UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Matemáticas y Computacionales

TÍTULO:

AN INTERACTIVE TOOL TO DATA ANALYSIS VISUALIZATION TECHNIQUES.

Autor:

Martín Vélez Falconí

Tutor:

Diego Hernán Peluffo-Ordóñez, Ph.D.

Urcuquí, octubre del 2020.

Urcuquí, 30 de septiembre de 2020

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES
CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN
ACTA DE DEFENSA No. UITEY-ITE-2020-00034-AD

A los 30 días del mes de septiembre de 2020, a las 15:00 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.
Miembro No Tutor	Dra. GUACHI GUACHI, LORENA DE LOS ANGELES , Ph.D.
Tutor	Dr. PELUFFO ORDONEZ, DIEGO HERMAN , Ph.D.

El(la) señor(ita) estudiante **VELEZ FALCONI, MARTIN** , con cédula de identidad No. **1724556681**, de la **ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES**, de la Carrera de **TECNOLOGÍAS DE LA INFORMACIÓN**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-43-No.496-2014**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **AN INTERACTIVE TOOL TO DATA ANALYSIS VISUALIZATION TECHNIQUES**, previa a la obtención del título de **INGENIERO/A EN TECNOLOGÍAS DE LA INFORMACIÓN**.

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

Tutor	Dr. PELUFFO ORDONEZ, DIEGO HERMAN , Ph.D.

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Miembro Tribunal De Defensa	Dra. GUACHI GUACHI, LORENA DE LOS ANGELES , Ph.D.	10,0
Tutor	Dr. PELUFFO ORDONEZ, DIEGO HERMAN , Ph.D.	10,0
Presidente Tribunal De Defensa	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.	8,0

Lo que da un promedio de: **9.3 (Nueve punto Tres)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

VELEZ FALCONI, MARTIN
Estudiante

Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.
Presidente Tribunal de Defensa

Dr. PELUFFO ORDONEZ, DIEGO HERMAN , Ph.D.
Tutor

Dra. GUACHI GUACHI, LORENA DE LOS ANGELES , Ph.D.
Miembro No Tutor

MEDINA BRITO, DAYSY MARGARITA
Secretario Ad-hoc

AUTORÍA

Yo, **Martín Vélez Falconí**, con cédula de identidad 1724556681, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el/la autora (a) del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, noviembre 2020.

Martín Vélez Falconí

CI: 1724556681

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Martín Vélez Falconí**, con cédula de identidad 1724556681, cedo a la Universidad de Investigación de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior

Urcuquí, noviembre 2020.

Martín Vélez Falconí

CI: 1724556681

Dedication

Dedicado a mi familia y a todos los que me ayudaron y me apoyaron, especialmente, a mi amigo y tutor Diego, y a mi perrita Pato.

Martín Vélez Falconí

Agradecimientos

Agradezco el apoyo brindado por el grupo de investigación "SDAS Research Group", y a, su director, Diego Peluffo por incluirme en este grupo y motivarme a investigar.A Federico Zertuche por los consejos que me ha dado para realizar este trabajo. Y finalmente, pero no menos importante, a mi familia, quien ha sido un soporte para mí.

Martín Vélez Falconí

Resumen

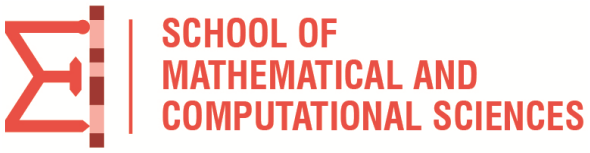
El área de dimensión de reducción (RD) tiene como propósito proveer maneras de aprovechar información de alta dimensionalidad, a través de la generación de una representación de en baja dimensionalidad, siguiendo algún criterio de preservación de estructura. En la literatura científica, se encuentran docenas de técnicas de reducción de la dimensionalidad. Sin embargo, la selección de un método adecuado para la reducción de dimensionalidad es una situación frecuente y no representa una tarea trivial. Para realizar una reducción adecuada, podría incorporarse el criterio de expertos en el proceso de análisis, de forma que los expertos necesitan interactuar dinámicamente con las representaciones de la baja dimensionalidad. Dicha interacción puede lograrse con los diferentes modelos interactivos descritos en la literatura. No obstante, aún hay problemas abiertos relacionados con la interacción dinámica del usuario con los datos. En este trabajo, se presenta un modelo interactivo, llamado "Inverse Data Visualization Framework" (IDVF), el cual es un modelo pionero de visualización interactiva que se basa en la aproximación, por métodos kernels, de una reducción de dimensionalidad dada por un experto. En términos generales, el modelo opera de la siguiente manera: Inicialmente, se muestra un gráfico de dispersión de los datos en baja dimensionalidad generado por métodos espectrales. Seguidamente, se solicita al usuario agrupar (a su criterio) algunos puntos del gráfico de acuerdo a la representación que considere más idónea. Una vez generado los datos, el modelo tratará de generar una representación de la misma dimensionalidad con una forma similar al creado por el usuario, mezclando diferentes aproximaciones de métodos espectrales en forma de matrices kernel y otras matrices kernels que son comúnmente usados para la reducción de dimensionalidad. Todo esto se desarrolla usando el método de análisis de componentes principales con kernel (Kernel PCA). La mezcla de las matrices kernel, después de la descomposición espectral de KPCA, deberá generar un gráfico de 2 dimensiones que resultará similar a la representación dada por el usuario.

Palabras clave: Reducción de dimensionalidad, modelo de interacción, funciones kernel, visualización de datos.

Abstract

Broadly, the area of dimensional reduction (DR) is aimed at providing ways to harness high dimensional (HD) information through the generation of lower dimensional (LD) representations, by following a certain data-structure-preservation criterion. In literature there have been reported dozens of DR techniques, which are commonly used as a preprocessing stage within exploratory data analyses for either machine learning or information visualization (IV) purposes. Nonetheless, the selection of a proper method is a nontrivial and -very often- toilsome task. In this sense, a readily and natural way to incorporate an expert's criterion into the analysis process, while making this task more tractable is the use of interactive IV approaches. Regarding the incorporation of experts' prior knowledge there still exists a range of open issues. In this degree thesis, we introduce a here-named Inverse Data Visualization Framework (IDVF), which is an initial approach to make the in-put prior knowledge directly interpretable. Our framework is based on 2D-scatter-plots visuals and spectral kernel-driven DR techniques. To capture either the user's knowledge or requirements, users are requested to provide changes or movements of data points in such a manner that resulting points are located where best convenient according to the user's criterion. Next, following a Kernel Principal Component Analysis approach and a mixture of kernel matrices, our framework accordingly estimates an approximate LD space. Then, the rationale behind the proposed IDVF is to adjust as accurately as possible the resulting LD space to the representation while fulfilling users' knowledge and requirements. Results are greatly promising and open the possibility to novel DR-based visualizations.

Keywords: Dimensionality reduction, interaction model, kernel functions, data visualization.



YACHAY TECH UNIVERSITY

DEGREE THESIS

AN INTERACTIVE TOOL TO DATA ANALYSIS VISUALIZATION TECHNIQUES

Author:
Martin
VÉLEZ FALCONI

Advisor:
Prof. Diego Hernán
PELUFFO-ORDÓÑEZ

A thesis submitted in fulfillment of the requirements for the degree of Information Technologies

in the

SCHOOL OF MATHEMATICAL AND COMPUTATIONAL SCIENCES

December 3, 2020

Dedicado a mi familia y a todos los que me ayudaron y me apoyaron, especialmente, a mi amigo y tutor Diego, y a mi perrita Pato...

Agradecimientos

Agradezco el apoyo brindado por el grupo de investigación "SDAS Research Group", y a, su director, Diego Peluffo por incluirme en este grupo y motivarme a investigar.
A Federico Zertuche por los consejos que me ha dado para realizar este trabajo. Y finalmente, pero no menos importante, a mi familia, quien ha sido un soporte para mí.

Resumen

El área de dimensión de reducción (RD) tiene como propósito proveer maneras de aprovechar información de alta dimensionalidad, a través de la generación de una representación de en baja dimensionalidad, siguiendo algún criterio de preservación de estructura. En la literatura científica, se encuentra docenas de técnicas de reducción de la dimensionalidad. Sin embargo, la selección de un método adecuado para la reducción de dimensionalidad es una situación frecuente y no representa una tarea trivial. Para realizar una reducción adecuada, podría incorporarse el criterio de expertos en el proceso de análisis, de forma que los expertos necesitarían interactuar dinámicamente con las representaciones de la baja dimensionalidad. Dicha interacción puede lograrse con los diferentes modelos interactivos descritos en la literatura. No obstante, aún hay problemas abiertos relacionados con la interacción dinámica del usuario con los datos. En este trabajo, se presenta un modelo interactivo, llamado "Inverse Data Visualization Framework" (IDVF), el cual es un modelo pionero de visualización interactiva que se basa en la aproximación, por métodos kernels, de una reducción de dimensionalidad dada por un experto. En términos generales, el modelo opera de la siguiente manera: Inicialmente, se muestra un gráfico de dispersión de los datos en baja dimensionalidad generado por métodos espectrales. Seguidamente, se solicita al usuario agrupar (a su criterio) algunos puntos del grafico de acuerdo a la representación que considere más idónea. Una vez generado los datos, el modelo tratará de generar una representación de la misma dimensionalidad con una forma similar al creado por el usuario, mezclando diferentes aproximaciones de métodos espectrales en forma de matrices kernel y otras matrices kernels que son comúnmente usados para la reducción de dimensionalidad. Todo esto se desarrolla usando el método de análisis de componentes principales con kernesl (Kernel PCA). La mezcla de las matrices kernel, después de la descomposición espectral de KPCA, deberá generar un gráfico de 2 dimensiones que resultará similar a la representación dada por el usuario.

Palabras clave: Reducción de dimensionalidad, modelo de interacción, funciones kernel, visualización de datos.

Abstract

Broadly, the area of dimensional reduction (DR) is aimed at providing ways to harness high dimensional (HD) information through the generation of lower dimensional (LD) representations, by following a certain data-structure-preservation criterion. In literature there have been reported dozens of DR techniques, which are commonly used as a preprocessing stage within exploratory data analyses for either machine learning or information visualization (IV) purposes. Nonetheless, the selection of a proper method is a nontrivial and -very often- toilsome task. In this sense, a readily and natural way to incorporate an expert's criterion into the analysis process, while making this task more tractable is the use of interactive IV approaches. Regarding the incorporation of experts' prior knowledge there still exists a range of open issues. In this degree thesis, we introduce a here-named Inverse Data Visualization Framework (IDVF), which is an initial approach to make the input prior knowledge directly interpretable. Our framework is based on 2D-scatter-plots visuals and spectral kernel-driven DR techniques. To capture either the user's knowledge or requirements, users are requested to provide changes or movements of data points in such a manner that resulting points are located where best convenient according to the user's criterion. Next, following a Kernel Principal Component Analysis approach and a mixture of kernel matrices, our framework accordingly estimates an approximate LD space. Then, the rationale behind the proposed IDVF is to adjust as accurately as possible the resulting LD space to the representation while fulfilling users' knowledge and requirements. Results are greatly promising and open the possibility to novel DR-based visualizations.

Keywords: Dimensionality reduction, interaction model, kernel functions, data visualization.

Contents

List of Figures

List of Tables

List of Abbreviations

Notation	Description
IDVF	Inverse Data Visualization Framework
DR	Dimensional Reduction
HD	High Dimensional
LD	Lower Dimensional
2D	Two Dimensional
IV	Information Visualization
PCA	Principal Component Analyisis
KPCA	Kernel PCA
Isomap	Isometric Mapping
CMDS	Classical Multidimensional Scaling
LLE	Locally Linear Embedding
LE	Laplacian Eigenmaps
KIsomap	Kernel Matrix approximation Isometric Mapping
KCMDS	Kernel Matrix approximation Classical Multidimensional Scaling
KLLE	Kernel Matrix approximation Locally Linear Embedding
KLE	Kernel Matrix approximation Laplacian Eigenmaps
IM	Interactive Models

TABLE 1: Acronyms considered in this work.

List of Abbreviations

Notation	Description
D	Dimension of the high-dimensional, input matrix
d	Dimension of the low-dimensional matrix ($D \geq d$)
m	Number of Kernels Matrices
$\boldsymbol{X}_{n \times d}$	High dimensional matrix
$\boldsymbol{Y}_{n \times D}$	Low dimensional matrix produced by KPCA
$\widehat{\boldsymbol{Y}}_{n \times D}$	Low dimensional matrix selected by the user
$\boldsymbol{\alpha}_m$	Vector of m coefficients
$\boldsymbol{I}_n$	$n \times n$ Identity Matrix
$\boldsymbol{K}_{n \times n}^{(m)}$	m-th kernels matrix
$\widetilde{\boldsymbol{K}}_n$	Mixture (Kernel Matrix constructed by the linear combination of the given kernels matrices)
$\boldsymbol{M}_n$	Matrix generated by eigenvectors and eigenvalues matrix.
$\boldsymbol{V}_n$	Eigenvector matrix.
$\boldsymbol{D}_n$	Eigenvalue matrix.
$\widehat{\boldsymbol{M}}_n^*$	Approximation matrix from the d eigenvectors of the d maximum eigenvalues of $\widehat{\boldsymbol{Y}}$
$\widehat{\boldsymbol{M}}_n$	Matrix generated by the vectors $\widehat{\boldsymbol{Y}}$
λ_i	i-th eigenvalue.
$\boldsymbol{v}_i$	i-th eigenvector.
$\boldsymbol{y}_i$	Vector from dimensional reduction.
$\widehat{\boldsymbol{y}}_i$	Vector from the user expected representation.
R_{NX}	Average Agreement Rate.

TABLE 2: Mathematical Notation

Chapter 1

Introduction

High-dimensional (HD) data requires an arduous and extensive analysis that exceeds the human senses and may even deceive the human perception. The wide and ubiquitous field of Computer Science refers to the HD data analysis as an Information Visualization (IV) problem. The area of IV aims to generate natural visual representations to simplify the data interpretation by the user. Particularly, the visualization approaches powered by low-dimensional (LD) spaces (mainly at 2D or 3D) represent a very appealing and outstanding alternative. In this connection, the Dimensionality Reduction (DR) techniques have taken place as a crucial stage for such approach, here named as DR-based IV. According to [1], the most remarkable DR techniques reported by literature are the original versions and variants of Principal Component Analysis (PCA), Classical Multidimensional Scaling (CMDS), locally linear embedding (LLE), Laplacian eigenmaps (LE), Stochastic Neighbor Embedding (SNE). Besides strengths and weaknesses of each DR method, there are some approaches to select a DR algorithm [2], [3]. As a result, there is growing necessity for an interactive approaches enabling (even non-expert) users assess each method and select the one(s) that best fit(s) the data set.

The state of the art reports some approaches to add interactivity through a so-called Interaction Models (IM) such as: geometric [4], equalizer-like [5], color-based [6], [7], geodesic [8] models, and among others, as extensively reviewed in [9]. The intuition behind these models is to find a kernel matrix, create a linear combination between kernel matrices, and adjust the weights in order to obtain the desired representation [10]. Since methods as LLE, LE, and ISOMAP [11] are susceptible to be represented as kernel matrices, it is possible to explore the DR of the data from a linear combination of kernel matrices using a Kernel DR method, namely Kernel PCA (KPCA) [12]

1.1 Problem statement

The task of selecting an appropriate kernel (or multiples kernels) and their criteria within a KPCA framework is challenging -even for experts. More detailed interactive tools -as those based on IM and kernel matrices [9]- propose diverse interfaces for dynamic selection of kernel combination approaches. In order to pick to the best kernel, the users should explore multiple options -e.g. try several kernels and their combinations. As natural, the option that a kernel perfectly suits the user's needs does not exist is also possible. However, these mechanisms are not efficient enough when user has no a proper interpretation of the data structure.

1.2 Contribution

As an alternative to tackling these issues, in this work, we present a novel interactive DR based on choosing/setting a suitable kernel for KPCA from the lower dimensional space defined by a user. Specifically, we introduce a here-named Inverse Data Visualization Framework (IDVF). Broadly, IDVF works as follows: It uses spectral DR techniques based on kernels. Its visualization consist of 2D-scatter-plots. Then, to capture either the user's knowledge or requirements, users are requested to provide changes or movements of data points in such a manner that resulting points are located where best convenient according to the user's criterion. Subsequently, we estimates an approximate LD space from a a mixture of kernel matrices inputting to a KPCA approach. Therefore, the main goal of IDVF is to adjust as accurate as possible the resulting LD space to the representation fulfilling users' knowledge and requirements. Results are greatly promising and open the possibility to novel DR-based visualizations approaches.

1.3 Document Organization

This work is divided into six Chapters as follows:

- Chapter 1 (Introduction) outlines the general aspects of the work, the problem statement 1.1, the contribution 1.2, and the objectives 1.4.
- Chapter 2 (Overview and Background) explains the main idea of interactive model for dimensional reduction (Section 2.1), the previous interactive models (Section 2.2), the dimensional spectral techniques

(Section 2.3) and their kernel approximation (Section 2.5.1), the kernel matrix used in this work (Section 2.5), the generation of the mixture kernel (Section 2.6), and finally the quality used to evaluate the proposed model (Section 2.7).

- Chapter 3 (Methodology) gathers stages for the operation of the IDVF. The methodology itself and the flow-chart are explained in the section 3.2.

- Chapter 4 (Experimental setup) describes the considered databases (Section 4.1), the kernels used for the experiments (Section 4.2), and the metrics (Section 4.3).

- Chapter 5 (Results) is divided into three sections 5.1, 5.2, and 5.3. Each section has three trials and at each trial the plot of the representation and their quality curve are depicted.

- Chapter 6 (Conclusion and future work) draws the final remarks as follows: Conclusions in Section 6.1, and the future work 6.2.

1.4 Objectives

1.4.1 General Objective

To develop an interactive model for dimensionality-reduction-based visualization able to produce a similar low-dimensional space to that beforehand given by an user through applying the best combination of kernel matrices into a Kernel Principal Component Analysis (KPCA) framework.

1.4.2 Specific Objectives

- To create an interactive scatter plot enabling the user to drag and move points in order to create an interface for the interactive model.

- To combine different dimensionality reduction methods and kernels matrices to produce a set of methods according the intuition of the user.

- To develop a mathematical model based on two vectors and high-dimensional datasets, aimed at approximating the KPCA outcomes to two 2D scatter plot giving by an user.

Chapter 2

Overview and Background

2.1 Context

Dimensional reduction (DR) transform a high dimesional data into a substantial representation of lower dimensionality, an optimal representation is an inherent dimensional depiction of data, in other words the minimum number of parameters that the data can preserve their features [13]. However, interactive models described in [5] have an additional aimed, which is to produce an approximation of non-legible data to a more comprehensible representation according to the expert judgement. This thesis focuses on interactive models of weighting factors commonly used to obtain the best DR set from a linear combination of kernel matrices, which becomes a mixture Kernel Matrix ($\widetilde{K}$). Such a Kernel Matrix is used as an input of a kernel-based DR method -e.g. Kernal Pricnipal Component Analysis (KPCA). The linear combination of weighting factors are manipulated by an Interactive Model (IM) [5].

Previous works [4]–[7], [14] have reported different interaction models based on linear mixtures, wherein the users may select the methods within an intuitive approach. Yet, the use of these interfaces to determine structure of the data represents is an ardours task.

2.2 Previous interactive models

This section introduces active interactive models based on combination of kernels approximation methods. These models are based on the combination of the weighting factors of kernels matrices.

2.2.1 Geometrical Homotopy Model

This method is based on the combination of different kernel methods, each vertex of a polygon with M vertexes represent a kernel [15]. The model only allows a pairwise combination of kernels. The relation between the 2 kernel matrices is tuned by a given parameter λ, which is manipulated according to its position in the edge, a visibly representation of it is in Figure. 2.1.

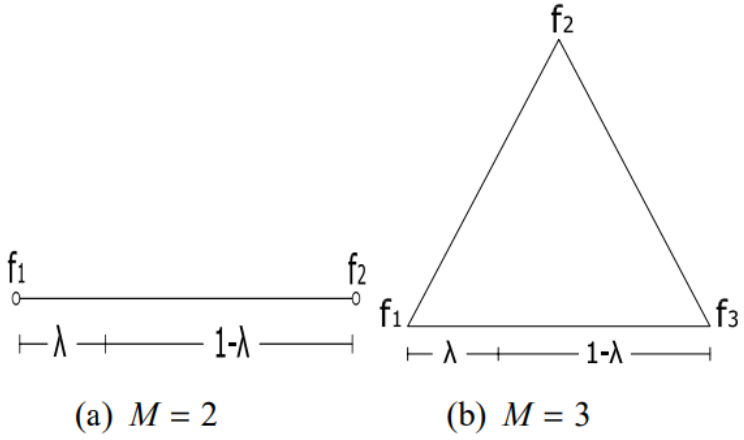


FIGURE 2.1: Representation of Homotopy model with $M = 2$ and $M = 3$. Source: [15].

2.2.2 Color-Based Model

The color-based model described in [7] uses weighting factors and three reference colors (red, blue, and green). Each color represents a kernel matrix, the interface allows for selecting a colors combination among them. The weighting factors are calculated with the percentage of the reference colors inside the selected color, as shown in Figure 2.2.

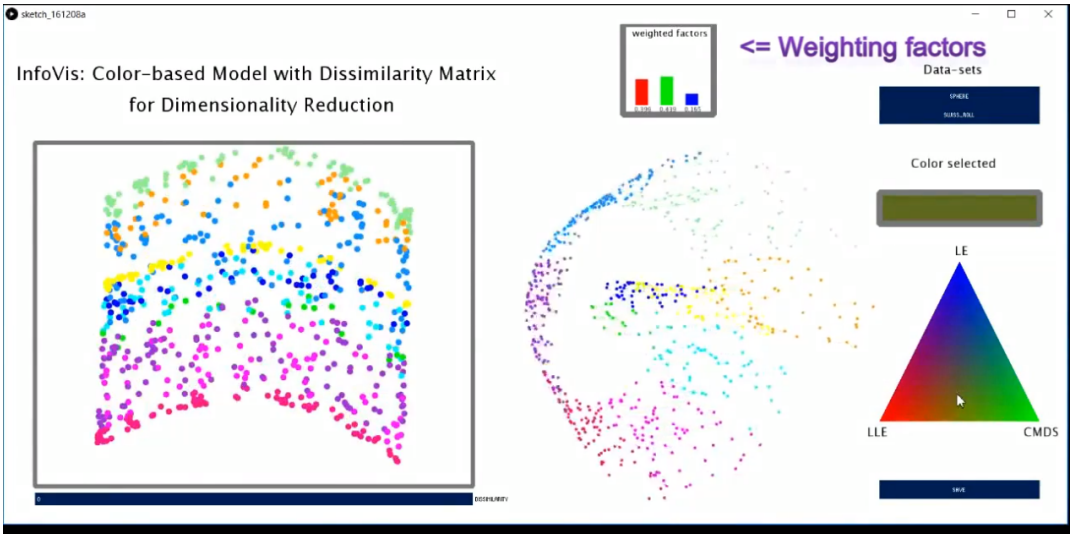


FIGURE 2.2: GUI of color-based model. Source: [7].

2.2.3 DataVisSim

Likewise, `DataVisSim` is another model based on weighting factors. The rate of the weighting factor is giving by an equalizer-bar. Each bar represented the coefficient value of a kenel matrix [6]. Figure 2.3 shows the interface of `DataVisSim`.

FIGURE 2.3: GUI of `DataVisSim`. Source: [6].

2.3 Dimensional Reduction Spectral Techniques

There are many different techniques to reduce the dimensional of dataset. The aim of the thesis is particularly interested in non linear techniques, specially on spectral techniques based in similitude, dissimilarities, and kernel.

The concerning techniques which are clearly related with this thesis are explained in the subsections 2.3.1, 2.3.2, 2.3.3, 2.3.4, and 2.4.

2.3.1 (Classical) Multi-Dimensional Scaling

(Classical) Multi-Dimensional Scaling (CMDS/MDS) generates a dense features maps or an embedding coordinates set from matrix of similarity generated between pairs of data points. This method can be understood a scaling process over a multidimensional data set regarding a target space [16]. A restriction of CMDS is that the objective space must be an Euclidean space. In contrast, MDS uses any kind of similarity (also kernel) matrix, which can be analyzed with non-linear techniques [17].

Additionally, Principal Component Analysis has the same minimization function metric of MDS, as is shown in the master thesis [17].

2.3.2 Isomap

Isometric Mapping works in a similar way to MDS. The similarity matrix used for Isomap is generated by the geodesic distance of the points [18]. The geodesic distance measures the topological distance [17].

2.3.3 Laplacian Eigenmaps (LE)

Laplacian Eigenmaps is a DR algorithm, which holds low-computational complexity and robustness to noise and outliers. The algorithm generates a Laplacian graph followed from the inherent geometric structure of the manifold [19]. This is justified by the fact that the Laplacian graph can be seen as an approximation of the Laplacian operators providing an optimal embedding space.

2.3.4 Locally Linear Embedding

Locally Linear Embedding (LLE) is a non-supervised DR method that seeks the smallest-neighborhood-preserving embeddings. It computes a low-dimensionality

linear reconstruction from linear local symmetries of a non-linear, high-dimensional embedding space [20].

2.4 Kernel PCA

PCA has been extended in different non-linear generalizations. One of the most remarkable ones is Kernel PCA (KPCA), which is based on a mapping of the original data onto a higher-dimensional space, and the kernel trick to estimate the principal components from a non-linear representation. The principles of KPCA are widely described in [21]–[24].

The kernel trick maps the input samples onto a so-named high-dimensional feature space $\mathcal{F}$. The kernel function $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2)$ satisfies that:

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2) \rangle, \tag{2.1}$$

where $\phi(\cdot)$ maps $\mathbb{R}^{n \times D}$ onto $\mathcal{F}$ and $\langle \cdot, \cdot \rangle$ stands for inner product.

The benefit of the kernel trick is the fact that the inner product on the feature space can be replaced by a kernel function and reduce a problem from $\mathcal{F}$ to $\mathbb{R}^n$. KPCA optimization problem can be understood as an eigenvalue problem of selecting the eigenvectors associated to the largest eigenvalues of a Gram matrix $\boldsymbol{K}$ holding pairwise kernel function values. A fully matrix development of KPCA is introduced in [24].

2.5 Kernels Matrices

This work takes advantage of the equivalent matrices as performing a DR process, when using neighborhood structure preservation through KPCA, as explained in [12]. Table 2.1 gathers the kernel matrices used in this work. In addition to DR-based kernel matrices, we use other common kernels [25] described in Table 2.2.

2.5.1 Kernels based in DR methods

DR Technique	Description	Reference

Kernel LE	Kernel LE constructs an Laplacian matrix $\boldsymbol{L}$, from the neighborhood relations graph $\boldsymbol{W}$ and its corresponing degree $\boldsymbol{D}$, as follows: $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$. Then, the KLE is defined as the pesudoinverse of the Laplacian matrix as: $$\boldsymbol{K}_{\text{LE}} = \boldsymbol{L}^{\dagger} \quad (2.2)$$	[12], [26], [27]
Kernel Isomap	To calculate the kernel matrix the two first steps of the Isomap algorithm are performed. It constructs the neighborhood graph and compute the geodesic distances which is defined as $\boldsymbol{D}^2$. Then, it constructs the kernel matrix $\boldsymbol{K}\left(\boldsymbol{D}^2\right) = -\frac{1}{2}\boldsymbol{H}\boldsymbol{D}^2\boldsymbol{H}$. However, this matrix $\boldsymbol{K}$ is no positive definite, in order to get a kernel, the largest eigenvalues c^* of $\begin{pmatrix} 0 & 2\boldsymbol{K}(\boldsymbol{D}^2) \\ a & -4\boldsymbol{K}(\boldsymbol{D}) \end{pmatrix}$ must be calculated. Then, the Mercer Kernel matrix is applied as: $$\boldsymbol{K}_{\text{Isomap}} = \boldsymbol{K}\left(\boldsymbol{D}^2\right) + 2c\boldsymbol{K}\left(\boldsymbol{D}\right) + \frac{1}{2}c^2\boldsymbol{H} \quad (2.3)$$ $\boldsymbol{K}_{\text{Isomap}}$ is positive definite if $c \geq c^*$.	[11], [28], [29]
Kernel LLE	It builds a weight matrix $\boldsymbol{W}$ and defines a matrix $\boldsymbol{M}$, so that $\boldsymbol{M} = (\boldsymbol{I}_n - \boldsymbol{W}^{\top})(\boldsymbol{I} - \boldsymbol{W})$ and compute the maximum eigenvalue c of $\boldsymbol{M}$. Then, its kernel matrix can be written as: $$\boldsymbol{K}_{\text{LLE}} = (c\boldsymbol{I}_n - \boldsymbol{M}) \quad (2.4)$$	[30]–[32]

Kernel CMDS	The Kernel version approximation of CMDS is based in a distance matrix $\boldsymbol{D}$, and it is double centered. For this article $\boldsymbol{D}$ is based on the euclidean distance. $$\boldsymbol{K}_{\text{CMDS}} = \frac{1}{2}(\boldsymbol{I} - \mathbf{1}_N\mathbf{1}_N^T)\boldsymbol{D}(\boldsymbol{I} - \mathbf{1}_N\mathbf{1}_N^T) \quad (2.5)$$	[24]

TABLE 2.1: Kernel matrices representing spectral DR techniques.

2.5.2 Classic kernels functions

Kernel Function	Definition
Linear	$x^\top y$
Polynomial	$(\gamma x^\top y + c)^D$
Sigmoid	$\tanh(\gamma x^\top y + c)$
Radial basis function	$\exp(-\gamma \parallel x - y \parallel^2)$
Laplacian	$\exp(-\gamma \parallel x - y \parallel_1)$

TABLE 2.2: Kernel Functions and its definitions

2.6 Mixtures of Kernel Matrices

A kernel matrix or Gram matrix is a square, symmetric positive definite matrix, such that its entries can be represented by a kernel function [33].

In this sense, as discussed in [34], the linear combination of kernel matrices is a also a kernel Matrix [34]. By taking advantage of this property, multiple kernel analysis for DR [2], and interaction models for visualization [35] have been proposed.

Let us define the matrix $\widetilde{\boldsymbol{K}}$ as a linear combination of kernels matrices, as follows:

$$\widetilde{\boldsymbol{K}}\left(\boldsymbol{X}_{n\times D}\right) = \sum_{i=1}^{m} \alpha_i \boldsymbol{K}^{(i)}\left(\boldsymbol{X}_{n\times D}\right) \quad \text{if } \alpha_i \geq 0 \text{ and if } \alpha_i \in \mathbb{R} \ . \tag{2.6}$$

2.7 Quality Curve

Lee and Verleysen [36], introduced a formal evaluation measurement of dimensionality reduction in form of a ranking-based metric. Let as denote $L < D$ and $|\cdot|$ the set cardinality, the authors represent δ_{ij} as the distance between two elements $(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$ of high dimensional dataset $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_i, \ldots, \boldsymbol{\xi}_n\} \subset \mathbb{R}^D$. Analogously, d_{ij} is the distance between two elements (x_i, x_j) of the low dimensions dataset $\boldsymbol{X} = \{\boldsymbol{x}_i, \ldots, \boldsymbol{x}_j\} \subset \mathbb{R}^L$. Then, the rank of ξ_j with respect to ξ_i in $\mathbb{R}^D$ is given by $\rho_{ij} = |\{k|\delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leq k < j \leq n\})|$, while the Rank of x_j with respect to x_i in $\mathbb{R}^L$ is given by $r_{ij} = |\{k|d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq n\})|$. Thus, reflexive ranks are set to zero $(\rho_{i,i} = r_{i,i} = 0)$ and non-reflexive ranks belong to $\{1, \ldots, n-1\}$. The definition of a co-ranking matrix allows to compare different rank based criteria. It is defined as:

$$q_{kl} = |\{(i,j) : \rho_{ij} = k \wedge r_{ij} = l\}|. \tag{2.7}$$

The core of the *Quality Criterion* of [36] as expressed by [37] is the matrix Q_{NX}:

$$Q_{NX}(K) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{l=1}^{L} q_{kl}. \tag{2.8}$$

Q_{NX} describes the agreement among HD and LD neighborhood, where $Q_{NX} = 1$ represents the ideal K-ary neighborhood agreement and $Q_{NX} = 0$ represents the absence of agreement.

The overall result is the curve generated by the following equation: [38], [39]:

$$R_{NX}(K) = \frac{(N-1)Q_{NX}(K) - K}{N - 1 - K}. \tag{2.9}$$

Finally, an indicator of quality is its area under the curve.

Chapter 3

Methodology

3.1 Outline

Inverse Data Visualization Framework (IDVF) tunes the coefficients of the kernel matrices in order to achieve the best approximation of DR structure provided by the user. In such vein, the user avoids manual adjustment of the weights.

Given a scatter plot of a DR structure of a High Dimensional Data set, the main idea of IDVF is to find the advantageously Mixture Kernel Matrix $\widetilde{\boldsymbol{K}}$. The expected result is to get a similar low dimensional data set to the structure by applying KPCA ($\widetilde{\boldsymbol{K}}$).

In broad terms, the rationale of IDVF is to estimate the coefficients or weighting factors for a mixture of kernel matrices, which is aimed to map $\boldsymbol{X}_{n\times D}$ onto $\boldsymbol{Y}_{n\times d}$ such that $d < D$, and $\boldsymbol{Y}_{n\times d}$ approximates the desired space (pointed out graphically by the user) $\widehat{\boldsymbol{Y}}_{n\times d}$. For the proposed of this thesis, the low dimensional space represent with the letter "d" is two.

3.2 Method flowchart

The IDVF works as follows: Considering a high-dimensional (3D for quick test), input data matrix, we calculate a set of kernel matrices. Then, we can estimate a low-dimensional representation and plot an initial lower-dimensional (2D) representation. Over such representation, the user can pick up a data point and decide its final location. Next, the algorithm seeks for the weighting factors or coefficients, which best approximate the desired low-dimensional

representation by following a KPCA-based DR applied over a linear combination of kernel matrices regarding obtained coefficients. This process iterates until the user manually stops. Figure 3.1 depicts the IDVF workflow.

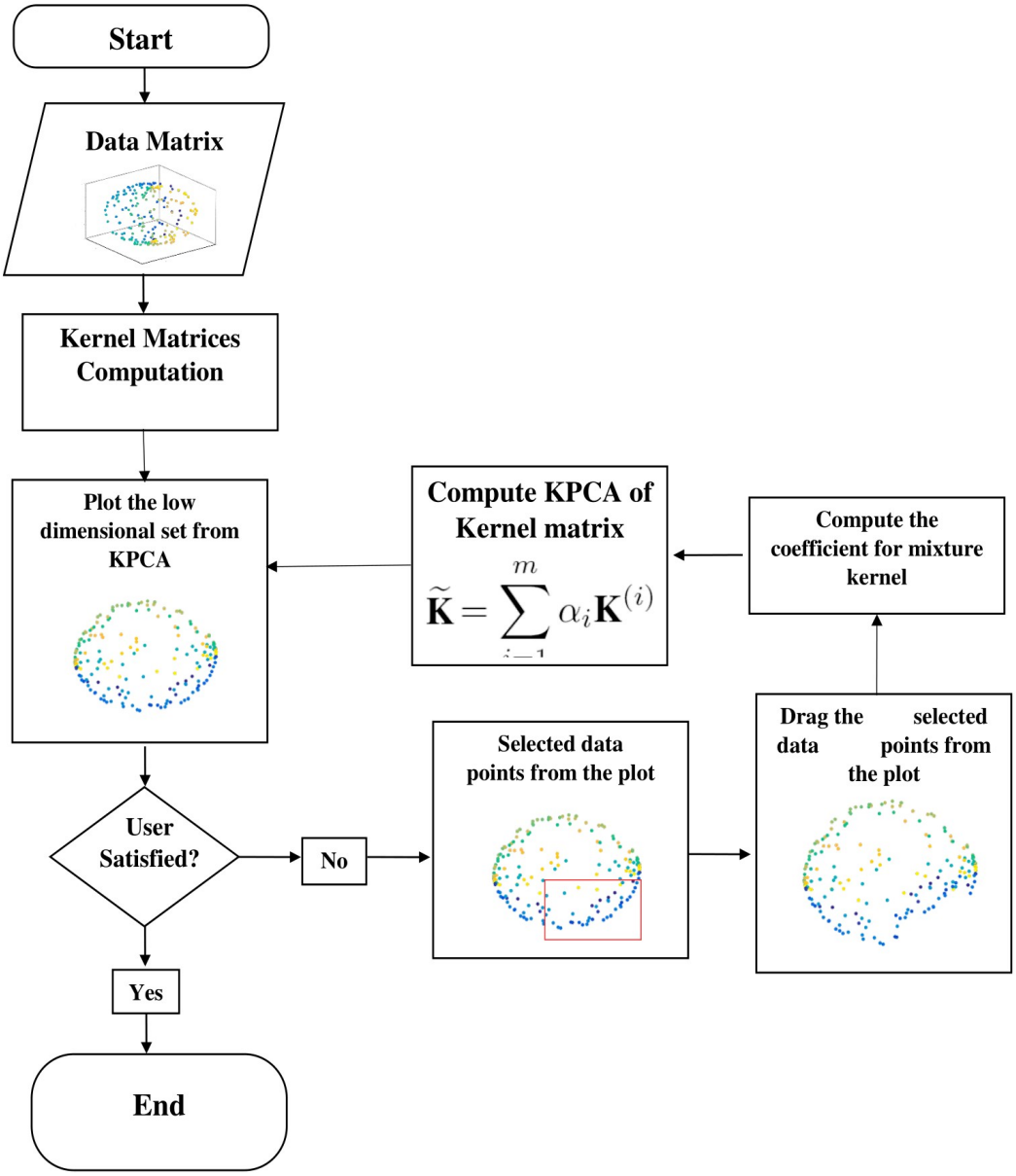


FIGURE 3.1: Proposed IVDF flowchart. It seeks for the best coefficients, which used as weighting factors for a mixture of kernel matrices best represent a desired, low-dimensional space when applying KPCA.

3.2.1 Data Matrix and Kernel Matrices Computation

The data use in this work $\boldsymbol{X}_{n\times D}$ is specified in the section 4.1. Once the data is load, the algorithm proceed to compute their kernel matrices and construct the multidimensional matrix $\boldsymbol{K}^{(m)}_{n\times n}$ as is specified in the section 4.2. Next, the

kernel matrix correspondent to the approximation of CMDS is calculated the d eigenvalues λ and eigenvectors $\boldsymbol{v}$.

3.2.2 Plot the low dimensional set from KPCA

Given some matrix kernel or mixture kernel $\widetilde{\boldsymbol{K}}$ calculate the fast approximation of d maximum eigenvalues λ_i, and eigenvectors $\boldsymbol{v_i}$ to construct d dimensional set of point $\boldsymbol{Y}$, as shown in Eq. 3.1:

$$\text{KPCA}(\boldsymbol{K}(\boldsymbol{X}_{n\times D})) = \boldsymbol{Y}_{n\times d} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d\}, \tag{3.1}$$

where the d terms from $\boldsymbol{Y}$ are constructed by the d eigenvalues and eigenvectors, according to the simplified expression used in [18], as follows:

$$\boldsymbol{y}_j = \sqrt{\lambda_j}\boldsymbol{v}_j. \tag{3.2}$$

3.2.3 Select and drag data points from the plot.

Given two dimensional scatter plot generate by spectral decomposition generate by some kernel matrix. The user thought an interface can edit the position of the points. To do this task, the user must drag his mouse in order to specify the area of point that he wants to move. Once the points are selected, the selected points are manipulate by the user, and they can be moved throw the x and y plane. This process can be repeat until the scatter satisfied the representation expected by the user. Finally, the set of points adulterated by the user are call $\widehat{\boldsymbol{Y}}$, a clear explanation of $\widehat{\boldsymbol{Y}}$ is in the equation 3.6.

3.2.4 Compute the coefficients for the mixture of kernels

Once the user define the low dimensional $\widehat{\boldsymbol{Y}}$ through a graphic interface, we need to estimate coefficients of the mixture matrix $\widetilde{\boldsymbol{K}}$ to subsequently apply $\text{KPCA}(\widetilde{\boldsymbol{K}})$ for producing a matrix $\boldsymbol{Y}$ that best approximates to $\widehat{\boldsymbol{Y}}$. However, building the matrix $\widetilde{\boldsymbol{K}}$ is not a trivial task, and there is no a linear or straightforward relationship between $\widetilde{\boldsymbol{K}}$ and $\widehat{\boldsymbol{Y}}$.

Our proposed solution is to generate a square and symmetric matrix $\widehat{\boldsymbol{M}}$, whose spectral decomposition is similar to $\widehat{\boldsymbol{Y}}$. To achieve this goal, $\boldsymbol{M}$ is defined as the spectral decomposition of $\boldsymbol{Y}$, as can be seen in Eq. 3.3:

$$\boldsymbol{M} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1} = \sum_{i=0}^{D} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^{\top}. \tag{3.3}$$

Notwithstanding, the result of KPCA($\boldsymbol{K}$) is a matrix $\boldsymbol{Y}$ holding d elements. Then, to compute $\boldsymbol{M}$, it is necessary to count on D eigenvectors and eivenvalues. Since, $\boldsymbol{Y}$ is a low-dimensional set of d elements, it is not possible to generate a exact spectral decomposition, as expressed in Eq. 3.3. Alternatively, it is possible to generate an approximation generated by the spectral decomposition $\widehat{\boldsymbol{M}}^{*}$, such that d eigenvalues and eigenvectors are different from zero, and the remaining ones are null, as described in Eq. 3.4:

$$\widehat{\boldsymbol{M}}^{*}_{n\times n} = \sum_{j=1}^{d} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\top} \text{ with } d \leq D. \tag{3.4}$$

Matrix $\widehat{\boldsymbol{M}}$ is the approximation generated by d eigenvectors $\boldsymbol{v}_j$ and d eigenvalues λ_j. At this extent, the original values of $\boldsymbol{Y}$ are explained in the subsection 3.2.2.

Therefore from equations 3.4 and 3.2, an approximation of $\boldsymbol{M}$ can be expressed as:

$$\widehat{\boldsymbol{M}}^{*} = \sum_{j=0}^{d} \sqrt{\lambda_j}\sqrt{\lambda_j} \boldsymbol{v}_j \boldsymbol{v}_j^{\top} = \sum_{j=0}^{d} \boldsymbol{y}_j \boldsymbol{y}_j^{\top} \text{ with } \boldsymbol{v}_j\lambda_j \in \boldsymbol{Y}. \tag{3.5}$$

Since $\widehat{\boldsymbol{Y}}$ is constructed by the user modification of $\boldsymbol{y}_j$, $\widehat{\boldsymbol{y}}_j$ can be expressed as presented in Eq. 3.6:

$$\widehat{\boldsymbol{y}}_j = \sqrt{\hat{\lambda}_j}\widehat{\boldsymbol{v}}_j, \tag{3.6}$$

where $\widehat{\boldsymbol{v}}_j$ is any vector that multiply by a constant $\hat{\lambda}_j$ return the vector $\widehat{\boldsymbol{y}}_j$. $\widehat{\boldsymbol{M}}$ is constructed by applying Eq. 3.5 on Eq. 3.6, so:

$$\widehat{\boldsymbol{M}} = \sum_{j=0}^{d} \sqrt{\hat{\lambda}_j}\sqrt{\hat{\lambda}_j} \widehat{\boldsymbol{v}}_j \widehat{\boldsymbol{v}}_j^{\top} = \sum_{j=0}^{d} \widehat{\boldsymbol{y}}_j \widehat{\boldsymbol{y}}_j^{\top} \text{ with } \widehat{\boldsymbol{y}}_j \in \widehat{\boldsymbol{Y}}. \tag{3.7}$$

Then, $\widehat{\boldsymbol{M}}$ is expected to approximate $\widetilde{\boldsymbol{K}}$, and can be defined as:

$$\operatorname{vec}\left(\widetilde{\boldsymbol{K}}\left(\boldsymbol{X}_{n\times D}\right)\right)=\sum_{i=1}^{m}\alpha_i\boldsymbol{K}^{(i)}\left(\boldsymbol{X}_{n\times D}\right)=\begin{pmatrix}\boldsymbol{K}^{(0)} & \dots & \boldsymbol{K}^{(m)}\end{pmatrix}\begin{pmatrix}\alpha_0\\ \vdots\\ \alpha_m\end{pmatrix}. \tag{3.8}$$

From previous equation, we can appreciate that there exists a linear relationship between $\widehat{\boldsymbol{M}}$ and $\widetilde{\boldsymbol{K}}$ as expressed in Eq. 3.9:

$$\begin{pmatrix}\operatorname{vec}(\boldsymbol{K}^{(0)}) & \dots & \operatorname{vec}(\boldsymbol{K}^{(m)})\end{pmatrix}\begin{pmatrix}\alpha_0\\ \vdots\\ \alpha_m\end{pmatrix}=\operatorname{vec}(\widehat{\boldsymbol{M}}). \tag{3.9}$$

Since $\widehat{\boldsymbol{M}}$ and $\boldsymbol{K}^{(i)}$ are beforehand calculated, the unique unknown variable is $\boldsymbol{\alpha}$. A simplest approach to estimated $\boldsymbol{\alpha}$ is as follows:

$$\begin{pmatrix}\alpha_0\\ \vdots\\ \alpha_m\end{pmatrix}=\begin{pmatrix}\operatorname{vec}(\boldsymbol{K}^{(0)}) & \dots & \operatorname{vec}(\boldsymbol{K}^{(m)})\end{pmatrix}^{+}\operatorname{vec}(\widehat{\boldsymbol{M}}). \tag{3.10}$$

3.2.5 Compute KPCA of Kernel matrix

Once the values of alpha was calculated using as it is shown in equation 2.6, $\widetilde{\boldsymbol{K}}$ is calculated in the following equation:

$$\widetilde{\boldsymbol{K}}=\begin{pmatrix}\boldsymbol{K}^{(0)} & \dots & \boldsymbol{K}^{(m)}\end{pmatrix}\begin{pmatrix}\alpha_0\\ \vdots\\ \alpha_m\end{pmatrix}. \tag{3.11}$$

Thus, KPCA$(\widetilde{\boldsymbol{K}})$ generate a d dimensional set $\widehat{\boldsymbol{Y}}$ being similar to $\boldsymbol{Y}$.

3.2.6 Compute the quality curve

The LD space generate by KCMS, CMDS, the mixture kernel matrix $\widetilde{\boldsymbol{K}}$, and the low dimensional manipulation of the user is compare with the original data, using the curve generate by R_{NX}, which is explained in the section 2.7.

If the K-ary neighbourhood are similar between the LD and HD spaces, the quality curve has higher values of area under the curve. Biggest curves are mostly related with original data. On the other hand, similar curves shows similar distribution of distance among data points.

Chapter 4

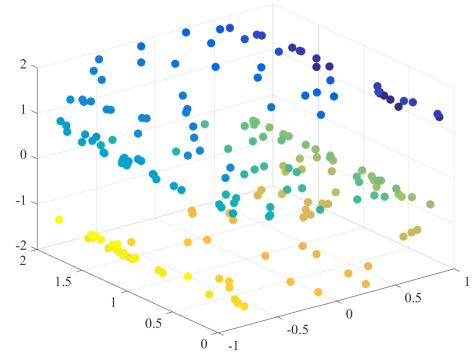
Experimental setup

4.1 Databases

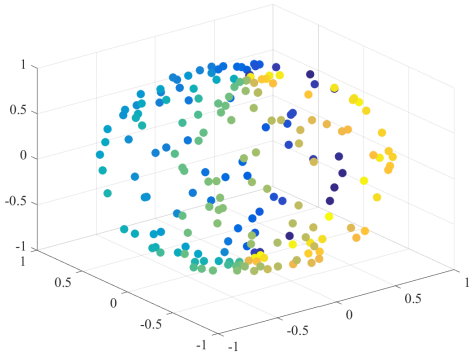
For experiments, the considered data-sets are three toy 3D figures database, they holds some topology and geometry relevant for DR purposes, and they are frequently used in the literature [6], [7], [14]. There are describe:

1. `S 3D`: A surface in form of the uppercase letter S.
2. `Spherical Shell`: A spherical surface.
3. `Swiss Roll`: A Swiss-roll-like manifold.

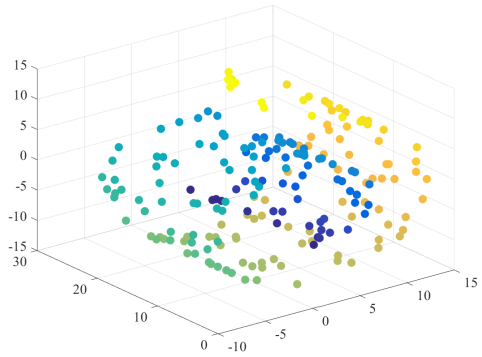
All previous datasets are generated by random points, setting $N = 200$ and $D = 3$. Figure 4.1 shows views of the datasets scatter plots.



(a) `S 3D`



(b) `Spherical Shell`



(c) `Swiss Roll`

FIGURE 4.1: Databases used for the experiment

4.2 Parameter settings and method

The experiment was executed with 100 different kernel matrices $\{\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(100)}\}$, the kernel used in this work are specify in the Section 2.5. To complete the

100 kernels matrices, some of the kernel matrices were obtained by varying the hyper-parameters, as is shown in the Table 4.2

$\boldsymbol{K}^{(100)}$ follow the next distribution:

TABLE 4.1: Kernel Functions and its definitions

Kernel Position	**Kernel Matrices**	**Hyper parameter**
$\boldsymbol{K}^{(1)}$	$\boldsymbol{K}_{\text{CMDS}}$	
$\{\boldsymbol{K}^{(2)}, \ldots, \boldsymbol{K}^{(10)}\}$	$\boldsymbol{K}_{\text{LE}}$	with neighborhood from 2 to 10.
$\{\boldsymbol{K}^{(11)}, \ldots, \boldsymbol{K}^{(19)}\}$	$\boldsymbol{K}_{\text{LLE}}$	with neighborhood from 2 to 10.
$\{\boldsymbol{K}^{(20)}, \ldots, \boldsymbol{K}^{(28)}\}$	$\boldsymbol{K}_{\text{Isomap}}$	with neighborhood from 2 to 10.
$\boldsymbol{K}^{(29)}$	$\boldsymbol{K}_{\text{linear}}$	
$\boldsymbol{K}^{(30)}$	$\boldsymbol{K}_{\text{polynomial}}$	setting the degree as 3.
$\boldsymbol{K}^{(31)}$	$\boldsymbol{K}_{\text{sigmoid}}$	σ=1.
$\boldsymbol{K}^{(32)}$	$\boldsymbol{K}_{\text{laplacian}}$	σ=1 .
$\{\boldsymbol{K}^{(33)}, \ldots, \boldsymbol{K}^{(100)}\}$	$\boldsymbol{K}_{\text{RBF}}$	with γ chosen from $\{0.0001, \ldots, 1.2\}$

TABLE 4.2: Table of Kernel Matrices used for experimental results.

4.3 Quality measure

To validate proposed approach and quantify the results of experiments, we evaluate the curve generate by R_{NX} in the different K-ary neighborhood, which is explained in Sections 3.2.6 and 2.7. Specifically, to quantify the quality of the similarity of embedding representations.

Chapter 5

Analysis of results

The experiments performed in this article seek two aims: The first one is to compare the DR result made by user manipulation with the DR outcomes produced by KPCA using the mixture kernel $\widetilde{K}$. The second one is to contrast the quality curve of the original space of the data with the DR produced by CMDS, KPCA inputted with the CMDS kernel, and the mixture kernel.

Consequently, Chapter 5 is divided into three subsections (5.1, 5.2, and 5.3), one for every database mentioned in subsection 4.1.

Each subsection starts with two plots illustrating the two dimensional representations of the database using CMDS and K_{CMDS} (Figures 5.1(a), 5.8, and 5.15). These results was used to compared the quality curves for the user manipulation of the DR and their approximation using a mixture Kernel.

5.1 Results for S 3D

Figure 5.1(a) shows a two dimensional representation of `S 3D` using CMDS, and the 5.1(b) the a two dimensional representation of `S 3D` using the Kernel approximation of CMDS.

(a) CMDS (b) K_{CMDS}

FIGURE 5.1: Dimentionality Reduction of `S 3D`

Trial 1

Figure 5.2(a) is the Figure 5.1(a) adulterated by the user. The Figure 5.2(b) is an approximation using a computed, mixture kernel matrix based on the representation shown in Figure 5.2(a).

(a) DR manipulation by the user

(b) Approximation from kernel mixture matrix to the user manipulation DR representation

FIGURE 5.2: Experimental results of the first trial for `S 3D`

Figure 5.3 shows similar quality curves in the kernel methods and R_{nx} values, excepting when the value of K is around 10^2 or higher, since at that value the curves are distinguishable from each other.

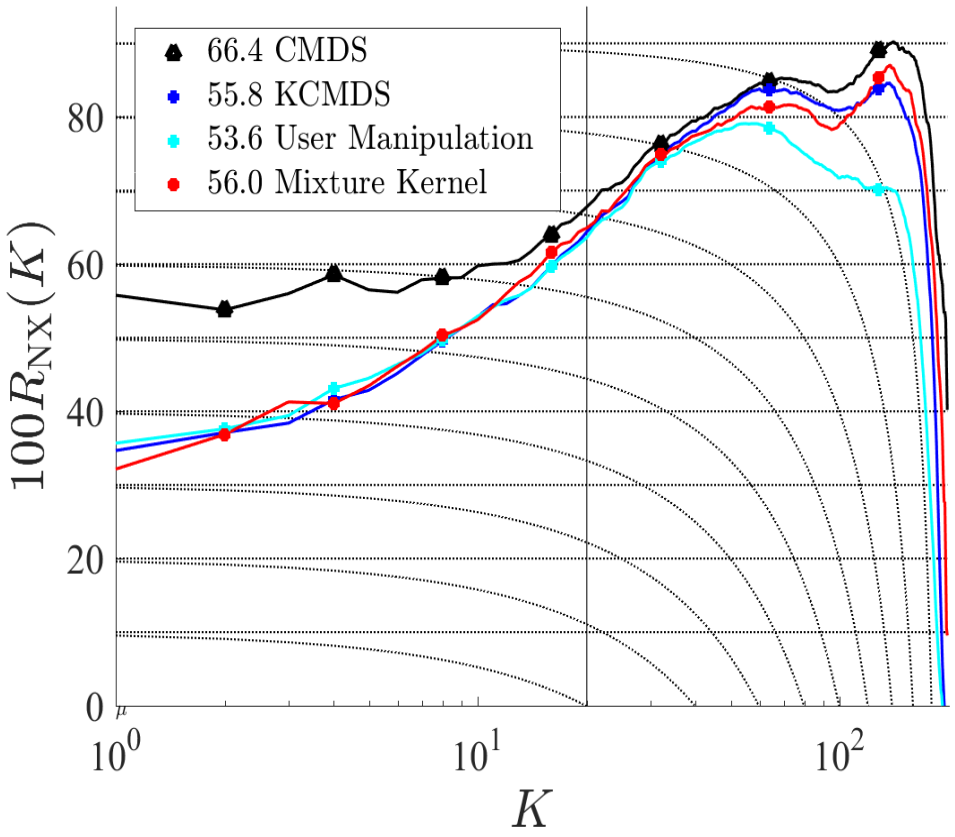
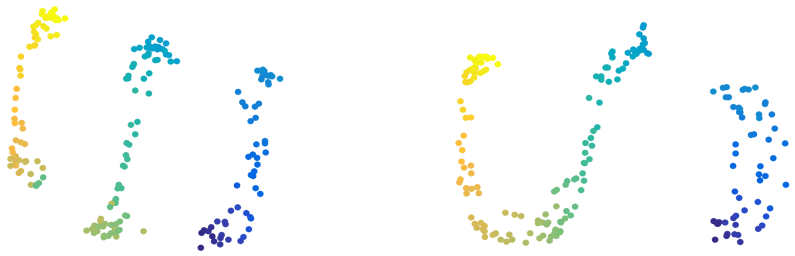


FIGURE 5.3: Quality Curves of CMDS, $\boldsymbol{K}_{\text{CMDS}}$, DR manipulated by the user, and kernel mixture matrix

Trial 2

Figure 5.4(a) is a two-dimensional representation made by the user, in this trial the user attempts to create 3 groups of data points. However, the approximation presented in Figure 5.4(b) reaches seemingly two separate groups, as the two first groups (from left to right) remain visually as one.



(a) DR manipulation by the user

(b) Approximation from kernel mixture matrix to the user manipulation DR representation

FIGURE 5.4: Experimental results of the second trial for `S 3D`

The quality curves of Fig. 5.5 shows that the mixture kernel matrix has an area under the curve higher than the DR manipulate by the user. However, the curves mentioned before seem parallel respect each other.

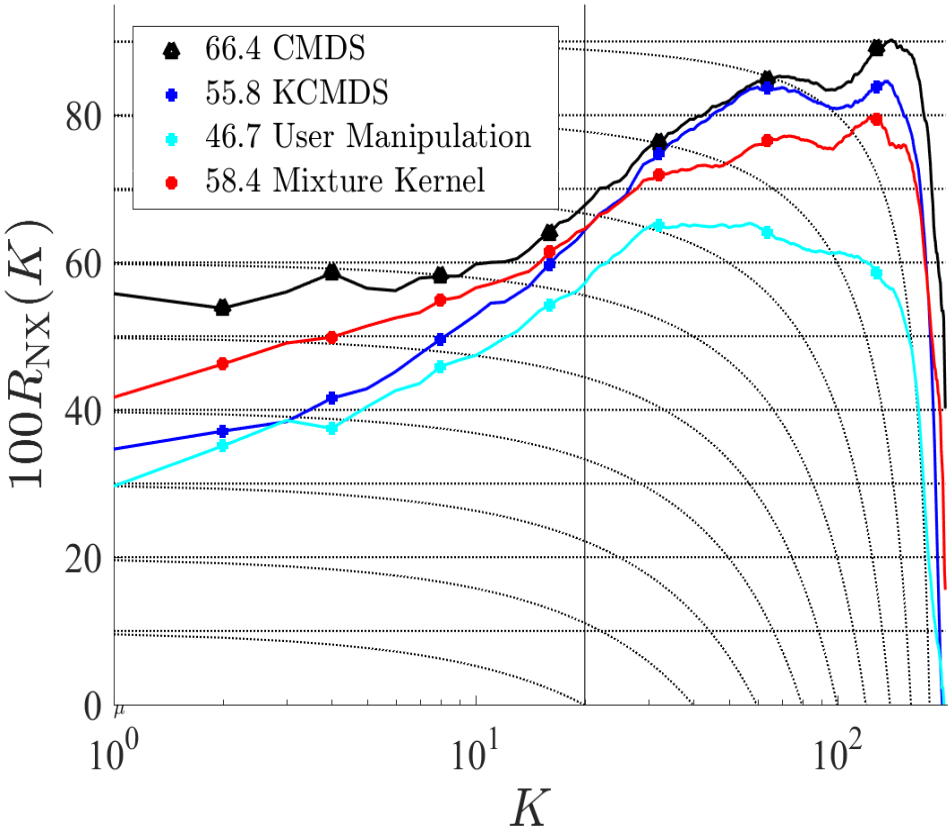
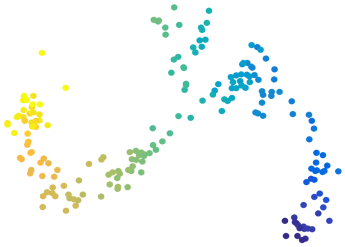


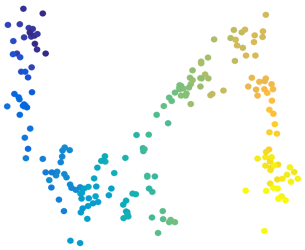
FIGURE 5.5: Quality Curves of CMDS, $\boldsymbol{K}_{\text{CMDS}}$, DR manipulated by the user, and Kernel Mixture Matrix.

Trial 3

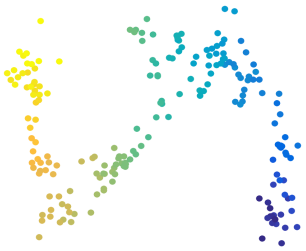
At a first glance, Figures 5.6(a) and 5.6(b) are not resembling. Nonetheless, if Figure 5.6(b) is rotated 180 degrees, the Figures can now be compared visually, and they appear to be similar, as is shown in Figure 5.6(c).



(a) DR manipulation by the user



(b) Approximation from kernel mixture matrix to the user manipulation DR representation



(c) Approximation from kernel mixture matrix to the user manipulation DR representation rotate 180 degrees

FIGURE 5.6: Experimental results of the third trial for `S 3D`

The quality curve of 5.6(a) and 5.6(b) are closer together, however for higher number K the curves tends to separate from each other, and the area under the curve of mixture kernel shown in Figure 5.7 is higher than user manipulation.

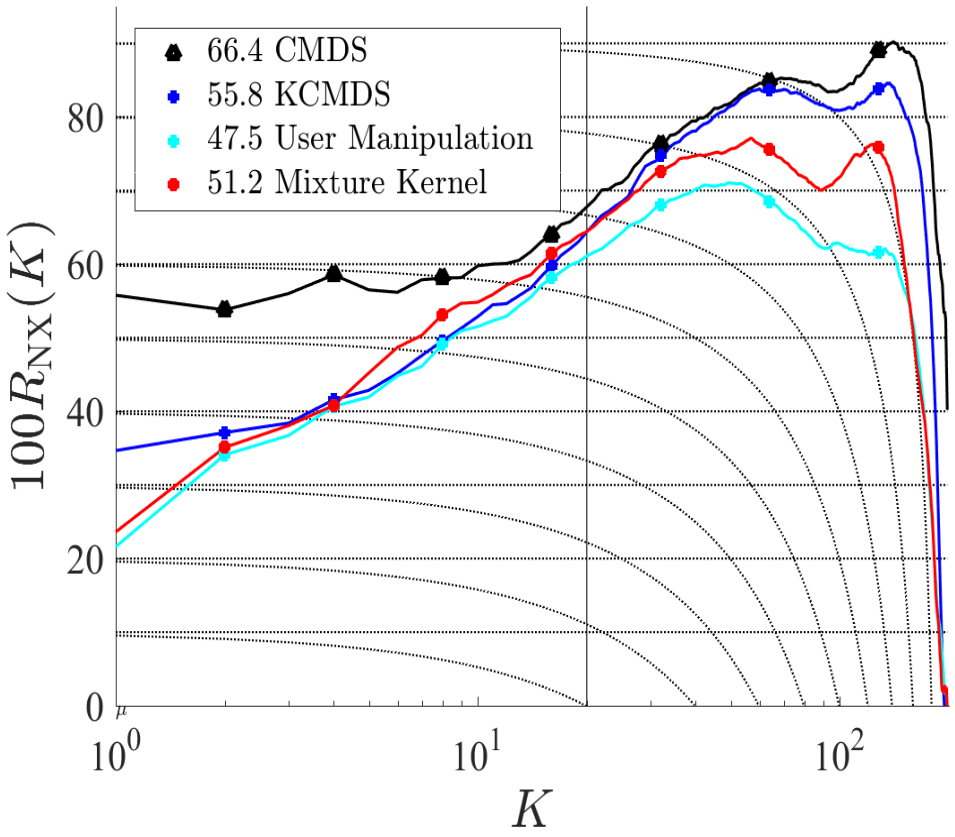
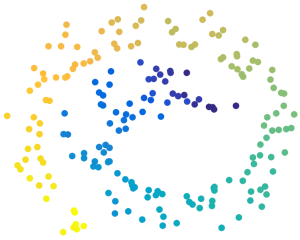


FIGURE 5.7: Quality Curves of CMDS, $\boldsymbol{K}_{\text{CMDS}}$, DR manipulated by the user, and kernel mixture matrix

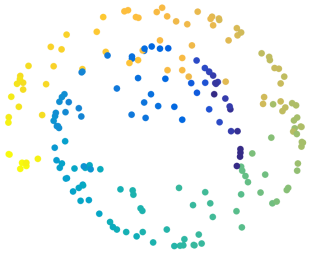
5.2 Results for Swiss Roll

In the case of the `Swiss Roll` database, the DR of the CMDS method and K_{CMDS} are distinguishable from one another. On one hand, once properly rotated 180 degrees, the two-dimensional representations become comparable and look seemingly similar, as is shown in Figure 5.8(c). On the other hand, the gap among points of Figure 5.8(a) is smaller than the that of the points in Figure 5.8(b).



(a) CMDS

(b) K_{CMDS}

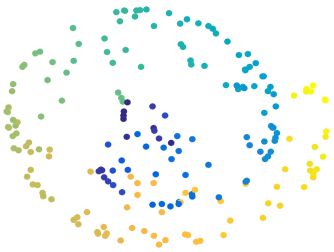


(c) K_{CMDS} rotate 180 degrees

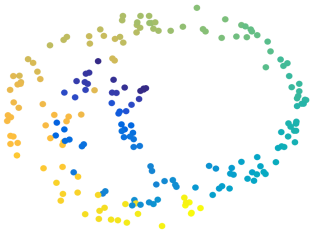
FIGURE 5.8: Dimentionality Reduction of `S 3D`

Trial 1

In the same way that Section 5.1 the user modified the position of certain points of the two dimensional representation of the `Swiss Roll` .



(a) DR manipulation by the user



(b) Approximation from kernel mixture matrix to the user manipulation DR representation

FIGURE 5.9: Experimental results of the first trial for `Swiss Roll`

As result, the quality curve of the mixture kernel matrix, user manipulation, and $\boldsymbol{K}_{\text{CMDS}}$ are similar to each other.

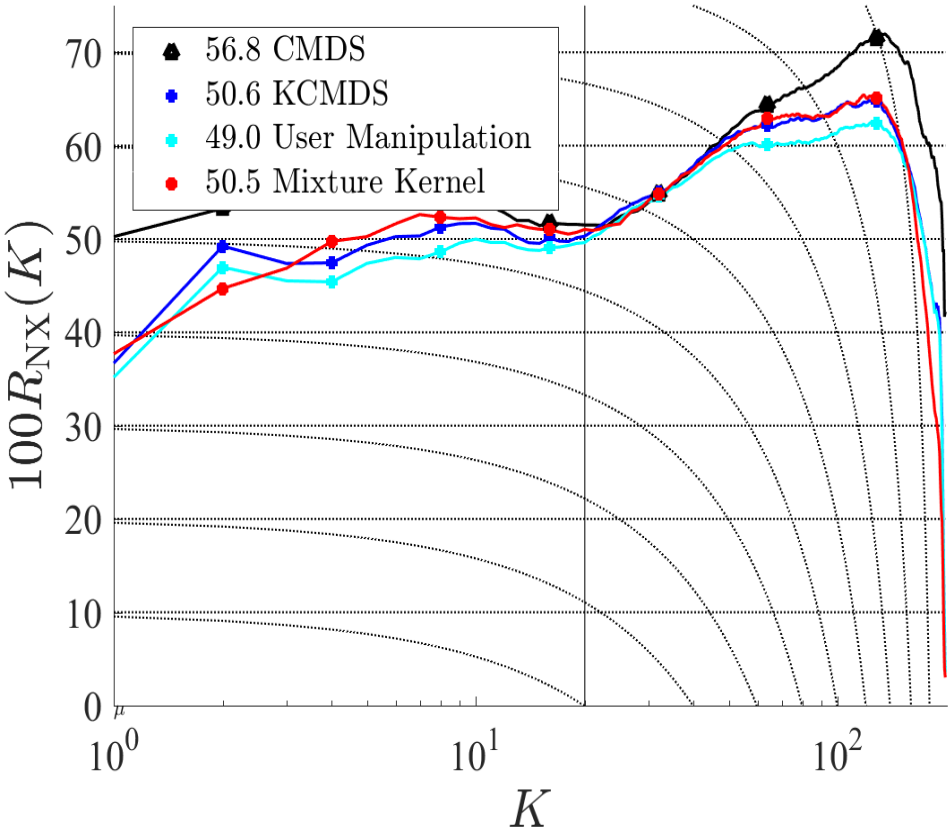
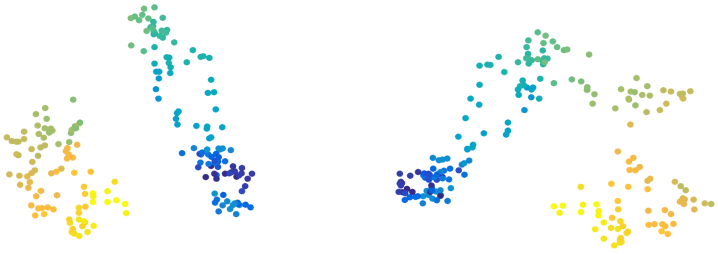


FIGURE 5.10: Quality Curves of the embeddings resulting from CMDS, $\boldsymbol{K}_{\text{CMDS}}$, DR manipulated by the user, and kernel mixture matrix.

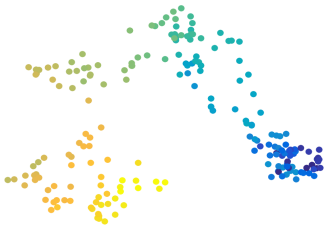
Trial 2

Figure 5.11(b) is an approximation of the 5.11(a) generated by a mixture kernel, however to check their affinity, the Figure 5.11(b) must suffer a reflection in the Y axis and a rotation of 180 degrees, as shown in Figure 5.11(c).



(a) DR manipulation by the user

(b) Approximation from kernel mixture matrix to the user manipulation DR representation



(c) Approximation from kernel mixture matrix to the user manipulation DR representation rotate 90 degrees and reflected in Y axis.

FIGURE 5.11: Experimental results of the second trial for `Swiss Roll`.

The qualities curves of presented in Figure 5.12, shows that the curve generate by mixture kernel and user manipulation are almost equal, however they have low R_{nx} values for high values of neighbors.

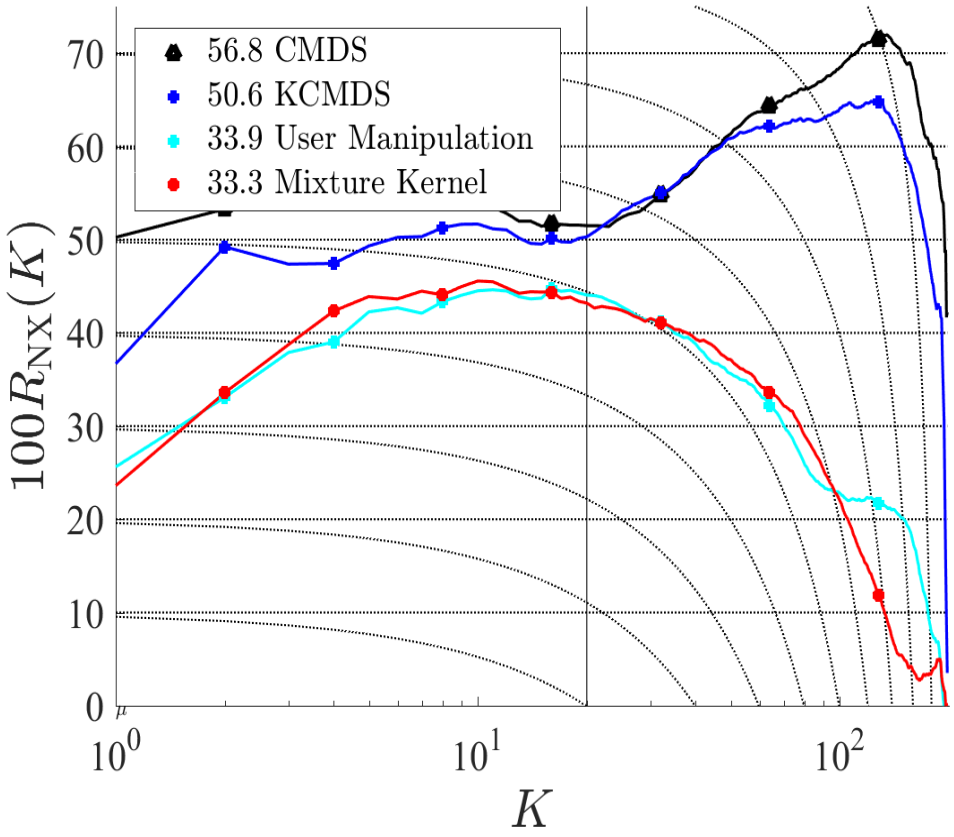
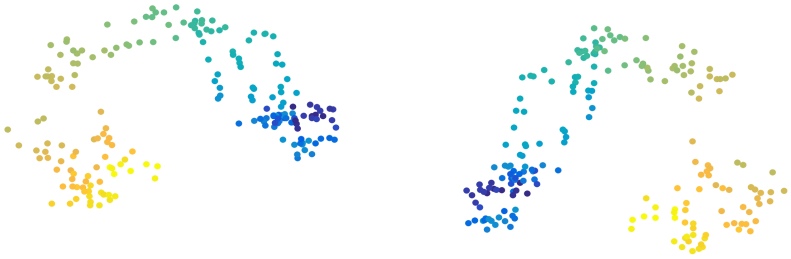


FIGURE 5.12: Quality Curves of CMDS, K_{CMDS}, DR manipulated by the user, and kernel mixture matrix.

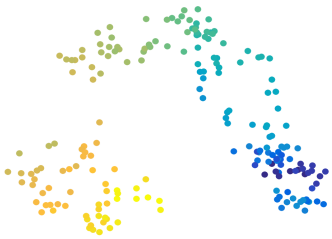
Trial 3

The result of DR approximation shown in Figure 5.13(b) is similar to a reflection in Y axis and rotation 180 degrees of the Figure 5.13(a).



(a) DR manipulation by the user

(b) Approximation from kernel mixture matrix to the user manipulation DR representation



(c) Approximation from kernel mixture matrix to the user manipulation DR representation rotate 90 degrees and reflected in Y axis

FIGURE 5.13: Experimental results of the third trial for `Swiss Roll`

The quality curves of Figures 5.13(a) and 5.13(a) are presented in Figure 5.14. The curves represented are similar. Likely in the above mentioned case, the R_{nx} for a K$>10^1$ are low.

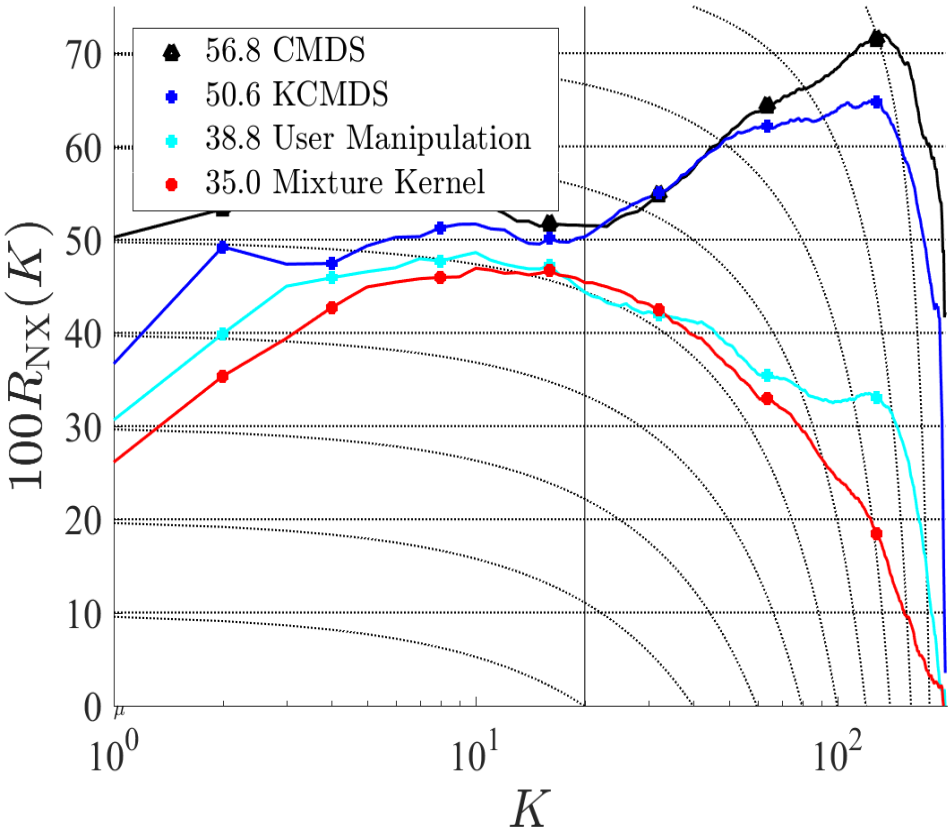
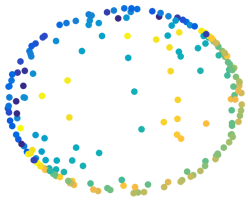


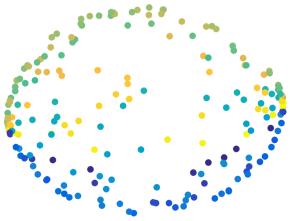
FIGURE 5.14: Quality Curves of CMDS, $\boldsymbol{K}_{\text{CMDS}}$, DR manipulated by the user, and kernel mixture matrix

5.3 Results for Spherical Shell

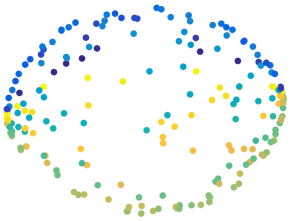
Similar to the Section 5.2, the two dimensional representation between K_{CMDS} and CMDS show in Figure 5.15 have a rotation of 180 degrees from each other and different dispersion among their points.



(a) CMDS



(b) $\boldsymbol{K}_{\text{CMDS}}$

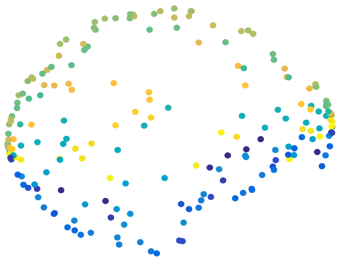


(c) $\boldsymbol{K}_{\text{CMDS}}$ rotate 180 degrees

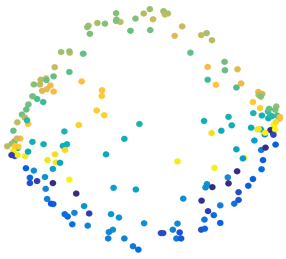
FIGURE 5.15: Dimentionality Reduction of `Spherical Shell`

Trial 1

The figure 5.16(a) is alteration of the K_{CMDS} representation, where some points from the right corner were dragged to the center. However, the approximation result shown in figure 5.16(b) has the upper and button corner points of the figure closely to the center of the image.



(a) DR manipulation by the user



(b) Approximation from kernel mixture matrix to the user manipulation DR representation

FIGURE 5.16: Experimental results of the first trial for `Spherical Shell`

The qualities curves of the figure 5.17 are closer from each other.

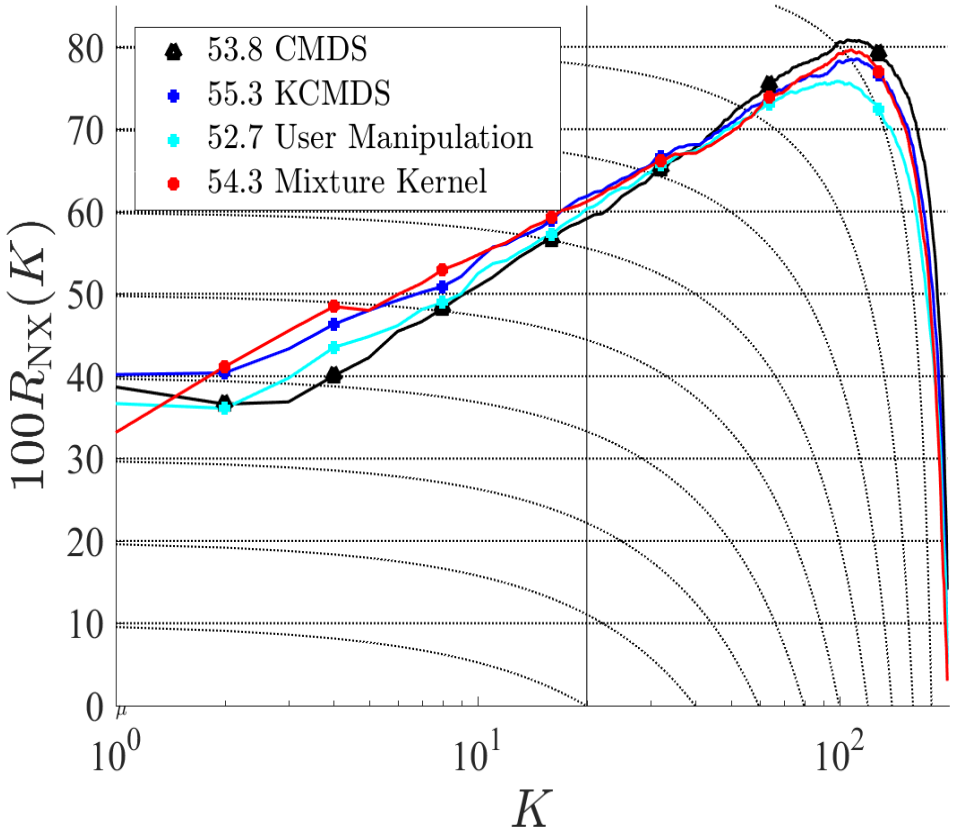
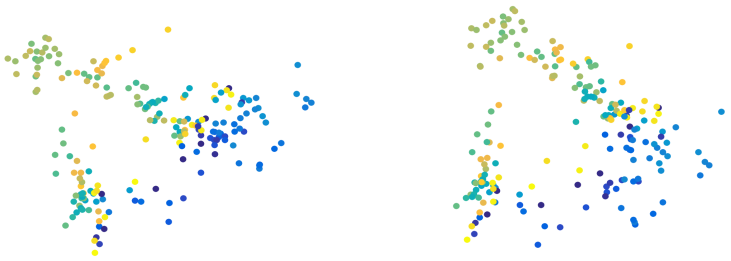


FIGURE 5.17: Quality Curves of CMDS, $\boldsymbol{K}_{\mathrm{CMDS}}$, DR manipulated by the user, and kernel mixture matrix

Trial 2

In the figure 5.18(a) presented an 2D representation made by user which not follow a circular shape, the approximation represented by figure 5.18(b) in the same way represent a non circular shape.



(a) DR manipulation by the user

(b) Approximation from kernel mixture matrix to the user manipulation DR representation

FIGURE 5.18: Experimental results of the second trial for `Spherical Shell`

From Figure 5.19 can be appreciated that the quality curve of Figures 5.18(a) and 5.18(b) are similar and they exhibit the smallest values of area under the curve R_{nx} than the other the curves.

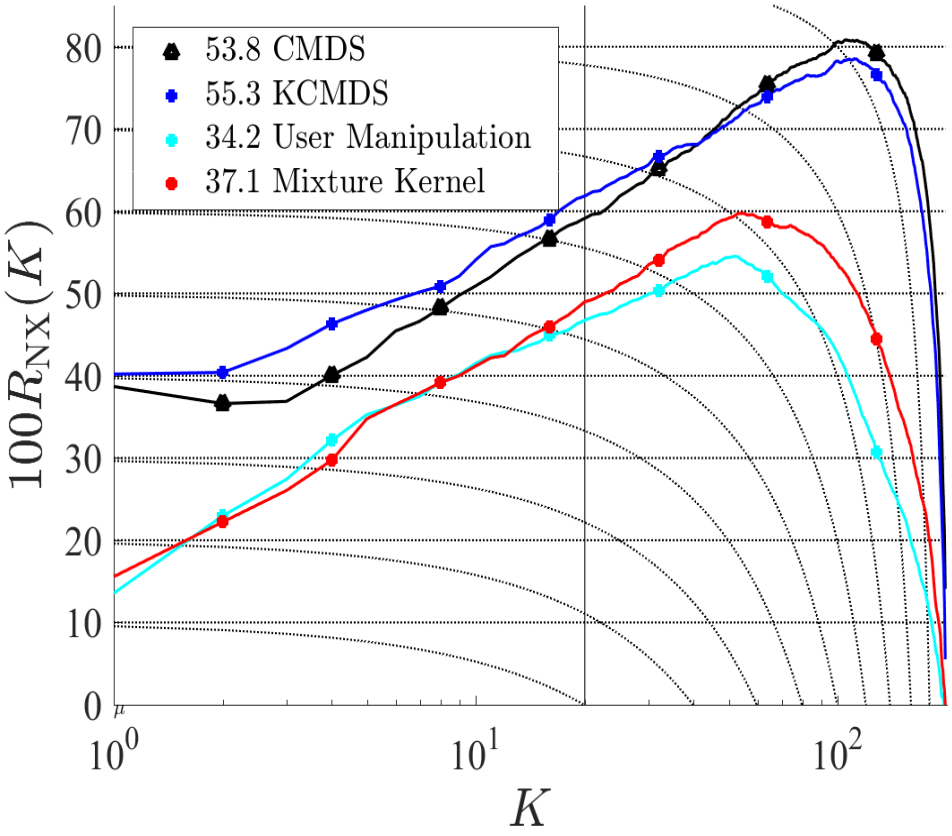
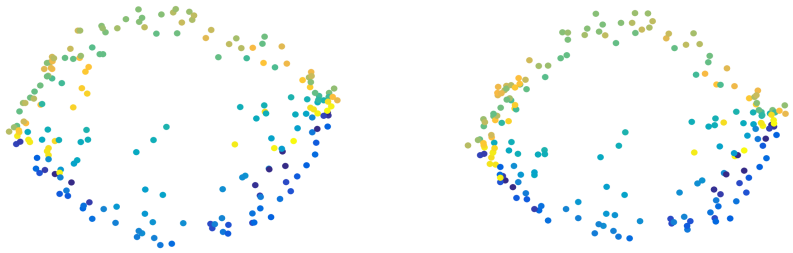


FIGURE 5.19: Quality Curves of CMDS, K_{CMDS}, DR manipulated by the user, and kernel mixture matrix

Trial 3

In the figure 5.20(a) presented an 2D representation made by user which is similar to a circle with less density of points in the center, the approximation represented by figure 5.18(b) in the same way is similar to a circle with less density of points in the center, moreover the upper corner is squashed to the center.



(a) DR manipulation by the user

(b) Approximation from kernel mixture matrix to the user manipulation DR representation

FIGURE 5.20: Experimental results of the third trial for `Spherical Shell`

The figure 5.21 show that the quality curve of figure 5.20(a) and 5.20(b) are similar and they have smaller values of R_{nx} than the other the curves.

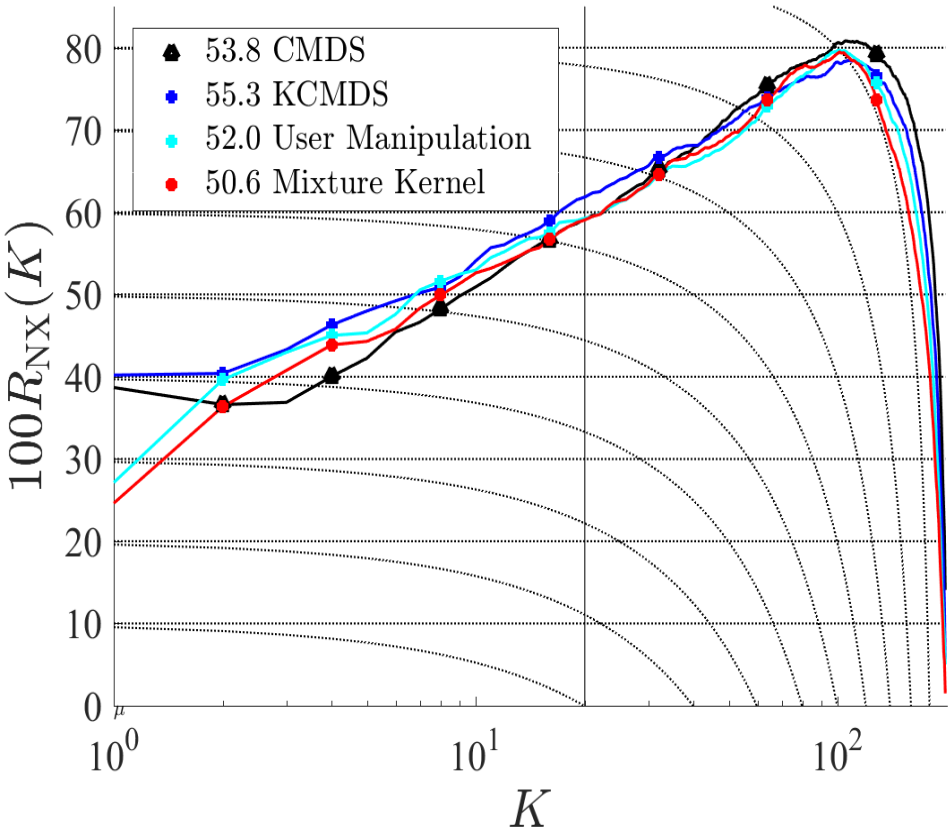


FIGURE 5.21: Quality Curves of CMDS, K_{CMDS}, DR manipulated by the user, and kernel mixture matrix

Chapter 6

Conclusion and Future work

6.1 Conclusions

In this work, we introduce the Interactive Data Visualization Framework (IDVF), which opens the possibility to formally develop new interactive data visualization approach based on mixtures of dimensional reduction (DR) techniques. Our IDVF allows for readily incorporating the users' knowledge and expertise into the data exploration and visualization processes. What makes this approach appealing and essentially different from other similar works is the fact that the user can directly handle data and dynamically accomplish a new representation. IDVF seeks to best fit the user's data interpretation and accordingly find the best combination of kernels representing DR methods. Although it counts on already promising results, this thesis is still a first approach to produce IDVF. Indeed, some of the two-dimensional approximations are not resembling to those made by the user. Moreover, in some cases, the results differ from the user expectations as well as their quality curves exhibit different behaviors. It is caused by many factors, such as: having not enough number of kernels to avoid the resulting ill-posed problem, KPCA outcomes reaches a no suitable representation with the current kernels matrices, and a poor approximation of the matrix required to estimate the eigenvalues. Notwithstanding, in many occasions, there was possible to successfully reach a similar representation of the user dimensional reduction data, as can be observed from the experimental results.

In terms of visualization, the representation quality of two-dimensional scatter plots depends on the expertise or prior knowledge about data that users may hold. As a matter of fact, because the user directly manipulates the data points themselves, not optimal resulting embedding spaces can be yielded when user-provided, desired embedding is not meaningful. As a consequence, the corresponding quality curves will produce poor values of area under the curve.

6.2 Future Works

For future works, we are aimed at exploring the possibility of developing novel kernel representations arising from other dimensional reduction methods, improves in the equations system and the approximation of eigenvalues, as well as an inverse framework more robust to different data point variations and datasets. In addition, further developing of GUI for a top-notch user experience is also to be explored, which should be able to deal with dimensionality reduction of three-dimensional spaces.

References

[1] D. H. Peluffo Ordoñez, J. A. Lee, and M. Verleysen, "Recent methods for dimensionality reduction: A brief comparative analysis", in *2014 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014)*, 2014.

[2] D. H. Peluffo-Ordóñez, A. E. Castro-Ospina, J. C. Alvarado-Pérez, and E. J. Revelo-Fuelagán, "Multiple kernel learning for spectral dimensionality reduction", in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, A. Pardo and J. Kittler, Eds., Cham: Springer International Publishing, 2015, pp. 626–634.

[3] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade", *IEEE transactions on visualization and computer graphics*, vol. 23, no. 3, pp. 1249–1268, 2016.

[4] J. A. Salazar-Castro, Y. C. Rosas-Narváez, A. D. Pantoja, J. C. Alvarado-Pérez, and D. H. Peluffo-Ordóñez, "Interactive interface for efficient data visualization via a geometric approach", in *2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)*, IEEE, 2015, pp. 1–6.

[5] P. Rosero-Montalvo, P. Diaz, J. A. Salazar-Castro, D. F. Peña-Unigarro, A. J. Anaya-Isaza, J. C. Alvarado-Pérez, R. Therón, and D. H. Peluffo-Ordóñez, "Interactive data visualization using dimensionality reduction and similarity-based representations", in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, C. Beltrán-Castañón, I. Nyström, and F. Famili, Eds., Cham: Springer International Publishing, 2017, pp. 334–342, ISBN: 978-3-319-52277-7.

[6] D. F. Peña-ünigarro, J. A. Salazar-Castro, D. H. Peluffo-Ordóñez, P. D. Rosero-Montalvo, O. R. Oña-Rocha, A. A. Isaza, J. C. Alvarado-Perez, and R. Theron, "Interactive visualization methodology of high-dimensional data with a color-based model for dimensionality reduction", in *2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, 2016, pp. 1–7.

[7] P. D. Rosero-Montalvo, D. F. Peña-Unigarro, D. H. Peluffo, J. A. Castro-Silva, A. Umaquinga, and E. A. Rosero-Rosero, "Data visualization using interactive dimensionality reduction and improved color-based interaction model", in *Biomedical Applications Based on Natural and Artificial Computing*, J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, J. Toledo Moreo, and H. Adeli, Eds., Cham: Springer International Publishing, 2017, pp. 289–298.

[8] J. A. Salazar-Castro, D. Peña-Unigarro, D. H. Peluffo-Ordóñez, P. D. Rosero-Montalvo, H. M. Domínguez-Limaico, J. C. Alvarado-Pérez, and R. Therón, "Dimensionality reduction for interactive data visualization via a geo-desic approach", in *Computational Intelligence (LA-CCI), 2016 IEEE Latin American Conference on*, IEEE, 2016, pp. 1–6.

[9] A. C. Umaquinga-Criollo, D. H. Peluffo-Ordóñez, P. D. Rosero-Montalvo, P. E. Godoy-Trujillo, and H. Benítez-Pereira, "Interactive visualization interfaces for big data analysis using combination of dimensionality reduction methods: A brief review", in *Technology, Sustainability and Educational Innovation (TSIE)*, A. Basantes-Andrade, M. Naranjo-Toro, M. Zambrano Vizuete, and M. Botto-Tobar, Eds., Cham: Springer International Publishing, 2020, pp. 193–203, ISBN: 978-3-030-37221-7.

[10] K. Q. Weinberger, F. Sha, and L. K. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction", in *Proceedings of the twenty-first international conference on Machine learning*, ACM, 2004, p. 106.

[11] H. Choi and S. Choi, "Kernel isomap", English, *Electronics Letters*, vol. 40, 1612–1613(1), 25 2004, ISSN: 0013-5194. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/el_20046791.

[12] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds", in *Proceedings of the twenty-first international conference on Machine learning*, ACM, 2004, p. 47.

[13] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative", *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[14] J. A. Salazar-Castro, P. D. Rosero-Montalvo, D. F. Peña-Unigarro, A. C. Umaquinga-Criollo, Z. Castillo-Marrero, E. J. Revelo-Fuelagán, D. H. Peluffo-Ordóñez, and C. G. Castellanos-Domínguez, "A novel color-based data visualization approach using a circular interaction model

and dimensionality reduction", in *Advances in Neural Networks – ISNN 2018*, T. Huang, J. Lv, C. Sun, and A. V. Tuzikov, Eds., Cham: Springer International Publishing, 2018, pp. 557–567.

[15] D. H. Peluffo-Ordóñez, J. C. Alvarado-Pérez, J. A. Lee, M. Verleysen, *et al.*, "Geometrical homotopy for data visualization.", in *ESANN*, 2015.

[16] Y. Bahroun and A. Soltoggio, "Online representation learning with single and multi-layer hebbian networks for image classification", in *Artificial Neural Networks and Machine Learning – ICANN 2017*, A. Lintas, S. Rovetta, P. F. Verschure, and A. E. Villa, Eds., Cham: Springer International Publishing, 2017, pp. 354–363, ISBN: 978-3-319-68600-4.

[17] A. J. Anaya Isaza, "Metodología de visualización de datos utilizando métodos espectrales y basados en divergencias para la reducción interactiva de la dimensión", PhD thesis, 2018.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering", in *Advances in neural information processing systems*, 2002, pp. 585–591.

[20] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000, ISSN: 0036-8075. DOI: 10.1126/science.290.5500.2323. eprint: https://science.sciencemag.org/content/290/5500/2323.full.pdf. [Online]. Available: https://science.sciencemag.org/content/290/5500/2323.

[21] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel pca and de-noising in feature spaces", in *Advances in neural information processing systems*, 1999, pp. 536–542.

[22] Y. Washizawa, "Subset basis approximation of kernel principal component analysis", *Principal Component Analysis*, p. 67, 2012.

[23] Y. Bengio, P. Vincent, J.-F. Paiement, O Delalleau, M Ouimet, and N LeRoux, "Learning eigenfunctions of similarity: Linking spectral clustering and kernel pca", Technical Report 1232, Departement d'Informatique et Recherche Oprationnelle . . ., Tech. Rep., 2003.

[24] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen, "Generalized kernel framework for unsupervised spectral methods of dimensionality reduction", in *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*, IEEE, 2014, pp. 171–177.

[25] L. A. Belanche Muñoz, "Developments in kernel design", in *ESANN 2013 proceedings: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning: Bruges (Belgium), 24-26 April 2013*, 2013, pp. 369–378.

[26] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering", in *Advances in neural information processing systems*, 2002, pp. 585–591.

[27] ——, "Laplacian eigenmaps for dimensionality reduction and data representation", *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[28] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction", *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[29] H. Choi and S. Choi, "Robust kernel isomap", *Pattern Recognition*, vol. 40, no. 3, pp. 853 –862, 2007, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2006.04.025. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320306001804.

[30] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[31] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data", *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003.

[32] D. DeCoste, "Visualizing mercer kernel feature spaces via kernelized locally-linear embeddings", 2001.

[33] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming", *Journal of Machine learning research*, vol. 5, no. Jan, pp. 27–72, 2004.

[34] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York, NY: Springer, 2006, Softcover published in 2016. [Online]. Available: https://cds.cern.ch/record/998831.

[35] M. C. Ortega-Bustamante, W. Hasperué, D. H. Peluffo-Ordóñez, I. M.-R. M. Paéz-Jaime, P. Rosero-Montalvo, A. C. Umaquinga-Criollo, and M. Vélez-Falconi, "Introducing the concept of interaction model for interactive dimensionality reduction and data visualization", in *20th International Conference on Computational Science and its Applications*, Springer International Publishing, 2020.

[36] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria", *Neurocomputing*, vol. 72, no. 7-9, pp. 1431–1443, 2009.

[37] B. Mokbel, W. Lueks, A. Gisbrecht, and B. Hammer, "Visualizing the quality of dimensionality reduction", *Neurocomputing*, vol. 112, pp. 109–123, 2013.

[38] J. A. Lee and M. Verleysen, "Scale-independent quality criteria for dimensionality reduction", *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2248–2257, 2010.

[39] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen, "Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation", *Neurocomputing*, vol. 112, pp. 92–108, 2013.

Appendix A

Interactive model and kernel matrices

A.1 Interface

Fig. A.1 shows a view of an interface for proposed IDVF. The left-scatter-plot can be manipulated by the user to set the desired low-dimensional space, while the right-one depicts the obtained representation given by the mixture of kernels. Find a demo at https://sdas-group.com/gallery/.

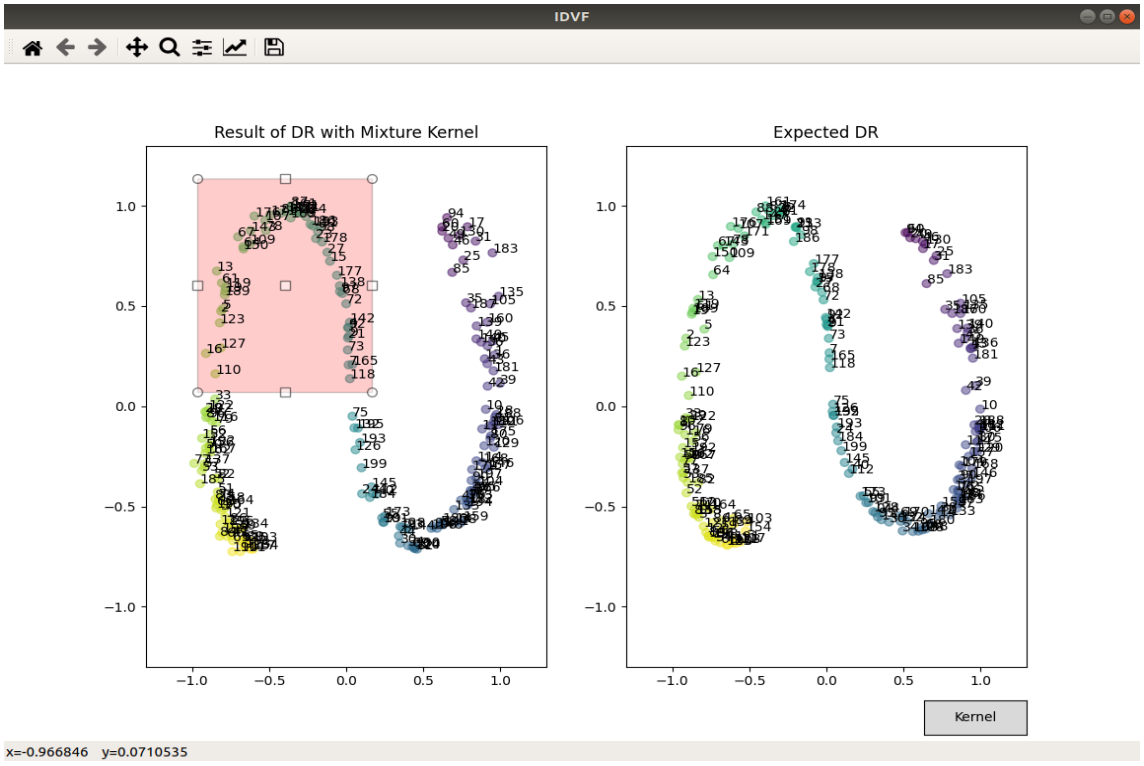


FIGURE A.1: View of the IDVF interface. Both the scatter plot of the desired and the obtained low-dimensional representations are displayed.

A.2 Python Implementation of the interactive model and kernels matrices

In this appendix are reference the code use to implement the flow chart figure 3.1, it contains the implementation of the different Kernel models explain in the section 2.5. For the complete code you can visit the github repository:

https://github.com/martinvelezf/Inverse-Data-Visualization-Framework-IDVF-Towards-a-prior-knowledge-driven-datavisualization

Appendix B

Academic products

B.1 Conference Papers

1. M. C. Ortega-Bustamante and W. Hasperué and D. H. Peluffo-Ordóñez and M. Paéz-Jaime, I. Marrufo-Rodríguez and P. Rosero-Montalvo and A. C. Umaquinga-Criollo and M. Vélez-Falconi *"Introducing the concept of interaction model for interactive dimensionality reduction and data visualization"* In:International Conference on Computational Science and its Applications 2020 `Accepted`

2. M. Vélez-Falconí, J. González-Vergara, D. H. Peluffo-Ordóñez *"Inverse Data Visualization Framework (IDVF): Towardsa prior-knowledge-driven data visualization"* In:International Conference on Applied Informatics 2020 `Accepted`