# 2
# Diffusion Maps

Grandes volumes de dados em alta dimensão surgem naturalmente em diferentes campos do conhecimento. Aplicações em diversas áreas como aquisição de sinais, processamento de imagens, classificação e aprendizado estatístico são alguns exemplos. Em tais aplicações, o que se pretende é fazer, em alguma escala, uma caracterização desses conjuntos de dados, geralmente através de alguma representação significativa em baixa dimensão. Revelar a estrutura global − geométrica ou estatística − desses conjuntos está entre os objetivos dessa caracterização.

Ao longo do tempo, várias técnicas foram desenvolvidas buscando este propósito. Um dos primeiros trabalhos neste sentido foi o paper *Embedding Riemannian Manifolds by Their Heat Kernel*, de Bérard et al.[3], publicado em 1994. Amadurecendo uma idéia apresentada em um preprint de 1986, os autores propuseram um método de imersão de variedades Riemannianas fechadas no espaço $L^2$ −das séries de quadrado integrável −, usando para isso um *heat kernel* e as autofunções do laplaciano da variedade em questão para realizar o mapeamento.

Em anos mais recentes, contribuições de Roweis e Saul[17], Belkin e Niyogi[2], Meila e Shi[13] e de Coifman e Lafon[7] permitiram obter aproximações discretas do laplaciano de uma variedade a partir de um *kernel* adequado e associar a este estudo o conceito de *random walk* em um grafo, por meio de uma matriz estocástica, mesmo que a variedade seja não-linear. O método chamado *Diffusion Maps* é hoje o ponto culminante desses estudos.

## 2.1
## Conectividade em um conjunto de dados

Considere um conjunto finito $X = \{x_i\}_{i=1}^{n}$ de pontos num espaço de dimensão $p$, ou seja, $x_i \in \mathbb{R}^p$, $\forall i$. Considere ainda que $X$ esteja sobre uma variedade $\mho \subset \mathbb{R}^p$, sendo $\mho$ não necessariamente linear, como mostra a figura a seguir.

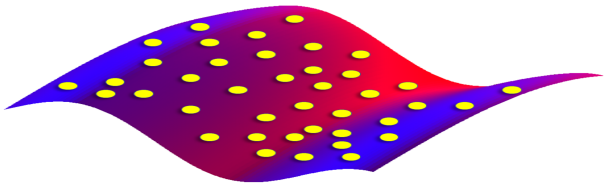

Figura 2.1: Conjunto finito sobre uma variedade.



É possível ver $X$ como um grafo $G = (V, E, w)$ − onde cada $x_i$ é um nó de $G$ −, desde que se defina uma medida $w$ de conectividade entre os seus elementos. Esta medida de conectividade − ou similaridade − é chamada de *kernel*, e será representada por $\mathbf{K}$: $X \times X \to \mathbb{R}$. Se dois pontos $x_i$ e $x_j$ de $X$ estiverem conectados por $\mathbf{K}$ , isso será indicado por $w_{ij} = k(x_i, x_j) \neq 0$. Finalmente, considere que $\mathbf{K}$ seja tal que apresente as seguintes propriedades:

**P1.** Simetria : $k(x_i, x_j) = k(x_j, x_i)$
**P2.** Não-negatividade : $k(x_i, x_j) \geq 0$

Na literatura, é muito comum o emprego do chamado *kernel* gaussiano − ou *heat kernel* − definido por

$$k(x_i, x_j) = e^{-\|x_i - x_j\|^2/\varepsilon} \tag{2-1}$$

que além de apresentar as propriedades **P1** e **P2**, tem $k(x_i, x_i) > 0$ , $\forall i \in 1, 2, \ldots, n$. Note que esta propriedade do *kernel* gaussiano faz com que o grafo $G$ associado ao conjunto $X$ tenha 1 laço em cada um de seus nós. Ao longo desta dissertação, será empregado apenas o *kernel* gaussiano.

O parâmetro $\varepsilon$ , que aqui será chamado de *largura*, controla a ve-

locidade de decrescimento exponencial do *kernel*. Desta forma, exerce papel fundamental na determinação do que pode ser interpretado como tamanho da vizinhança de um ponto $x_i$, segundo a idéia de conectividade. Falando de uma maneira não-formal, um $\varepsilon$ "grande" engrossa o *kernel*, fazendo com que as arestas entre $x_i$ e seus nós adjacentes $x_j$ tenham pesos maiores do que as arestas determinadas por um *kernel* cujo $\varepsilon$ é "pequeno". A próxima figura mostra representações do *heat kernel* para diferentes valores de $\varepsilon$.

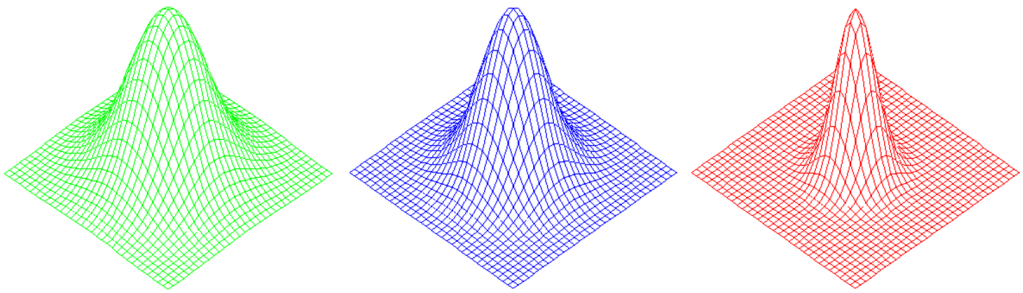

Figura 2.2: *Kernel* gaussiano com diferentes *larguras*. Da esquerda para a direita, os valores de $\varepsilon$ são decrescentes.



## 2.2
## Relação entre kernel e cadeia de Markov

Todas as avaliações do *kernel* $\mathbf{K}$: $X \times X \to \mathbb{R}$ podem ser armazenadas na matriz simétrica e não-negativa $K = [k_{ij}]_{n \times n}$ , de tal modo que $k_{ij} = k(x_i, x_j)$.

Além disso, é possível submeter $K$ a um processo de normalização que faz surgir uma matriz de transição de probabilidades $P$, onde $P$ está associada simultaneamente a uma cadeia de Markov sobre o espaço de estados $X$ e a um *random walk* sobre o grafo $G$.

Para isto, considere uma matriz diagonal $D_{n \times n}$ tal que $D_{ii} = \sum_j k_{ij}$ , ou seja, cada entrada $i$ de $D$ é a soma dos elementos da *i-ésima* linha de $K$. Deste modo, $P$ é obtida fazendo-se $D^{-1}K$.

Vale lembrar que, sendo $\mathbf{K}$ gaussiano, tem-se $k(x_i, x_i) > 0$ e, consequentemente, $p_{ii} = p(x_i, x_i) > 0$, que é condição suficiente para que a cadeia de Markov representada por $P$ seja aperiódica. Mais ainda: sendo $\varepsilon$ suficientemente grande, a cadeia em questão é irredutível, pois está associada

a um grafo conexo. Isto decorre do fato de $\varepsilon$ controlar o decrescimento do *kernel*, como explicado na seção anterior. Um $\varepsilon$ "pequeno demais" produz, após a normalização, probabilidades $p_{ij}$ também muito pequenas, por vezes tão desprezíveis numericamente que desconexões entre os nós são produzidas, podendo tornar a cadeia redutível. Apesar disso, a literatura atual não fornece meios de se conhecer, *a priori*, qual o melhor intervalo de valores de $\varepsilon$ para um determinado conjunto $X$, e essa escolha torna-se tarefa experimental.

De qualquer modo, supondo-se que se tem um $\varepsilon$ adequado, o que pode ser dito é que o *kernel* permite inferir sobre estruturas geométricas locais de $X$. Estruturas em outras escalas, até mesmo a estrutura global do conjunto, são resultantes da integração de suas informações locais, e podem ser obtidas pelas potências $P^t$ da matriz de transição. Coifman e Lafon[7] ilustram esta situação através do seguinte exemplo: um conjunto de pontos $X$ é gerado no plano, e partindo-se de um *kernel* gaussiano para estabelecer uma relação inicial de conectividade entre os pontos, é construída a matriz $P$ de probabilidades de transição correspondente. À medida que são feitas as potências $P^t$, a conectividade inicial é alterada. Isso faz com que as três nuvens de pontos − que podem ser chamadas de *clusters* − se misturem ao longo do tempo. Isto é, a geometria local sofre um processo de integração, até se admitir que o conjunto $X$ seja formado por um único agrupamento de pontos. Essa descrição é ilustrada pela figura a seguir, extraída de [7]. A disposição das cores (que não fazem parte dos dados de $X$) sugere tão-somente a característica multi-escala da estrutura geométrica do conjunto de pontos, revelada pelas potências $P^8$ , $P^{64}$ e $P^{1024}$.

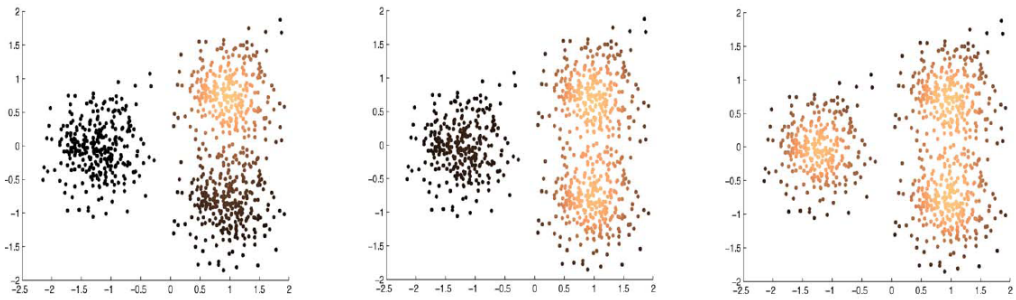

Figura 2.3: Estruturas multi-escala de um conjunto, reveladas pelas potências $P^t$ da matriz de transição. Da esquerda para a direita, tem-se $t = 8$, $t = 64$ e $t = 1024$, respectivamente. Imagens extraídas de Coifman e Lafon[7].

## 2.3
## Distância de difusão

A partir do que foi apresentado sobre os efeitos das potências de $P$, é correto admitir que a similaridade entre os pontos de $X$ pode ser quantificada de acordo com a evolução das probabilidades de transição do *random walk* correspondente. Neste sentido, Coifman e Lafon[7] definem, para um certo tempo $t$, uma distância de difusão entre os pontos $x_i$ e $x_j$, denotada por $\mathbf{D}_t(x_i, x_j)$, calculada de acordo com

$$\mathbf{D}_t^2(x_i, x_j) = \|p_t(x_i, \bullet) - p_t(x_j, \bullet)\|_\xi^2 \tag{2-2}$$

Nesta igualdade, $\bullet$ indica todos os pontos de $X$, $p_t(x_i, \bullet)$ é a probabilidade de transição do ponto $x_i$ para o ponto $\bullet$ e $\xi$ é um fator de ponderação adequado. Mais adiante, será visto como $\xi$ pode ser definido.

É fácil notar que a expressão $\|p_t(x_i, \bullet) - p_t(x_j, \bullet)\|$ indica, no tempo $t$, a diferença entre as linhas $i$ e $j$ da matriz de transição $P^t$. Ponderada pelo fator $\xi$, a norma desse vetor-diferença conduz à distância de difusão definida.

## 2.4
## Representação em coordenadas de difusão

A distância de difusão definida na seção anterior é capaz de estabelecer todas as relações de distância − ponderadas por $\xi$ − entre os pontos de $X$. Assim sendo, se $n(X) = n$, obter $\mathbf{D}_t$ requer o cálculo − para cada tempo $t$ − da norma de $\frac{n!}{2(n-2)!}$ vetores de dimensão $n$ − que são as diferenças entre as linhas da matriz $P^t$. Mas é possível mostrar que, mesmo sem calcular $\mathbf{D}_t$ diretamente, pode-se obter uma representação geométrica equivalente para as distâncias de difusão.

Para isto, o que o *Diffusion Maps* faz é definir, por meio dos autovetores de $P$, um sistema de coordenadas de difusão no espaço euclidiano $\mathbb{R}^s$ para onde os elementos de $X$ serão mapeados , ao mesmo tempo em que a estrutura intrínseca do conjunto de dados é preservada. Ou seja, é construída uma aplicação

$$\Psi_t : X \to \mathbb{R}^s \tag{2-3}$$

que, para um determinado instante $t$, "mergulha" os dados de $X$ para o espaço $\mathbb{R}^s$ com a norma euclidiana usual.

A aplicação $\Psi_t$ é obtida pelos autovalores e autovetores da matriz $P$. Pelos teoremas 1.3.1.3 e 1.3.1.4, pode-se garantir que os autovalores à direita de $P_{n\times n}$ formam a sequência

$$1 = \lambda_0 > |\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_{n-1}| \tag{2-4}$$

Associem-se a essa sequência os autovetores à direita $\psi_0, \psi_1, ..., \psi_{n-1}$. Deste modo, dado um ponto $x_i \in X$, a sua representação no espaço $\mathbb{R}^s$ em um certo instante $t$, através das coordenadas de difusão, se dará da seguinte forma:

$$\Psi_t(x_i) = \begin{bmatrix} \lambda_1^t\psi_1(x_i) \\ \lambda_2^t\psi_2(x_i) \\ \vdots \\ \lambda_s^t\psi_s(x_i) \end{bmatrix} \tag{2-5}$$

onde $\psi_q(x_i)$ significa a *i-ésima* componente do *q-ésimo* autovetor de $P$.

O autovalor $\lambda_0 = 1$ e seu correspondente autovetor $\psi_0$ em geral não são usados na construção de $\Psi$, pelo fato de todas as componentes de $\psi_0$ serem iguais.

Com o que se tem até aqui, é possível estruturar uma versão básica de um algoritmo para o método *Diffusion Maps*:

**Algoritmo 1 - Diffusion Maps em versão básica**

**1) Entre com** $\{x_1, x_2, ..., x_n\}$**,** $\varepsilon$ **,** $s$ **,** $t$

**2) Construa** $K_{n\times n}$ **tal que** $k(x_i, x_j) = e^{-\frac{\left\|x_i - x_j\right\|^2}{\varepsilon}}$**,** $\forall i, j \in \{1, 2, ..., n\}$

**3) Obtenha a diagonal** $D_{n\times n}$**, tal que** $D_{ii} = \sum_{j=1}^{n} k(x_i, x_j)$

**4) Obtenha a matriz de transição** $P = D^{-1}K$

**5) Calcule os autovalores** $\lambda_1, ..., \lambda_s$ **e autovetores** $\psi_1, ..., \psi_s$ **de** $P$

**6) Construa o mapeamento** $\Psi_t(x_i) = \begin{bmatrix} \lambda_1^t\psi_1(x_i) \\ \lambda_2^t\psi_2(x_i) \\ \vdots \\ \lambda_s^t\psi_s(x_i) \end{bmatrix}$**,** $\forall i \in \{1, 2, ..., n\}$

A aplicação $\Psi_t$ é empregada pelo fato de se poder mostrar que as relações de distância entre os pontos mapeados $\Psi_t(x_1), \Psi_t(x_2), \ldots, \Psi_t(x_n)$, segundo a norma euclidiana em $\mathbb{R}^s$ , são equivalentes às relações de distância (de difusão) entre os pontos $x_1, x_2, \ldots, x_n$ no conjunto original $X$ e ponderadas por $\xi$. Ou seja,

$$\|\Psi_t(x_i) - \Psi_t(x_j)\| = \mathbf{D}_t(x_i, x_j) \tag{2-6}$$

A próxima seção esclarece como isso acontece.

## 2.5
## Equivalência entre distâncias

Para mostrar a equivalência entre as distâncias tratadas na seção anterior, serão observados, de início, alguns aspectos essencialmente matriciais.

Em primeiro lugar, deve-se notar que $P$ não é simétrica, em virtude da normalização $D^{-1}K$. Mas é fácil obter $P_S = D^{1/2}PD^{-1/2}$, onde $P_S$ é simétrica.

A auto-decomposição de $P_S$ é tal que $P_S = U\Lambda U^T$, e imediatamente se tem $P = (D^{-1/2}U)\Lambda(U^T D^{1/2})$. Ou seja, $P$ e $P_S$ têm os mesmos autovalores.

Fazendo $u$ ser autovetor de $P_S$ e $\phi^T$ e $\psi$, respectivamente, autovetores à esquerda e à direita de $P$, são válidas as relações $\psi_k = D^{-1/2}u_k$ e $\phi_k^T = u_k^T D^{1/2}$. A associação dessas igualdades à auto-decomposição de $P$ permite escrever a matriz $P$ na base de autovetores à esquerda, ou seja, $P = \sum_k \lambda_k \psi_k \phi_k^T$.

Considere agora, num certo instante $t$, a aplicação $\Psi_t : X \to \mathbb{R}^n$ − suponha a utilização integral do espectro − de dois pontos $x_i$ e $x_j$ de $X$, ou seja,

$$\Psi_t(x_i) = \begin{bmatrix} \lambda_0^t \psi_0(x_i) \\ \lambda_1^t \psi_1(x_i) \\ \vdots \\ \lambda_{n-1}^t \psi_{n-1}(x_i) \end{bmatrix} \tag{2-7}$$

e também

$$\Psi_t(x_j) = \begin{bmatrix} \lambda_0^t \psi_0(x_j) \\ \lambda_1^t \psi_1(x_j) \\ \vdots \\ \lambda_{n-1}^t \psi_{n-1}(x_j) \end{bmatrix} \tag{2-8}$$

A distância euclidiana entre os pontos $\Psi_t(x_i)$ e $\Psi_t(x_j)$ pode ser calculada a partir do produto interno do vetor diferença $\Psi_t(x_i) - \Psi_t(x_j)$ por ele mesmo, isto é,

$$\|\Psi_t(x_i) - \Psi_t(x_j)\|^2 =$$

$$= \left\langle \begin{bmatrix} \lambda_0^t(\psi_0(x_i) - \psi_0(x_j)) \\ \lambda_1^t(\psi_1(x_i) - \psi_1(x_j)) \\ \vdots \\ \lambda_{n-1}^t(\psi_{n-1}(x_i) - \psi_{n-1}(x_j)) \end{bmatrix}, \begin{bmatrix} \lambda_0^t(\psi_0(x_i) - \psi_0(x_j)) \\ \lambda_1^t(\psi_1(x_i) - \psi_1(x_j)) \\ \vdots \\ \lambda_{n-1}^t(\psi_{n-1}(x_i) - \psi_{n-1}(x_j)) \end{bmatrix} \right\rangle \tag{2-9}$$

E então:

$$\|\Psi_t(x_i) - \Psi_t(x_j)\|^2 = \sum_k \lambda_k^{2t}(\psi_k(x_i) - \psi_k(x_j))^2 \tag{2-10}$$

Para que a distância de difusão (ponderada por $\xi$) e a distância de mapeamento (pela norma euclidiana usual) sejam equivalentes, ou seja, para que valha a relação

$$\mathbf{D}_t^2(x_i, x_j) = \|p_t(x_i, \bullet) - p_t(x_j, \bullet)\|_\xi^2 = \|\Psi_t(x_i) - \Psi_t(x_j)\|^2 \tag{2-11}$$

basta mostrar que

$$\mathbf{D}_t^2(x_i, x_j) = \|p_t(x_i, \bullet) - p_t(x_j, \bullet)\|_\xi^2 = \sum_k \lambda_k^{2t}(\psi_k(x_i) - \psi_k(x_j))^2 \tag{2-12}$$

De fato, considerando a auto-decomposição de $P$, vem

$$\|p_t(x_i, \bullet) - p_t(x_j, \bullet)\|_\xi^2 = \left\|\sum_k \lambda_k^t \psi_k(x_i)\phi_k^T - \sum_k \lambda_k^t \psi_k(x_j)\phi_k^T\right\|_\xi^2 =$$

$$= \left\|\sum_k \lambda_k^t \phi_k^T(\psi_k(x_i) - \psi_k(x_j))\right\|_\xi^2 \tag{2-13}$$

Lembrando que $\phi_k^T = u_k^T D^{1/2}$, transforma-se a igualdade acima em

$$\mathbf{D}_t^2(x_i, x_j) = \left\| \sum_k \lambda_k^t u_k^t (\psi_k(x_i) - \psi_k(x_j)) D^{1/2} \right\|_\xi^2 \tag{2-14}$$

Por fim, desenvolvendo o produto interno e fazendo $\xi$ ser definido por $D^{-1}$ , escreve-se

$$\mathbf{D}_t^2(x_i, x_j) = (\sum_k \lambda_k^t u_k^t (\psi_k(x_i) - \psi_k(x_j))) D^{1/2} D^{-1} D^{1/2} (\sum_k \lambda_k^t u_k^t (\psi_k(x_i) - \psi_k(x_j))) \tag{2-15}$$

e assim

$$\mathbf{D}_t^2(x_i, x_j) = \sum_k \lambda_k^{2t} (\psi_k(x_i) - \psi_k(x_j))^2 \tag{2-16}$$

## 2.6 Exemplos

### Exemplo 1 - Preservação de estruturas geométricas

Considere um conjunto $X$ de 200 pontos gerados uniformemente no quadrado $[0, 1]^2$ e representado na figura a seguir. Aleatoriamente, os pontos sorteados são classificados segundo dois critérios diferentes: cor (verde ou vermelho) e forma (cruz ou círculo).

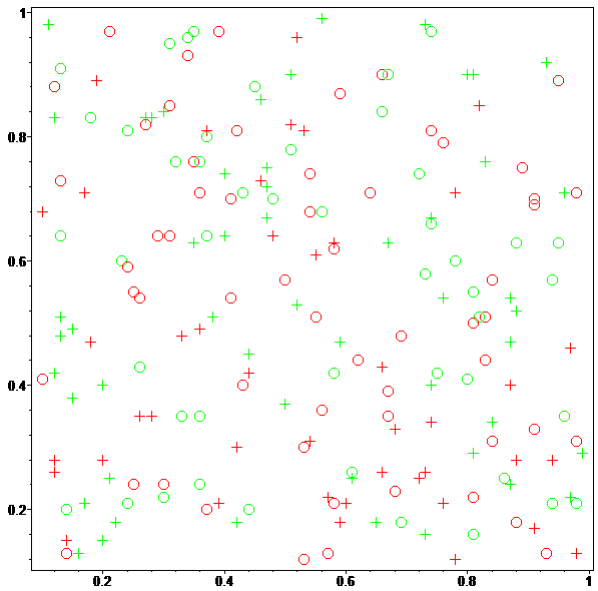


Figura 2.4: Conjunto aleatório de 200 pontos. Os eixos coordenados são **x** e **y**.

Primeiramente, o mapeamento destes pontos nas coordenadas de difusão será feito considerando-se apenas a posição. Assim, cada $x_i \in X$ é da forma $[\mathbf{x}_i, \mathbf{y}_i]$, com $i = 1, \ldots, 200$. A figura 2.5 mostra o resultado obtido através de um *kernel* gaussiano com parâmetro $\varepsilon = 0.8$.

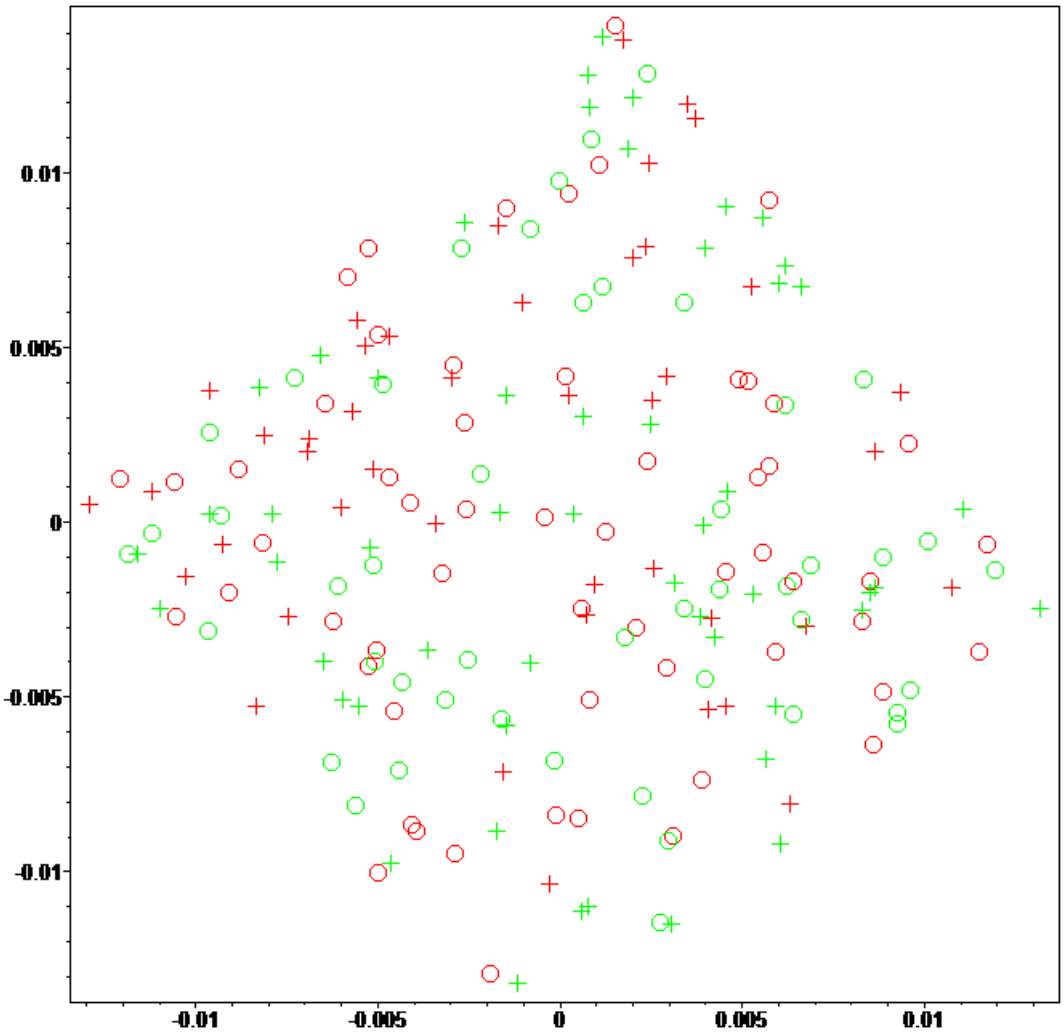


Figura 2.5: *Diffusion Maps* do conjunto $X$, considerando como dados apenas a posição dos pontos. Aqui, os eixos coordenados são os autovetores $\psi_1$ e $\psi_2$.

A aplicação do *Diffusion Maps* a este conjunto produz uma distribuição dos pontos, nas coordenadas de difusão, que preserva estruturas geométricas da distribuição original. Isto condiz com o que foi explicado sobre a equivalência entre distâncias na seção anterior. A figura 2.6 ilustra este fato.

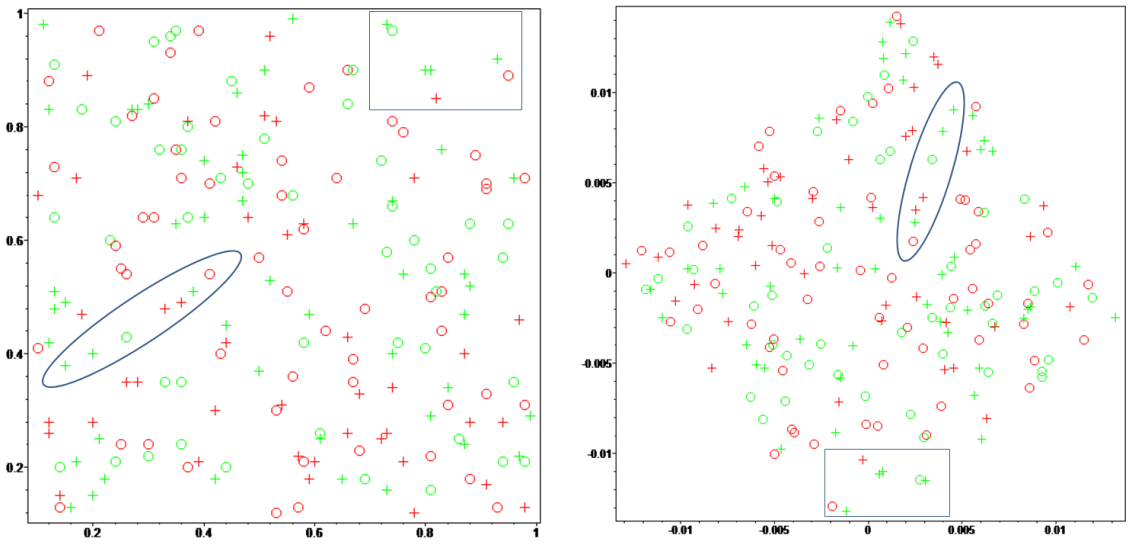


Figura 2.6: Estruturas geométricas do conjunto original preservadas nas coordenadas de difusão.

**Exemplo 2 - Capacidade de formação de *clusters***

Suponha agora que se queira agrupar − ou seja, *clusterizar* − os elementos do conjunto $X$ capturando a pré-classificação dada a eles de acordo com os critérios de cor e forma. Pode-se então fazer com que os dados de entrada assumam a forma $x_i = [\mathbf{x}_i, \mathbf{y}_i, \mathbf{cor}_i]$, para se conseguir uma separação dos elementos em dois grupos de cores diferentes; do mesmo modo, ao se usar $x_i = [\mathbf{x}_i, \mathbf{y}_i, \mathbf{forma}_i]$, pretende-se separar os elementos de $X$ em dois grupos de formas distintas; e para capturar a classificação quanto a cor e forma simultaneamente, pode-se empregar $x_i = [\mathbf{x}_i, \mathbf{y}_i, \mathbf{cor}_i, \mathbf{forma}_i]$.

As figuras 2.7, 2.8 e 2.9, mostradas a partir da próxima página, ilustram as *clusterizações* obtidas e representadas em duas e três dimensões. O parâmetro $\varepsilon$ utilizado é o mesmo do exemplo anterior.

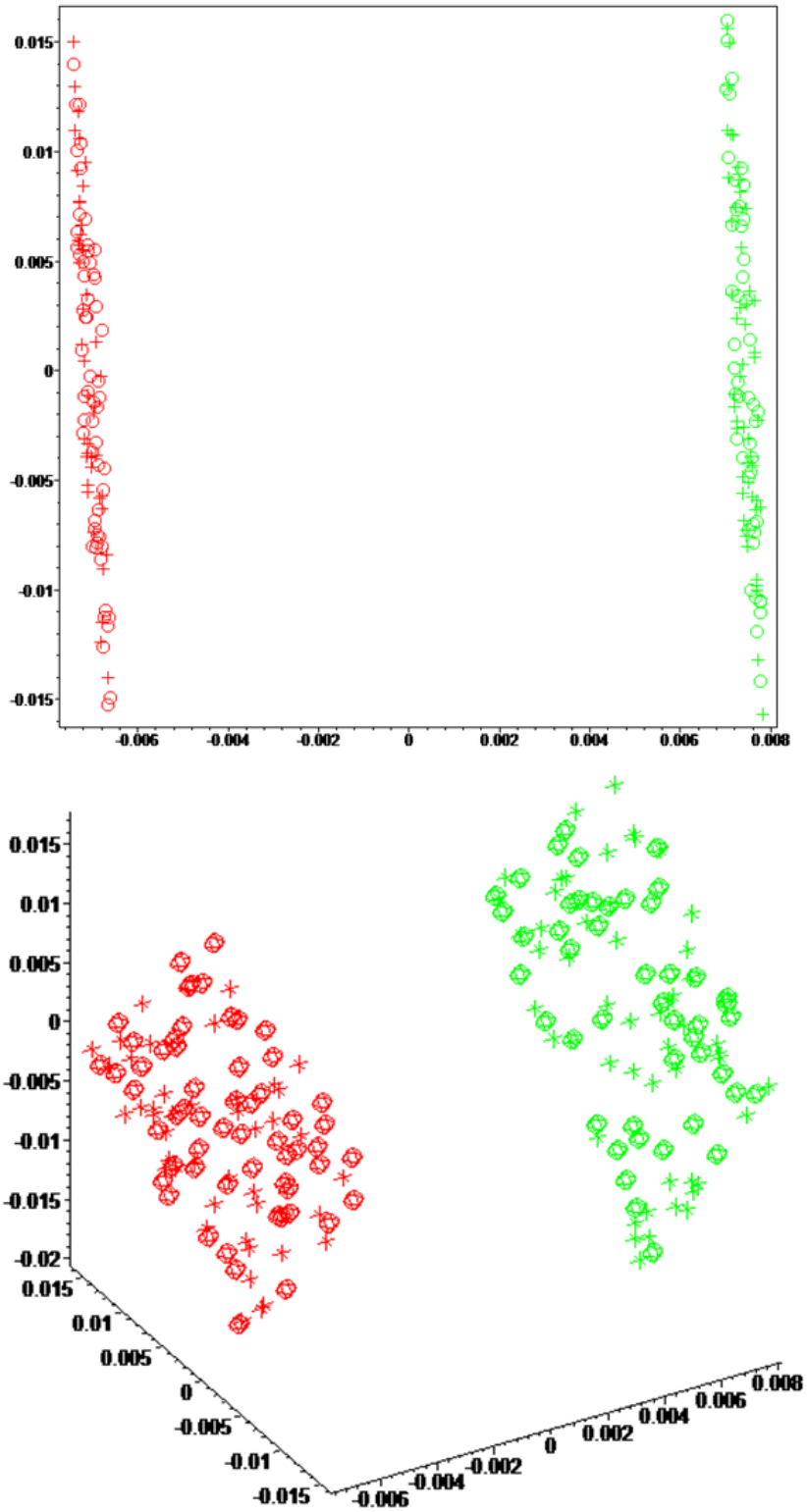


Figura 2.7: Clusterização segundo o critério de cor.

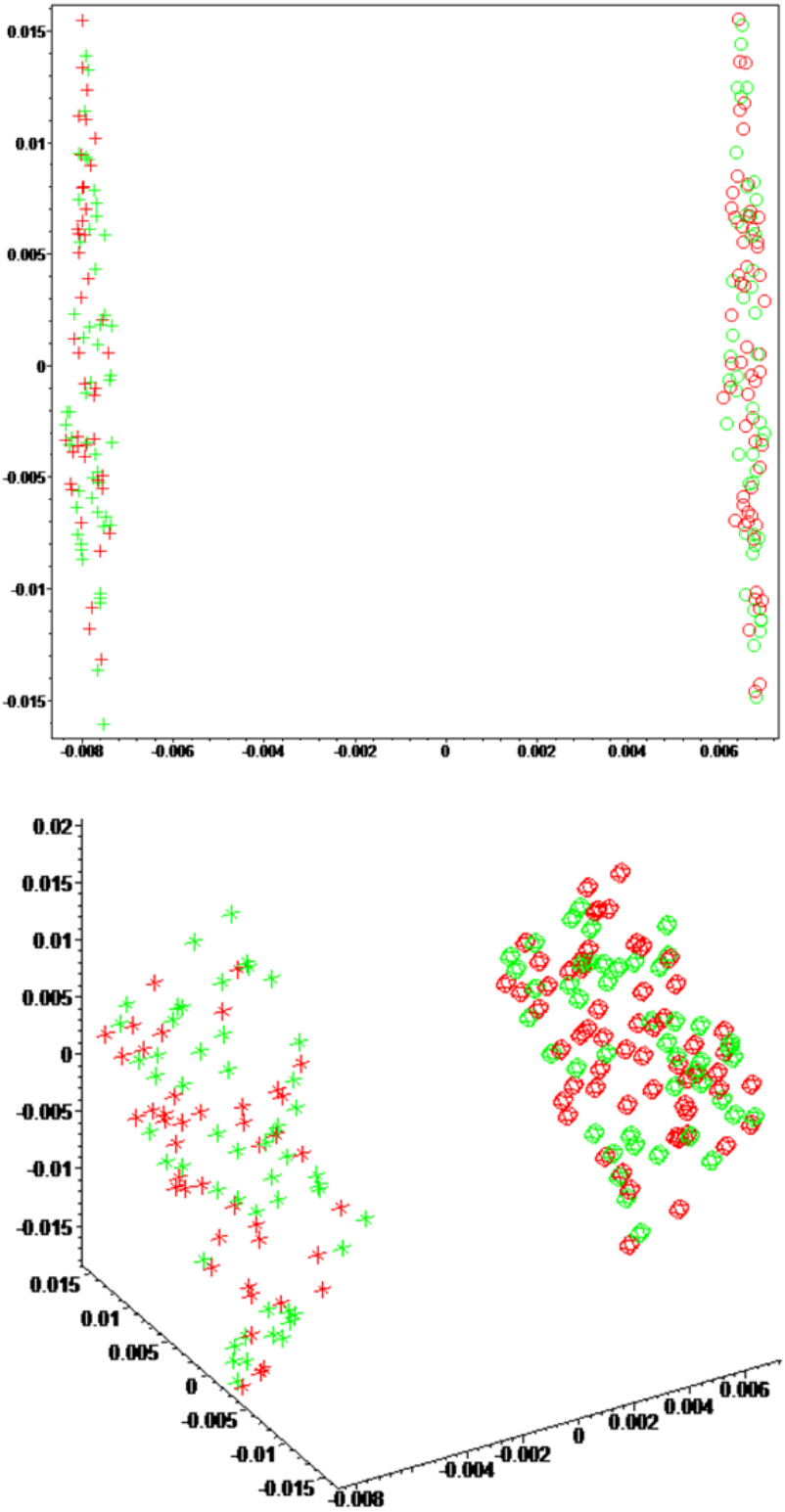


Figura 2.8: Clusterização segundo o critério de forma.

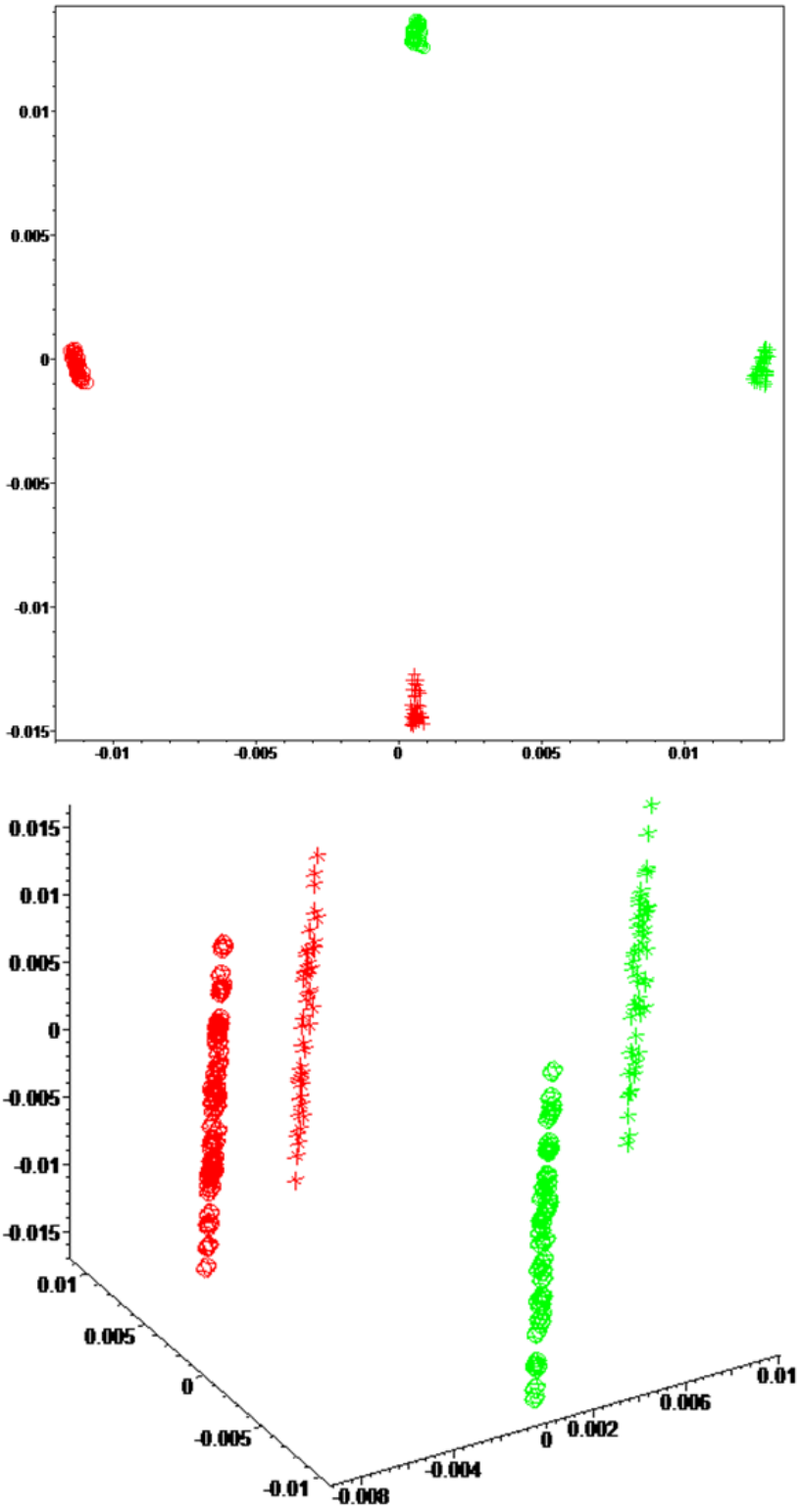


Figura 2.9: Clusterização segundo cor e forma.

**Exemplo 3 - Capacidade de organização**

Os conceitos de similaridade e distância de difusão permitem aplicar o *Diffusion Maps* na ordenação de uma sequência de movimentos, tal como os *frames* de um vídeo. Considerando-se que *frames* próximos entre si têm uma similaridade alta − e consequentemente pequena distância de difusão −, é possível recuperar a organização original da sequência de cenas através das coordenadas de difusão.

O exemplo a seguir trata de um vídeo em que um projétil atinge uma maçã. Uma permutação da sequência original de *frames* deste vídeo foi usada para se construir uma matriz de transição $27 \times 27$, onde cada um dos 27 elementos $x_i$ do conjunto $X$ de *frames* é de dimensão 18088 − cada *frame* tem 238 por 76 *pixels*. A próxima figura mostra os *frames* inicialmente desordenados.

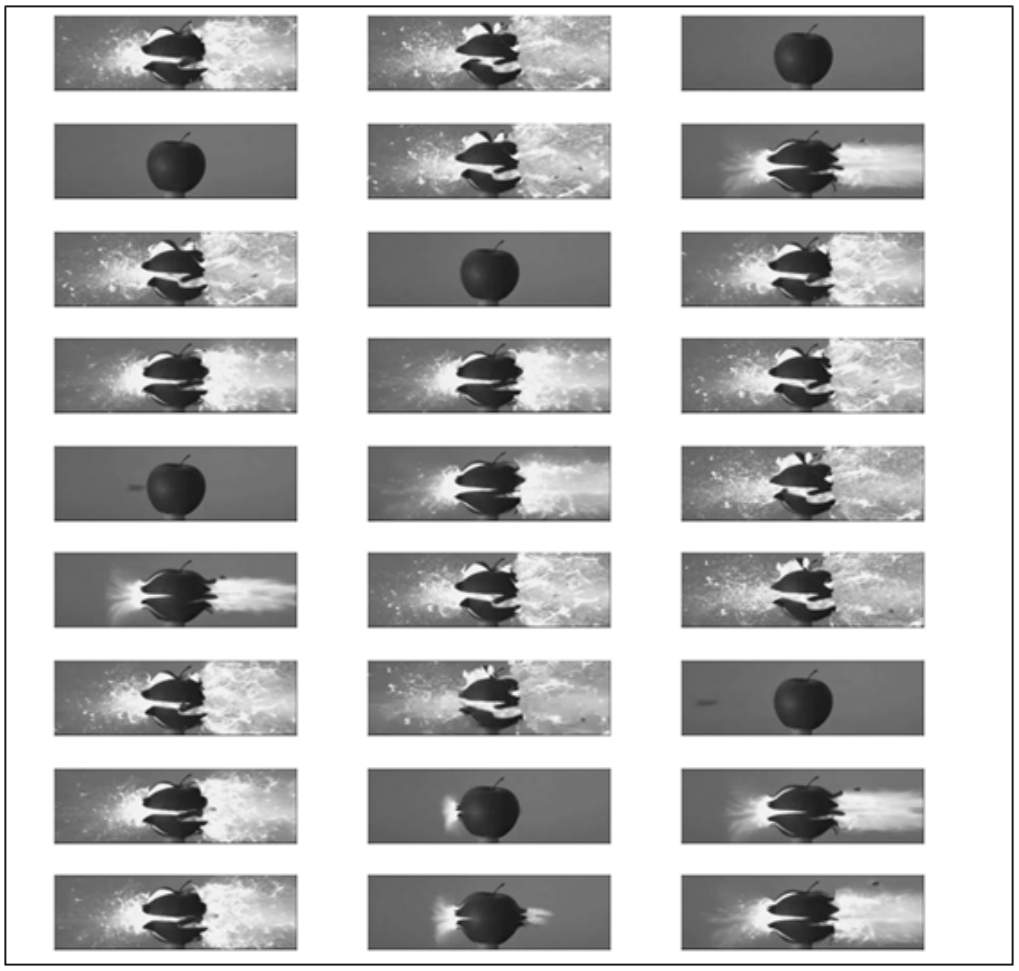

Figura 2.10: Sequência desordenada de 27 frames submetida à ordenação pelo *Diffusion Maps*. www.youtube.com

Na representação de $X$ em coordenadas de difusão, foram utilizados círculos de tamanhos diferentes, associados à posição dos *frames* na sequência, de modo que o primeiro *frame* fosse representado pelo menor círculo empregado, e o último *frame*, pelo maior dos círculos. Deve-se destacar aqui que isso é uma estratégia puramente gráfica: o tamanho dos círculos não são utilizados na construção do *kernel*.

As figuras a seguir mostram a representação das duas primeiras coordenadas de difusão do conjunto e a sequência recuperada de *frames*.

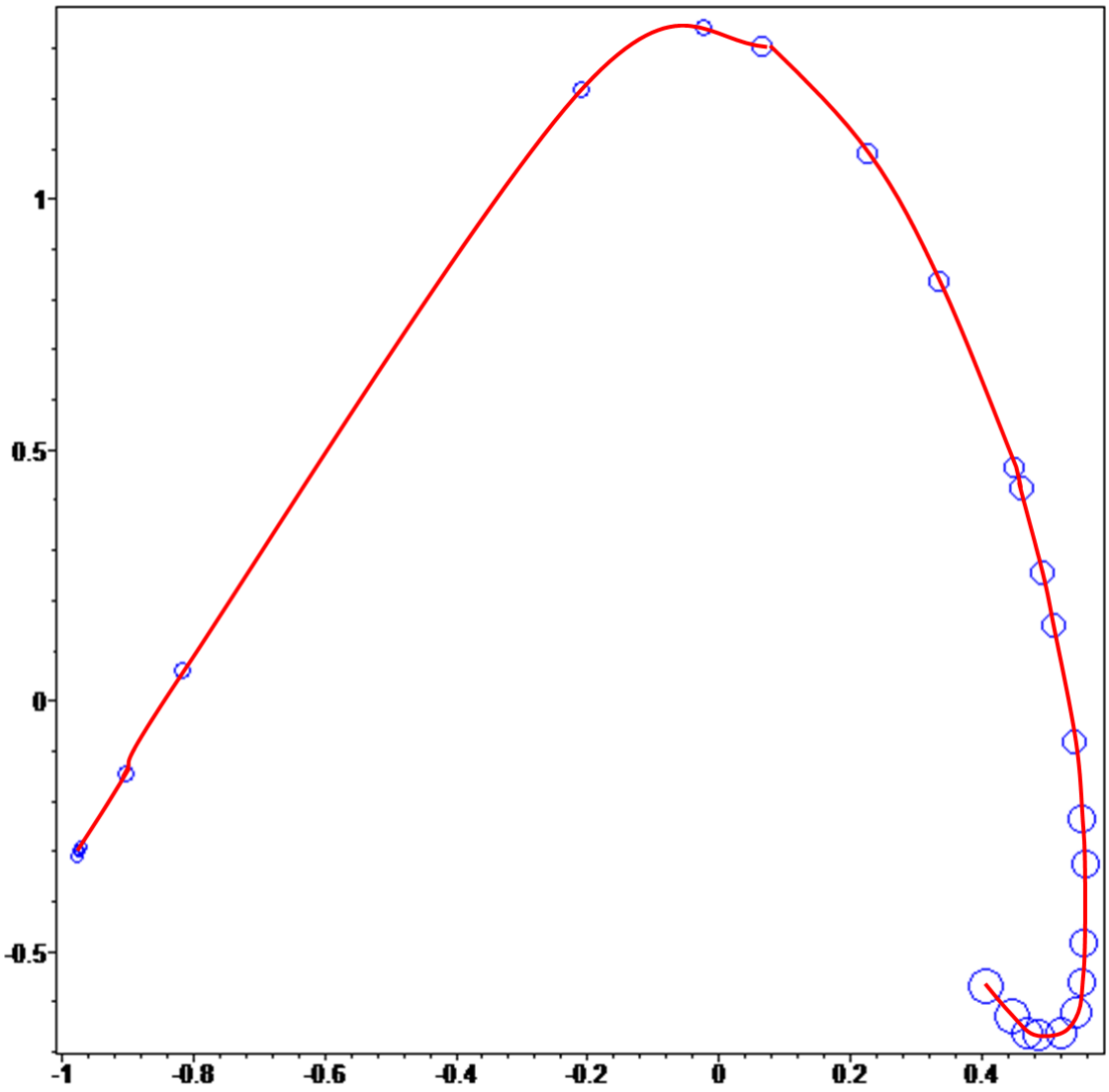


Figura 2.11: Coordenadas de difusão do conjunto de *frames*. A linha vermelha foi traçada apenas para se destacar a ordenação obtida para as cenas.

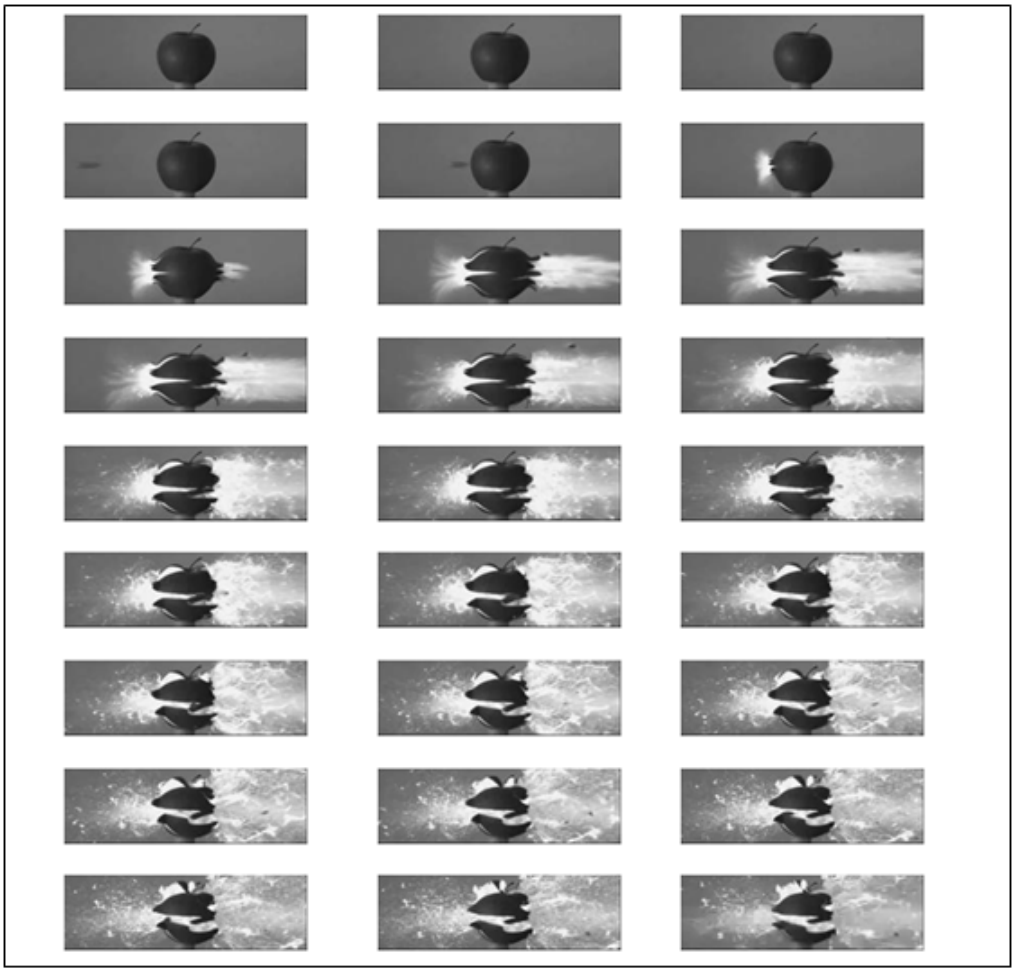

Figura 2.12: Recuperação integral da sequência original de *frames* do vídeo.

## 2.7
## O parâmetro de difusão $\varepsilon$

Na Seção 2.1, viu-se que o *kernel* gaussiano também recebe a denominação de *heat kernel*. Isso vem da sua relação com a equação do calor. Sabe-se que no espaço $\mathbb{R}^m$ a difusão de calor num meio cuja condutividade térmica é $\kappa$ pode ser descrita pela equação

$$\frac{\partial u(\mathbf{x}_1, ..., \mathbf{x}_m, t)}{\partial t} - \kappa \sum_{i=1}^{m} \frac{\partial^2 u(\mathbf{x}_1, ..., \mathbf{x}_m, t)}{\partial \mathbf{x}_i^2} = 0 \qquad (2\text{-}17)$$

Respeitando-se as condições inicial e de contorno do problema, e fazendo $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_m)$, a solução fundamental da equação acima pode ser expressa por

$$u(\mathbf{x}, t) = \frac{1}{(4\pi\kappa t)^{m/2}} e^{-\mathbf{x}.\mathbf{x}/4\kappa t} \tag{2-18}$$

A forma da solução fundamental revela o mesmo comportamento de decrescimento exponencial presente no *kernel* gaussiano. Considerando-se isso, é possível então dar um outro significado para o parâmetro $\varepsilon$ descrito anteriormente como largura do *kernel*. Pode-se ver $\varepsilon$ como o tamanho do passo de tempo do *random walk* entre os nós $x_i$ do grafo $G$ associado ao conjunto $X$ definido sobre uma variedade $\mho$.

Esta nova interpretação permite que se façam algumas considerações. Sendo $n = n(X)$ finito e $\varepsilon > 0$, as transições entre os estados $x_i$ ocorrem de acordo com um *random walk* discreto no espaço e no tempo. Fazendo-se $n \to \infty$ e mantendo-se $\varepsilon > 0$, passa-se a ter um *random walk* contínuo no espaço e discreto no tempo. Finalmente, para $n \to \infty$ e $\varepsilon \to 0$, tem-se a representação de um *random walk* contínuo no espaço e no tempo, ou seja, um processo de difusão. No estudo do método *Diffusion Maps*, o interesse teórico de se considerar um passeio aleatório a um parâmetro infinitesimal $\varepsilon$ reside no fato de se poder analisar a evolução das probabilidades de transição entre os estados como aproximações de classes de operadores de difusão. Isto será feito na próxima seção.

## 2.8 Famílias de difusão

No início da Seção 2.1, considerou-se um conjunto finito $X = \{x_i\}_{i=1}^n$ de pontos sobre uma variedade $\mho \subset \mathbb{R}^p$ como ponto de partida para a construção discreta do método *Diffusion Maps*. Mas a questão sobre como esses pontos estão distribuídos sobre $\mho$ não foi abordada, ou seja, nenhum tratamento sobre a densidade dos elementos de $X$ foi feito. O que se quer dizer aqui é que ter informação geométrica a respeito de $X$ ou de $\mho$ não significa ter, automaticamente, informação estatística sobre essa variedade. Ou ainda, sabendo-se que as coordenadas de difusão preservam estruturas geométricas, conforme ilustrado na Seção 2.6, Exemplo 1, é coerente perguntar se é possível fazer um mapeamento que aponte características estatísticas do conjunto em estudo. A resposta a esta pergunta é sim, e encontra-se na construção do que se chama de famílias de difusão, dependentes de um parâmetro que atribui maior ou menor importância à densidade da distribuição dos pontos sobre $\mho$.

A bem da verdade, deve-se dizer que os quatro primeiros passos do Algoritmo 1 apresentado na Seção 2.4 correspondem ao que a literatura chama de construção clássica do laplaciano normalizado de um grafo. Tais procedimentos são a base dos chamados métodos espectrais de clusterização, e construções semelhantes podem ser encontradas nos trabalhos de Shi e Malik[18] , Belkin e Niyogi[2] e Meila e Shi[13]. No entanto, nenhum destes aborda, em sua construção, o papel da densidade dos pontos amostrados que formam o conjunto $X$. Isto talvez seja o maior diferencial do *Diffusion Maps* em relação a outros métodos espectrais.

A convergência de um *random walk* para um processo de difusão sugerida na última seção − isto é, fazendo-se $n \to \infty$ e $\varepsilon \to 0$ − traz em si a possibilidade de se estender, para uma formulação contínua, a abordagem discreta que até então foi utilizada na construção do *Diffusion Maps*. Inserindo-se na formulação contínua a informação sobre a densidade de $\mho$, torna-se possível a obtenção das já citadas famílias de difusão.

Para isso, forme o *kernel*

$$k_\varepsilon(x,y) = e^{-\frac{\|x-y\|^2}{\varepsilon}} \tag{2-19}$$

e, assumindo que o conjunto $X$ seja toda a variedade $\mho$, obtenha

$$q_\varepsilon(x) = \int_X k_\varepsilon(x,y)q(y)dy \tag{2-20}$$

como aproximação da verdadeira densidade $q(x)$. Para um parâmetro real $\alpha$, forme um novo *kernel*

$$k_\varepsilon^{(\alpha)}(x,y) = \frac{k_\varepsilon(x,y)}{q_\varepsilon^\alpha(x)q_\varepsilon^\alpha(y)} \tag{2-21}$$

Definindo

$$d_\varepsilon^{(\alpha)}(x) = \int_X k_\varepsilon^{(\alpha)}(x,y)q(y)dy \tag{2-22}$$

normaliza-se o *kernel* obtido para então se obter

$$p_{\varepsilon,\alpha}(x,y) = \frac{k_\varepsilon^{(\alpha)}(x,y)}{d_\varepsilon^{(\alpha)}(x)} \tag{2-23}$$

Os elementos $p_{\varepsilon,\alpha}(x,y)$ formam assim o operador $P_{\varepsilon,\alpha}$ definido por

$$P_{\varepsilon,\alpha}f(x) = \int_X p_{\varepsilon,\alpha}(x,y)f(y)q(y)dy \tag{2-24}$$

que é o equivalente contínuo da matriz de transição de probabilidades.

Coifman e Lafon[7] demonstram que, no limite $\varepsilon \to 0$, as autofunções de $P_{\varepsilon,\alpha}$ são aproximações das autofunções do seguinte operador de Schrodinger

$$\Delta\phi - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}}\phi \tag{2-25}$$

onde $\phi$ representa as autofunções do operador de Laplace-Beltrami $\Delta$ que descreve a difusão do calor na variedade $\mho$. Este importante resultado permite analisar, para cada valor escolhido do parâmetro $\alpha$, a estrutura do operador produzido. Aqui será feita uma descrição rápida sobre os operadores produzidos por dois valores particulares de $\alpha$.

Para $\alpha = 0$, o operador assume a forma

$$\Delta\phi - \frac{\Delta(q)}{q}\phi \tag{2-26}$$

e deste modo, se a densidade $q$ for uniforme, obtém-se simplesmente o operador de Laplace-Beltrami sobre a variedade. Convém destacar que, ao se fazer $\alpha = 0$, a obtenção do operador de difusão correspondente se reduz imediatamente à construção clássica do laplaciano normalizado de um grafo, já mencionada anteriormente. Isto equivale aos resultados de Belkin em[2], desde que a densidade seja uniforme. É importante dizer que Belkin considera a hipótese de densidade uniforme em todas as suas aplicações, o que nem sempre acontece na prática.

Já para $\alpha = 1$, o operador assume automaticamente a forma $\Delta\phi$, significando que, novamente, quem está sendo aproximado é o operador de Laplace-Beltrami, mas agora, independentemente da densidade dos dados sobre a variedade.

Assim, ao se construir as coordenadas de difusão de uma amostra representativa de uma variedade, o que se deve esperar é que a utilização de $\alpha = 0$ produza um mapeamento que aponte as características estatísticas da amostra, revelando algo sobre a densidade dos pontos tomados. Por outro

lado, a utilização de $\alpha = 1$ deverá produzir um mapeamento que recupere tão-somente informações geométricas da amostra.

Como ilustração, considere uma amostra de 2000 pontos tomados sobre uma superfície esférica, de modo que uma maior densidade destes pontos seja verificada nos polos, como mostra a figura 2.13.

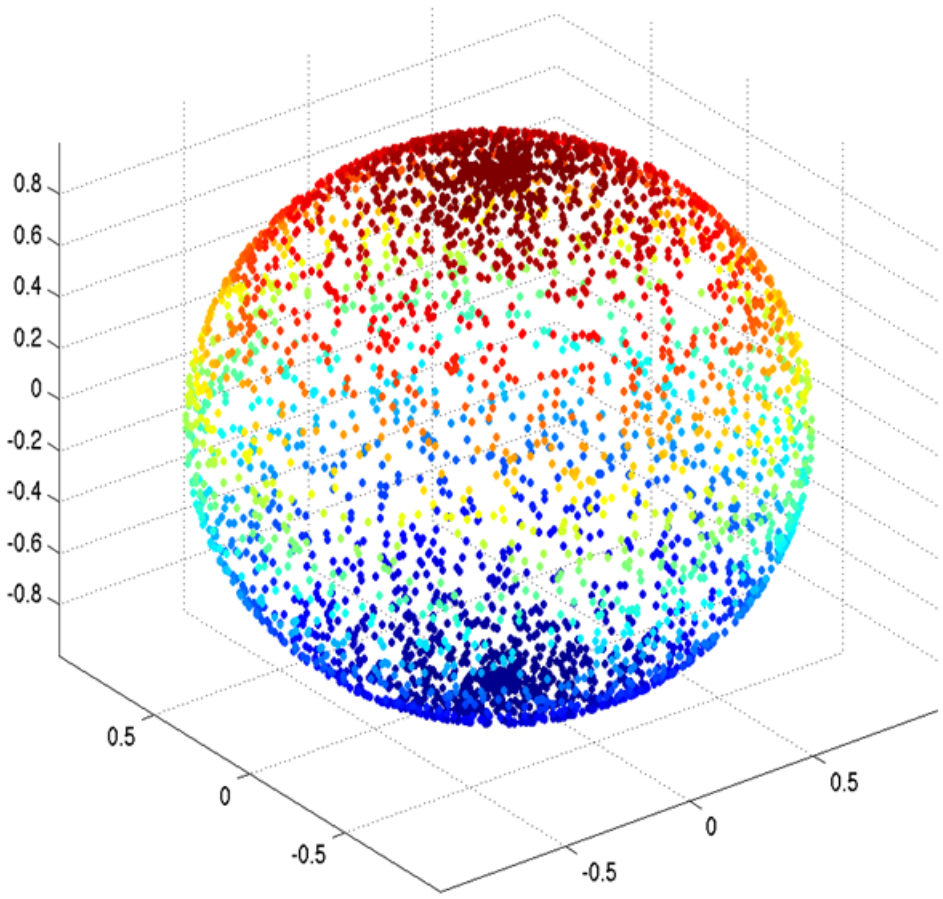


Figura 2.13: Amostra de 2000 pontos distribuídos não-uniformemente sobre uma esfera.

A figura 2.14 mostra resultados obtidos pela aplicação do *Diffusion Maps* a esta amostra.

Como explicado anteriormente, estes resultados são coerentes com a

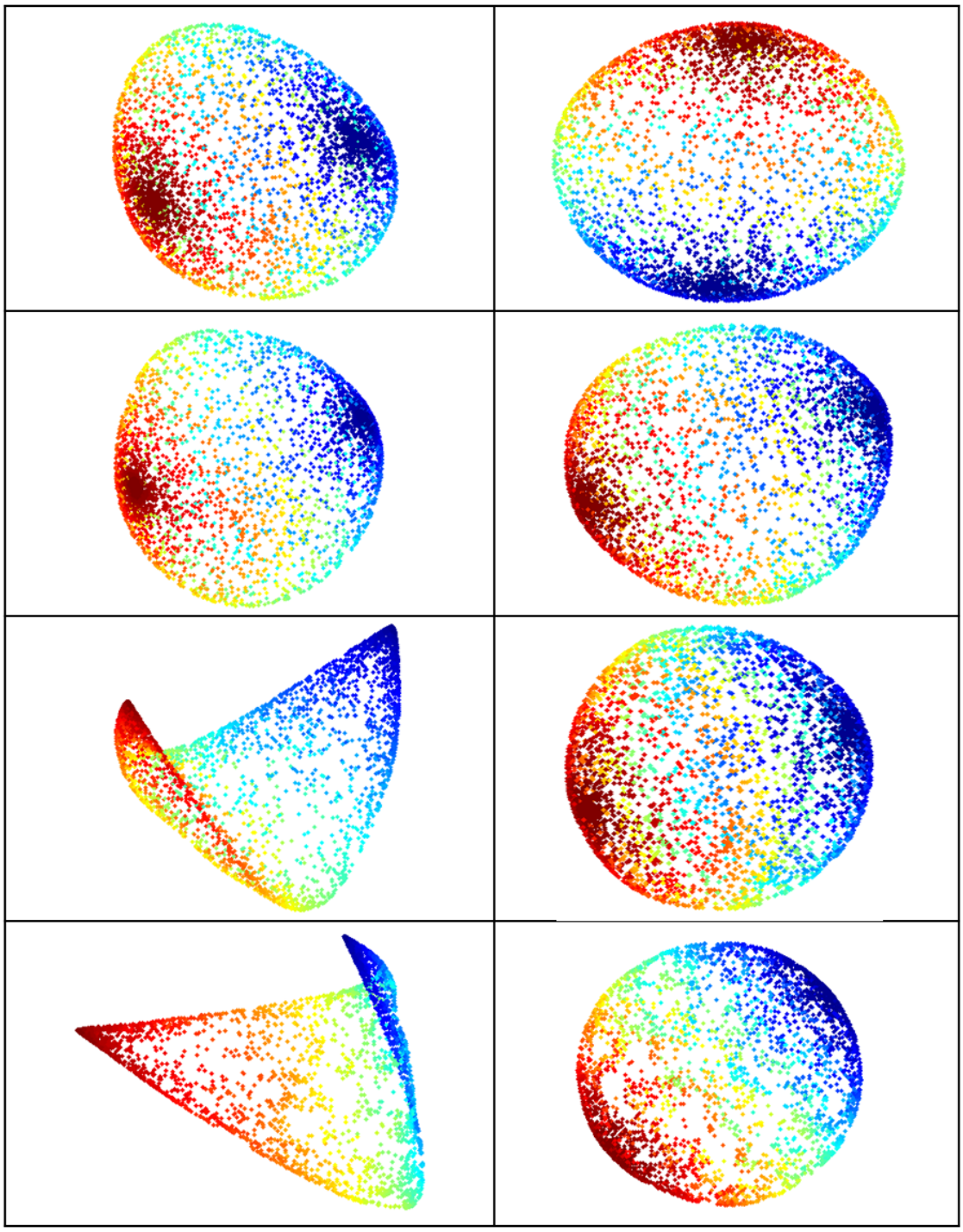

Figura 2.14: *Diffusion Maps* da amostra de pontos sobre a esfera. A coluna da esquerda mostra resultados obtidos para $\alpha = 0$; a da direita, para $\alpha = 1$. De cima para baixo, os valores de $\varepsilon$ são 0.5625, 0.25, 0.0625 e 0.015625.

teoria: para uma amostra significativa e $\varepsilon$ "pequeno", o parâmetro $\alpha = 0$ acentua as informações sobre a densidade contidas na amostra, ao passo que $\alpha = 1$ recupera integralmente as informações geométricas.