizem o erro de reconstrução usando os pesos ótimos

$$\Phi(Y) = \sum_{i=1}^{n} \| \, y_i - \sum_j w_{ij} y_j \, \|^2 \tag{113}$$

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

Algoritmo 3 Locally Linear Embedding

0: **function** LLE(X, k, d)
0: Construir o grafo KNN da dispersão de entrada $X_{m \times n}$.
0: **for** $\vec{x}_i \in X^T$ **do**
0: Calcular a matriz C_i $k \times k$:

$$C_i(j,k) = (\vec{x}_i - \vec{x}_j)^T(\vec{x}_i - \vec{x}_k) \tag{114}$$

0: Resolver o sistema linear $C_i\vec{w}_i = \vec{1}$ para estimar os pesos $\vec{w}_i \in R^k$.
0: Normalizar os pesos em $\vec{w}_i$ para que $\sum_j \vec{w}_i(j) = 1$.
0: **end for**
0: Construir a matriz W $n \times n$, cujas linhas são os $\vec{w}_i$ estimados.
0: Calcular $M = (I - W)^T(I - W)$.
0: Encontrar os autovetores e autovalores de M.
0: Selecionar os d menores autovetores não-nulos de M e definir a matriz Y, onde cada coluna é um autovetor.

$$Y = \begin{bmatrix} | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \\ \vec{v}_1 & \vec{v}_2 & \dots & \dots & \vec{v}_d \\ | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \end{bmatrix}_{n \times d} \tag{115}$$

0: **return** Y
0: **end function**=0

9.2.3 Laplacian Eigenmaps

O ISOMAP é considerado um método global por utilizar todo o conjunto para estimar a proximidade entre duas amostras. Já o LLE é considerado um método local por utilizar somente a vizinhança de uma amostra para mapear tal localidade na dispersão de saída. Esses algoritmos foram os primeiros métodos propostos na literatura que utilizam o grafo do conjunto de amostras como ferramenta de captura da estrutura geométrica para a realização do mapeamento.

Um terceiro algoritmo proposto alguns anos depois é o método da Imersão Laplaciana (BELKIN; NIYOGI, 2002; BELKIN; NIYOGI, 2003). Assim como o LLE, o *Laplacian Eigenmaps* preserva localidades, fazendo com que amostras originalmente próximas permaneçam próximas na nova representação. A intuição mais simples da justificativa de seu funcionamento baseia-se no fato de que, seu critério de minimização garante que a proximidade no grafo de entrada se reflita em proximidade no espaço de saída. Pode-se mostrar que tal critério é expresso em termos de uma forma quadrática envolvendo a matriz Laplaciana L do grafo de entrada.

Como primeiro passo, para capturar a proximidade entre as amostras, a matriz de adjacências W é utilizada. A partir de W, a matriz L é construída pela diferença entre a matriz D e W. A matriz D é uma matriz diagonal onde cada entrada corresponde

ao número de arestas de um nó. O método se baseia portanto na decomposição espectral do operador Laplaciano isto é, na definição dos autovetores associados aos menores autovalores não nulos de L. O Algoritmo 4 exibe o pseudocódigo do processo.

Apesar de ser definido em poucos passos, o método possui um embasamento matemático sofisticado que envolve conceitos de topologia matemática (CHUNG; GRIGOR'YAN; YAU, 2000) e teoria espectral dos grafos (CHUNG, 1997). A teoria espectral dos grafos estuda grafos como objetos matemáticos que fornecem aproximações discretas para espaços não Euclidianos, tais como variedades Riemannianas e superfícies curvas. Nessa área estuda-se o espectro dos grafos, ou seja, os autovalores e autovetores das matrizes que os definem, como por exemplo as matrizes W e L.

O argumento principal do método advém da observação de que o operador Laplace-Beltrani provê uma imersão ótima da variedade subjacente ao conjunto das amostras. O operador Laplaciano é conhecido por suas propriedades espectrais em grafos e variedades. Suas autofunções exibem propriedades desejáveis na imersão, pois o operador captura a variação dos pontos em relação à sua vizinhança local. A imersão seria então uma aproximação discreta de um mapa contínuo naturalmente definido pela geometria da variedade. A variedade é aproximada pelo grafo de adjacência calculado das amostras e a matriz Laplaciana do grafo constitui uma aproximação do operador Laplaciano da variedade.

Em outras palavras, o operador Laplace-Beltrani é aproximado por L dada uma escolha adequada dos pesos. A decomposição espectral de L pode ser interpretada como uma forma de difusão sobre o grafo, dado que a matriz L tem uma conexão com a equação do calor. A escolha do kernel Gaussiano para os pesos do grafo é justificada pelo papel do operador Laplaciano na equação do calor. Tal operador faz com que a escolha do kernel do calor como função de decaimento para a definição dos pesos, seja apropriada.

O entendimento mais profundo e a apresentação dos detalhes matemáticos estão fora do escopo desse capítulo, onde apresentamos somente a ideia principal de cada método acompanhada de seus passos algorítmicos. Caso o leitor deseje mais detalhes, recomendamos a leitura dos trabalhos originais da proposta (BELKIN; NIYOGI, 2002; BELKIN; NIYOGI, 2003) e suas referências.

9.2.4 t-Distributed Stochastic Neighbour Embedding

Para se entender o método t-SNE, é necessário apresentar seu método predecessor e motivador: *Stochastic Neighbor Embedding* (SNE) (HINTON; ROWEIS, 2003). Nesse método, as distâncias entre as amostras $\vec{x}_i$ são convertidas em probabilidades condicionais como forma de representar similaridade. A similaridade entre as amostras $\vec{x}_i$ e $\vec{x}_j$ é dada pela probabilidade condicional $p_{j|i}$ de $\vec{x}_i$ ser vizinho de $\vec{x}_j$ de acordo com uma pdf Gaussiana centrada em $\vec{x}_i$ com variância σ_i^2

Algoritmo 4 Laplacian Eigenmaps

0: **function** LaplaceEigen(X, k, d)
0: Construir o grafo KNN do conjunto de entrada $X_{m\times n}$.
0: Escolha os pesos para definir a matriz de adjacência W.

$$W_{ij} = exp\left\{-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{t}\right\} \qquad \text{se} \qquad v_j \in N(v_i) \tag{116}$$

0: Calcular a matriz diagonal D com os graus d_i para $i = 1, 2, ..., n$.

$$d_i = \sum_{j=1}^{n} W_{ij} \tag{117}$$

0: Calcular a matriz Laplaciana $L = D - W$
0: Selecionar os d menores autovetores de autovalores não-nulos de $D^{-1}L$ e definir a matriz Y, onde cada coluna é um autovetor.

$$Y = \begin{bmatrix} | & | & ... & ... & | \\ | & | & ... & ... & | \\ \vec{v}_1 & \vec{v}_2 & ... & ... & \vec{v}_d \\ | & | & ... & ... & | \\ | & | & ... & ... & | \end{bmatrix}_{n\times d} \tag{118}$$

0: **return** Y
0: **end function**=0

$$p_{j|i} = \frac{\exp\left(-\|\vec{x}_i - \vec{x}_j\|^2 / 2\sigma_i^2\right)}{\sum_{k\neq i} \exp\left(-\|\vec{x}_i - \vec{x}_k\|^2 / 2\sigma_i^2\right)} \tag{119}$$

Note que se $\vec{x}_i$ e $\vec{x}_j$ estão próximos o bastante, então $p_{j|i}$ possui um valor significativamente alto e, se estiverem bem separados, então $p_{j|i}$ tende a zero. É possível calcular uma probabilidade condicional similar $q_{j|i}$ nas novas representações $\vec{y}_i$ e $\vec{y}_j$ para essas amostras. Definindo a variância para $1/\sqrt{2}$, tal medida de similaridade é dada por

$$q_{j|i} = \frac{\exp\left(-\|\vec{y}_i - \vec{y}_j\|^2\right)}{\sum_{k\neq i} \exp\left(-\|\vec{y}_i - \vec{y}_k\|^2\right)} \tag{120}$$

Note que, as similaridades são calculadas par a par e portanto $p_{i|i} = q_{i|i} = 0$. A ideia é a de que as amostras $\vec{y}_i$ e $\vec{y}_j$ na nova representação modelem corretamente as similaridades entre as amostras originais $\vec{x}_i$ e $\vec{x}_j$, e portanto as probabilidades condicionais $p_{j|i}$ e $q_{j|i}$ sejam iguais.

O objetivo do método SNE portanto é mapear as amostras de forma a minimizar a distância entre as duas probabilidades, tornando-as o mais próximas possível. Uma medida estatística de proximidade entre duas distribuições de probabilidade é a divergência de Kullback-Leibler dada pela entropia relativa entre as duas distribuições. O método de

descida do gradiente é utilizado para minimizar a soma das divergências KL sobre todas as amostras. A função objetivo a ser minimizada é

$$C = \sum_{i=1}^{n} KL(P_i||Q_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{j|i}\log\frac{p_{j|i}}{q_{j|i}} \tag{121}$$

onde P_i representa a distribuição de probabilidade sobre todas as outras amostras dada a amostra $\vec{x}_i$, e Q_i representa a distribuição de probabilidade sobre todas as outras amostras dada a amostra $\vec{y}_i$. Portanto, o foco de tal função de custo é preservar a estrutura local das amostras originais para valores razoáveis de variância σ_i^2.

Não há somente um valor possível para σ_i^2 que seja ótimo em todos os casos. Menores valores de σ_i^2 são mais apropriados para regiões densas, enquanto que maiores valores são mais adequados para regiões esparsas. Um dado σ_i^2 induz uma distribuição de probabilidade P_i cuja entropia é dada em função da variância. A chamada medida de perplexidade é definida como (MAATEN; HINTON, 2008):

$$\text{Perp}(P_i) = 2^{H(P_i)} \tag{122}$$

onde $H(P_i)$ é a entropia de Shannon em bits:

$$H(P_i) = -\sum_{j=1}^{n} p_{j|i} log_2 p_{j|i} \tag{123}$$

No método SNE busca-se um valor de σ_i^2 que produza uma P_i de perplexidade fixa definida pelo usuário. A perplexidade pode ser interpretada como uma medida suavizada para a quantidade efetiva de vizinhos. A minimização da função objetivo (divergência KL) é feita através da descida do gradiente com momento para acelerar a conversão, isto é, após a inicialização, as coordenadas das novas amostras são iterativamente atualizadas por:

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} - \eta\frac{\partial C}{\partial \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right) \tag{124}$$

onde $\mathcal{Y}^{(t)}$ denota a solução na iteração t, η denota a taxa de aprendizado e $\alpha(t)$ representa o momento na iteração t.

O SNE é afetado pelo problema de *crowding* (aglomeração), caracterizado pelo fato de que, a área na nova representação utilizada para posicionar amostras originalmente distantes, pode não ser grande o suficiente em comparação à área disponível para mapear amostras originalmente próximas. Em termos práticos, para se modelar pequenas distâncias corretamente no mapeamento, muitas amostras originalmente distantes tem de ser mapeadas em posições muito distantes na nova representação.

Os autores (MAATEN; HINTON, 2008) sugerem que uma maneira de aliviar esse problema é obtida pela utilização de distribuições de caudas mais pesadas. Assim a distribuição t-Student é escolhida. A função de custo no t-SNE difere da do SNE utilizando

uma forma simetrizada com cálculo simplificado de gradiente (COOK et al., 2007) e utilizando uma distribuição t-Student no lugar da Gaussiana para calcular similaridade entre duas amostras na nova representação. O Algoritmo 5 sumariza o método t-SNE cujos parâmetros são: conjunto de amostras X, quantidade de iterações T, perplexidade $Perp$, taxa de aprendizado η, momento $\alpha(t)$.

Algoritmo 5 t-distributed Stochastic Neighboring Embedding

0: **function** T-SNE($X, Perp, T, \eta, \alpha(t)$)
0: Calcule as probabilidades $p_{j|i}$ e $p_{i|j}$ utilizando a Equação (119).
0: Defina o valor de p_{ij} como

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{125}$$

0: Amostre a solução inicial $\mathcal{Y}^{(0)} = \{\vec{y}_1, \vec{y}_2, ..., \vec{y}_n\}$ de $\mathcal{N}(0, 10^{-4}I)$.
0: **for** t = 1 to T **do**
0: Calcule as afinidades no espaço-alvo q_{ij}

$$q_{ij} = \frac{\left(1 + \|\vec{y}_i - \vec{y}_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|\vec{y}_k - \vec{y}_l\|^2\right)^{-1}} = \frac{w_{ij}^{-1}}{\sum_{k \neq l} w_{kl}^{-1}} = \frac{w_{ij}^{-1}}{Z} \tag{126}$$

0: Calcule o gradiente utilizando a equação

$$\begin{aligned} \frac{\partial C}{\partial \vec{y}_i} &= 4 \sum_{j=1}^{n} (p_{ij} - q_{ij})\, w_{ij}^{-1}\, (\vec{y}_i - \vec{y}_j) \\ &= 4 \sum_{j=1}^{n} (p_{ij} - q_{ij}) \left(1 + \|\vec{y}_i - \vec{y}_j\|^2\right)^{-1} (\vec{y}_i - \vec{y}_j) \end{aligned} \tag{127}$$

0: Atualize as coordenadas utilizando a Equação (124).
0: **end for**
0: **return** $\mathcal{Y}^{(T)}$
0: **end function**=0

9.2.5 Uniform Manifold Approximation and Projection

O algoritmo UMAP é uma modificação do t-SNE que utiliza o LE para inicializar o processo iterativo. O UMAP se baseia na assunção de que existe uma medida de distância que faz com que as amostras se distribuam de forma aproximadamente uniforme em uma variedade localmente conexa. Para assegurar a validade dessa suposição, uma medida que aproxima os vizinhos mais próximos de x_i é escolhida, criando assim uma noção de distância para cada x_i que deve ser unificada à uma estrutura global consistente. Para assegurar essa unificação, esses espaços métricos devem ser convertidos em conjuntos simpliciais fuzzy, computacionalmente implementados através de um grafo ponderado (MCINNES; HEALY; MELVILLE, 2020).

Seja $X = \{x_1, x_2, \ldots, x_n\}$ o conjunto de amostras com métrica (i.e. medida de similaridade) $d : X \times X \to R_{\geq 0}$. Dado o parâmetro k, para cada x_i calcula-se um conjunto $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$ dos k vizinhos mais próximos de x_i sob a métrica d. O UMAP utiliza um algoritmo de procura descendente dos vizinhos mais próximos (DONG; MOSES; LI, 2011). Para cada x_i, ρ_i e σ_i definidos. Seja:

$$\rho_i = \min \left(d\left(x_i, x_{i_j}\right) \mid 1 \leq j \leq k, d\left(x_i, x_{i_j}\right) > 0 \right), \tag{128}$$

e σ_i um valor tal que:

$$\sum_{j=1}^{n} \exp \left(\frac{-\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i} \right) = \log_2 (k) . \tag{129}$$

A escolha de ρ_i garante que x_i está conectado a ao menos uma outra amostra com aresta de peso 1, o que equivale ao conjunto simplicial fuzzy localmente conexo em x_i. A escolha de σ_i corresponde a um fator de normalização suavizante, que define a métrica Riemanniana local na amostra x_i.

Dessa forma define-se o grafo ponderado $\bar{G} = (V, E, w)$, de vértices V. Portanto o conjunto $E = \left\{ \left(x_i, x_{i_j}\right) \mid 1 \leq j \leq k, 1 \leq i \leq N \right\}$ de arestas direcionadas é formado, e a função de peso w é definida como:

$$w\left(\left(x_i, x_{i_j}\right)\right) = \exp \left(\frac{-\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i} \right) . \tag{130}$$

Dada uma amostra x_i, existe um grafo induzido de x_i de arestas conectadas a x_i. O peso de uma aresta pode ser considerado como a probabilidade de que essa aresta exista. Dado esse conjunto de grafos locais deseja-se representá-los por um único grafo direcionado e portanto um método para combinar tais grafos em uma representação topológica unificada é necessário. Seja A a matriz de adjacência de $\bar{G}$, considerando a matriz simétrica:

$$B = A + A^{\top} - A \circ A^{\top}, \tag{131}$$

onde $\circ$ é o produto Hadamard. Se o valor de A_{ij} é interpretado como a probabilidade de uma aresta entre x_i e x_j existir, então B_{ij} é a probabilidade de que ao menos uma aresta (de x_i a x_j e de x_j a x_i) exista. O grafo G no UMAP é portanto um grafo não direcionado, com matriz de adjacência B.

Na prática, o UMAP usa um grafo de forças atrativas aplicadas ao longo das arestas e forças repulsivas entre os vértices. A obtenção de tal grafo define um problema de otimização não-convexa e é obtido de maneira iterativa. A convergência para um mínimo local é garantido pela redução lenta das forças de maneira similar a abordagem na técnica *simulated annealing* (MCINNES; HEALY; MELVILLE, 2020). No UMAP, a força de atração entre os vértices i e j de coordenadas $\vec{y}_i$ e $\vec{y}_j$ respectivamente, é determinada por:

$$\frac{-2ab \left\| \vec{y}_i - \vec{y}_j \right\|_2^{2(b-1)}}{1 + \left\| \vec{y}_i - \vec{y}_j \right\|_2^2} w\left((x_i, x_i)\right) \left(\vec{y}_i - \vec{y}_j\right) , \tag{132}$$

onde a e b são parâmetros. As forças repulsivas são calculadas por amostragem devido a restrições computacionais. Assim, quando uma força atrativa é aplicada em uma aresta, o vértice dessa aresta é repelido por uma amostragem dos outros vértices. A força repulsiva é dada por:

$$\frac{2b}{\left(\epsilon + \left\|\vec{y}_i - \vec{y}_j\right\|_2^2\right)\left(1 + a\left\|\vec{y}_i - \vec{y}_j\right\|_2^{2b}\right)}\left(1 - w\left(\left(x_i, x_j\right)\right)\right)\left(\vec{y}_i - \vec{y}_j\right), \tag{133}$$

onde ϵ é um valor pequeno que evita divisão por zero.

As forças derivam do gradiente que minimiza a entropia cruzada das arestas entre o grafo ponderado G e seu equivalente H construído das amostras $\{\vec{y}_i\}_{i=1\ldots n}$, isto é, deseja-se posicionar as amostras y_i de uma maneira que o grafo ponderado induzido por elas se aproxime de G o tanto quanto possível. A diferença entre os grafos ponderados é medida pela entropia cruzada total sobre todas as probabilidades de existência das arestas. Dado que o grafo ponderado G captura a topologia da dispersão original, o grafo equivalente H das amostras $\{\vec{y}_i\}_{1\ldots n}$ define uma aproximação de tal topologia o tanto quanto a otimização permitir, provendo assim uma boa representação da topologia geral da dispersão de entrada.

Referências

ABOU-MOUSTAFA, K. T.; FERRIE, F. P. A note on metric properties for some divergence measures:the gaussian case. **Journal of Machine Learning Research**, v. 25, p. 1–15, 2012.

AMARI, S. **Differential-geometrical methods in statistics (Lecture notes in statistics)**. [S.l.]: Springer-Verlag, 1985.

ARWINI, K. A.; DODSON, C. T. J. **Information Geometry: Near Randomness and Near Independence**. [S.l.]: Springer, 2008.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: DIETTERICH, T. G.; BECKER, S.; GHAHRAMANI, Z. (Ed.). **Advances in Neural Information Processing Systems 14**. [S.l.]: MIT Press, 2002. p. 585–591.

_____. Laplacian eigenmaps for dimensionality reduction and data representation. **Neural Computation**, MIT Press, v. 15, n. 6, p. 1373–1396, jun. 2003.

BHATTACHARYYA, A. On a measure of divergence between two statistical populations defined by their probability distributions. **Bulletin of the Calcutta Mathematical Society.**, v. 35, p. 99–109, 1943.

CALIńSKI, T.; JA, H. A dendrite method for cluster analysis. **Communications in Statistics - Theory and Methods**, v. 3, p. 1–27, 01 1974.

CANDèS, E. J. et al. Robust principal component analysis? **Journal of the ACM**, Association for Computing Machinery, New York, NY, USA, v. 58, n. 3, jun. 2011.

CANONNE, C. L. **A short note on an inequality between KL and TV**. 2022.

CASELLA, G.; BERGER, R. L. **Statistical inference**. [S.l.]: Cengage Learning, 2001.

CHUNG, F. R. K. (Ed.). **Spectral Graph Theory**. [S.l.]: American Mathematical Society, 1997.

CHUNG, F. R. K.; GRIGOR'YAN, A.; YAU, S.-T. Higher eigenvalues and isoperimetric inequalities on riemannian manifolds and graphs. **Communications in Analysis and Geometry**, v. 8, p. 969–1026, 2000. Disponível em: <https://api.semanticscholar.org/CorpusID:639010>.

COOK, J. A. et al. Visualizing similarity data with a mixture of maps. In: **Proceedings of the 11 th International Conference on Artificial Intelligence and Statistics**. [S.l.: s.n.], 2007. v. 2, p. 67–74.

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

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-1, n. 2, p. 224–227, 1979.

DEBORAH, L. J.; BASKARAN, R.; KANNAN, A. A survey on internal validity measure for cluster validation. **International Journal of Computer Science & Engineering Survey**, v. 1, p. 85–102, 2010. Disponível em: <https://api.semanticscholar.org/CorpusID:15068419>.

DIJKSTRA, E. W. A note on two problems in connexion with graphs. **Numerische Mathematik**, v. 1, n. 1, p. 269–271, dez. 1959.

DONG, W.; MOSES, C.; LI, K. Efficient k-nearest neighbor graph construction for generic similarity measures. In: **Proceedings of the 20th International Conference on World Wide Web**. New York, NY, USA: Association for Computing Machinery, 2011. (WWW '11), p. 577–586. ISBN 9781450306324. Disponível em: <https://doi.org/10.1145/1963405.1963487>.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 2. ed. [S.l.]: Wiley-Interscience, 2000. 688 pages p.

FISHER, R. A. On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society of London, A**, v. 222, p. 309–368, 1922.

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

GIL, M.; ALAJAJI, F.; LINDER, T. Rényi divergence measures for commonly used univariate continuous distributions. **Information Sciences**, v. 249, p. 124 – 131, 2013.

GUPTA, P.; SEHGAL, N. K.; ACKEN, J. M. Practical aspects in machine learning. In: _____. **Introduction to Machine Learning with Security: Theory and Practice Using Python in the Cloud**. Cham: Springer International Publishing, 2025. p. 281–330. ISBN 978-3-031-59170-9. Disponível em: <https://doi.org/10.1007/978-3-031-59170-9_9>.

HAVRDA, J.; CHARVAT, F. Quantification method of classification processes. **Kiberbetika Cislo**, v. 1, n. 3, p. 30–34, 1967.

HINTON, G. E.; ROWEIS, S. T. Stochastic neighbor embedding. In: BECKER, S.; THRUN, S.; OBERMAYER, K. (Ed.). **Advances in Neural Information Processing Systems 15**. [S.l.]: MIT Press, 2003. p. 857–864.

HOANG, H. G. et al. The cauchy-schwarz divergence for poisson point processes. **IEEE Trans. on Information Theory**, v. 61, n. 8, p. 4475–4485, 2015.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. Aberdeen, UK: Springer, 2002. 487 p.

KULLBACK, S.; LEIBLER, R. On information and sufficiency. **Annals of Mathematical Statistics**, v. 22, n. 1, p. 79–86, 1951.

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

LEVADA, A. L. Parametric pca for unsupervised metric learning. **Pattern Recognition Letters**, Elsevier, v. 135, p. 425–430, 2020.

LEVADA, A. L. M. Pca-kl: a parametric dimensionality reduction approach for unsupervised metric learning. **Advances in Data Analysis and Classification**, v. 15, n. 4, p. 829–868, 2021. Disponível em: <https://EconPapers.repec.org/RePEc:spr:advdac:v:15:y:2021:i:4:d:10.1007_s11634-020-00434-3>.

LI, D.; TIAN, Y. Survey and experimental study on metric learning methods. **Neural Networks**, v. 105, p. 447–462, 2018. ISSN 0893-6080.

LIU, Y. et al. Understanding of internal clustering validation measures. In: **2010 IEEE International Conference on Data Mining**. [S.l.: s.n.], 2010. p. 911–916.

MAATEN, L. van der; HINTON, G. Visualizing high-dimensional data using t-sne. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008.

MARKOPOULOS, P. P. et al. Efficient l1-norm principal-component analysis via bit flipping. **IEEE Transactions on Signal Processing**, v. 65, n. 16, p. 4252–4264, 2017.

MCINNES, L.; HEALY, J.; MELVILLE, J. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2020.

NAKAO, E. K.; LEVADA, A. L. M. Entropic principal component analysis using cauchy–schwarz divergence. **Knowledge and Information Systems**, v. 65, n. 12, p. 5375–5385, Dec 2023. ISSN 0219-3116. Disponível em: <https://doi.org/10.1007/s10115-023-01940-4>.

_____. Information theory divergences in principal component analysis. **Pattern Analysis and Applications**, v. 27, n. 1, p. 19, Feb 2024. ISSN 1433-755X. Disponível em: <https://doi.org/10.1007/s10044-024-01215-w>.

NIELSEN, F. The many faces of information geometry. **Notices of the American Mathematical Society**, v. 69, p. 36–45, 01 2022.

NIELSEN, F.; NOCK, R. A closed-form expression for the sharma–mittal entropy of exponential families. **Journal of Physics A: Mathematical and Theoretical**, IOP Publishing, v. 45, n. 3, p. 032003, dec 2011.

_____. On rényi and tsallis entropies and divergences for exponential families. 2011. Cite arxiv:1105.3259Comment: 7 pages. Disponível em: <http://arxiv.org/abs/1105.3259>.

Nielsen, F.; Sun, K. Guaranteed deterministic bounds on the total variation distance between univariate mixtures. In: **2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)**. [S.l.: s.n.], 2018. p. 1–6.

PARDO, L. **Statistical Inference Based on Divergence Measures**. [S.l.]: Chapman and Hall/CRC, 2006.

PARZEN, E. On estimation of a probability density function and mode. **The Annals of Mathematical Statistics.**, v. 33, n. 3, p. 1065–1076, 1962.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RAO, C. Information and accuracy attainable in estimation of statistical parameters. **Bulletin of the Calcutta Mathematical Society**, 1945.

ROSENBLATT, M. Remarks on some nonparametric estimates of a density function. **The Annals of Mathematical Statistics.**, v. 27, n. 3, p. 832–837, 1956.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Comp. and Appl. Math.**, v. 20, p. 53–65, 1987.

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

SCHöLKOPF, B.; SMOLA, A.; MüLLER, K. R. Kernel principal component analysis. In: **Advances in Kernel Methods – Support Vector Learning**. [S.l.]: MIT Press, 1999. p. 327–352.

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

SHARMA, S. **Applied Multivariate Techniques**. Wiley, 1995. ISBN 9780471310648. Disponível em: <https://books.google.com.br/books?id=6iURRAAACAAJ>.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 8, p. 888–905, 2000.

Spurek, P.; Palka, W. Clustering of gaussian distributions. In: **2016 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2016. p. 3346–3353.

SUáREZ, J. L.; GARCíA, S.; HERRERA, F. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. **Neurocomputing**, v. 425, p. 300–322, 2021. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231220312777>.

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

TSALLIS, C. Possible generalization of boltzmann-gibbs statistics. **Journal of Statistical Physics**, v. 52, p. 479–487, 1988.

van Erven, T.; Harremos, P. Rényi divergence and kullback-leibler divergence. **IEEE Transactions on Information Theory**, v. 60, n. 7, p. 3797–3820, 2014.

Verdu, S. Total variation distance and the distribution of relative information. In: **2014 Information Theory and Applications Workshop (ITA)**. [S.l.: s.n.], 2014. p. 1–3.

WANG, F.; SUN, J. Survey on distance metric learning and dimensionality reduction in data mining. **Data Min. Knowl. Discov.**, v. 29, n. 2, p. 534–564, mar. 2015. ISSN 1384-5810.

WEINBERGER, K. Q.; SAUL, L. K. Unsupervised learning of image manifolds by semidefinite programming. **International journal of computer vision**, Springer, v. 70, p. 77–90, 2006.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics Bulletin**, v. 1, n. 6, p. 80–83, 1945.

YI, S. et al. Joint sparse principal component analysis. **Pattern Recognition**, v. 61, p. 524 – 536, 2017.

YOUNG, T. Y.; CALVERT, T. W. **Classification, Estimation and Pattern Recognition**. [S.l.]: Elsevier, 1974.