UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

CHARACTERIZATION STUDY OF METRICS FOR DATA SOURCE STABILITY IN BIOMEDICAL REPOSITORIES

AUTHOR: PABLO FERRI BORREDÀ

SUPERVISOR: JUAN MIGUEL GARCÍA GÓMEZ

SUPERVISOR: CARLOS SÁEZ SILVESTRE

Academic year: 2016-17

“Engineering is the art of trade-offs” (Anonymous)

Agradecimientos

Agraïments

En primer lugar, me gustaría agradecer este trabajo fin de grado a mis tutores, el Dr. Carlos Sáez Silvestre y el Dr. Juan Miguel García Gómez, por las enseñanzas transmitidas en todos los niveles, tanto el teórico como el pragmático, así como a nivel personal. Agradezco asimismo su confianza y apoyo omnipresente en cada una de las fases de este proyecto.

En segundo lugar, debo mostrar mi gratitud hacia el programa Generación Espontánea, el cual me brindó la oportunidad de conocer de primera mano el día a día en el grupo de investigación IBIME, iniciando un proceso de colaboración que se ha extendido durante casi dos años, y que espero que prosiga en el futuro.

Finalment, m'agradaria agrair especialment aquest projecte als meus pares, Amparo i José, sense els quals cap meta acadèmica o personal en la meua vida haguera sigut possible. Ells han aconseguit transmetre'm la seua constància i esforç, així com la il·lusió per continuar aprenent cada dia.

Abstract

The degree of homogeneity among data distributions is a critical issue when reusing data integrated from different sources, since the introduction of undesired variabilities may lead to misleading results. Therefore, assessing this data source variability is of utmost importance to ensure a confident data reuse.

In the biomedical field, this issue is even more critical, due to the risk of serious consequences if data is used improperly. Furthermore, in spite of the existence of classical statistical tools which attempt to deal with this task, there are still several aspects to be addressed due to inherent complexity of biomedical data, such as the intrinsic multimodality of data distributions.

New statistical metrics have been recently developed to overcome this challenge, including the Global probabilistic deviation (GPD) and the Source probabilistic outlyingness (SPO). These metrics allow a multivariate analysis of data source variability without assuming any underlying distribution and without being restricted to certain types of data.

However, when implementing them, users must decide among different options related to data preprocessing techniques, as data representation, reduction and normalization. This fact is not a limitation, but needs to be addressed. In this work, an in-depth study of the influence of preprocessing techniques over the multi-source variability metrics is performed, and procedures to overcome the problematic phenomena encountered are proposed and evaluated.

Once understood the influence of the different settings, the potential biases introduced by pre-set factors to the GPD and SPO metrics, such as the number of sources or the number of data, are evaluated. Results of this assessment suggest the robustness of GPD and SPO to these factors.

Finally, new procedures are proposed to find patterns in multi-source biomedical repositories and offer suggestions for data reuse based on the different data source variability structures. A new clustering algorithm for detecting data source variability patterns is proposed, and its evaluation over multi-source biomedical repositories has led to the discover of four main stability patterns: the Global stability pattern (GSP), the Local stability pattern (LSP), the Sparse stability pattern (SSP) and the Instability pattern (IP). These new procedures increase the added value of the multi-source variability framework for biomedical data characterization.

Keywords: *data science, data quality, data variability, integrated data repositories, density estimation, curse of dimensionality, multiple regression, clustering analysis.*

Resum

El grau d’homogeneïtat entre distribucions de dades és una circumstància crítica quan es reutilitzen dades provinents de diferents fonts, ja que la introducció de variabilitats no desitjades podria conduir a resultats enganyosos. Per tant, avaluar aquesta variabilitat produïda per la font d’on provenen les dades esdevé clau de cara a assegurar una reutilització segura d’aquestes.

En l’àmbit biomèdic, aquest problema és encara més important, a causa del risc de conseqüències greus si les dades son utilitzades de forma inadequada. A més, a pesar de l’existència d’eines estadístiques clàssiques que intentar afrontar aquesta tasca, encara existeixen diversos aspectes que requereixen ser tractats i que són fruit de la inherent complexitat de les dades biomèdiques.

Noves mètriques estadístiques han sigut recentment desenvolupades per a afrontar aquest repte, incloent la Global probabilistic deviation (GPD) i la Source probabilistic outlyingness (SPO). Aquestes mètriques permeten una anàlisi multivariant de la variabilitat de la font de dades sense assumir cap distribució subjacent i sense estar restringides a certs tipus de dades.

Tanmateix, a l’hora d’implementar-les, els usuaris han de decidir entre diferents opcions relacionades amb tècniques de preprocessament. Aquesta circumstància no suposa una limitació, però requereix ser tractada. En aquest treball, es du a terme un estudi en profunditat de la influència de les tècniques de preprocessament sobre les mètriques de variabilitat multi font, i es proposen procediments per a superar els fenòmens adversos trobats.

Una vegada entesa la influència de les diferents configuracions, s'avaluen els potencials biaixos introduïts per factors preestablerts, sobre la GPD i la SPO, com ara el nombre de fonts o el nombre de dades. Els resultats d'aquesta anàlisis suggereixen la robustesa de GPD i SPO front aquests factors.

Finalment, es proposen nous procediments per a trobar patrons en repositoris biomèdics multi font, així com suggeriments per a la reutilització de dades en funció de les diferents estructures de variabilitat multi font encontrades. Es presenta un nou algorisme d’anàlisi clúster per a detectar patrons de variabilitat multi font, i la seva avaluació sobre repositoris biomèdics multi font ha permès descobrir quatre patrons d'estabilitat principals: el Global stability pattern (GSP), el Local stability pattern (LSP), el Sparse stability pattern (SSP) i el Instability pattern (IP). Aquests nous procediments augmenten el valor afegit del marc teòric de variabilitat multi font per a la caracterització de dades biomèdiques.

Paraules clau: *ciència de dades, qualitat de dades, variabilitat de dades, repositoris de dades integrats, estimació de densitat, maledicció de la dimensionalitat, regressió múltiple, anàlisi clúster.*

Resumen

El grado de homogeneidad entre distribuciones de datos es una circunstancia crítica cuando se reutilizan datos provenientes de diferentes fuentes, ya que la introducción de variabilidades no deseadas podría conducir a resultados engañosos. Por lo tanto, evaluar esta variabilidad producida por la fuente de donde provienen los datos es clave de cara a asegurar una reutilización segura de los mismos.

En el ámbito biomédico, este problema es aún más importante, debido al riesgo de consecuencias graves si los datos son utilizados de forma inadecuada. Además, a pesar de la existencia de herramientas estadísticas clásicas que intentar afrontar esta tarea, todavía existen varios aspectos que requieren ser tratados, fruto de la inherente complejidad de los datos biomédicos.

Nuevas métricas estadísticas han sido recientemente desarrolladas para afrontar este reto, incluyendo la Global probabilistic deviation (GPD) y la Source probabilistic outlyingness (SPO). Estas métricas permiten un análisis multivariante de la variabilidad introducida por la fuente de datos sin asumir ninguna distribución subyacente y sin estar restringidas a ciertos tipos de datos.

Sin embargo, a la hora de implementarlas, los usuarios tienen que decidir entre diferentes opciones relacionadas con técnicas de preprocesamiento. Esta circunstancia no supone una limitación, pero debe ser abordada. En este trabajo, se lleva a cabo un estudio en profundidad de la influencia de las técnicas de preprocesamiento sobre las métricas de variabilidad multi fuente, y se proponen procedimientos para superar los fenómenos adversos encontrados.

Una vez entendida la influencia de las diferentes configuraciones, se evalúan los potenciales sesgos introducidos por factores preestablecidos, sobre la GPD y la SPO, tales como el número de fuentes o el número de datos. Los resultados de este análisis sugieren la robustez de GPD y SPO frente estos factores.

Finalmente, se proponen nuevos procedimientos para encontrar patrones en repositorios biomédicos multi fuente, así como sugerencias para la reutilización de datos en función de las diferentes estructuras de variabilidad multi fuente encontradas. Se presenta un nuevo algoritmo de clustering para la detección de patrones de variabilidad multi fuente, y su evaluación sobre repositorios biomédicos multi fuente ha permitido descubrir cuatro patrones de estabilidad principales: el Global stability pattern (GSP), el Local stability pattern (LSP), el Sparse stability pattern (SSP) y el Instability pattern (IP). Estos nuevos procedimientos aumentan el valor añadido del marco teórico de variabilidad multi fuente para la caracterización de datos biomédicos.

Palabras clave: *ciencia de datos, calidad de datos, variabilidad de datos, repositorios de datos integrados, estimación de densidad, maldición de la dimensionalidad, regresión múltiple, clustering.*

Glossary

Acronyms

AC Autoencoder

DBSCAN Density-based spatial clustering of applications with noise

GPD Global probabilistic deviation

GSP Global stability pattern

HFA Hill finder algorithm

IP Instability pattern

JS Jensen-Shannon divergence

JSD Jensen-Shannon distance

KDE Kernel density estimation

KL Kullback-Leibler divergence

LR Lasso regression

LSP Local stability pattern

MDS Multidimensional scaling

MR Multiple regression

PCA Principal component analysis

PDF Probability distribution function

PLSR Partial least squares regression

SAC Sparse autoencoders

SPO Source probabilistic outlyingness

SR Stepwise regression

SSP Sparse stability pattern

RC Rate of convergence

RR Ridge regression

Contents

Report

Report contents

Figures

Chapter 1

Introduction

This chapter offers an overview of this final degree project. Firstly, the main motivations of this project will be presented. Regarding with these rationales, the objectives to accomplish in this work will be established. Afterwards, the findings and contributions of this project will be mentioned. Finally, a brief explanation of the followed outline will be offered.

1.1. Motivation

The Biomedical Data Science Lab (BDSLab) is an interdisciplinary research line of the ITACA institute at Universitat Politècnica de Valencia (UPV). It is focused on solving biomedical problems by means of techniques from pattern recognition, machine learning, modelling and computational prediction, as well as on the development of tools for biomedical data processing. It comprises five main research lines: Clinical decision support & Predictive Analytics, Service delivery intelligence, Big Data technologies in health care, Multiparametric tissue signatures and Biomedical Data Quality.

The author of this final degree project joined BDSLab in 2015 thanks to the program Generación Espontánea UPV. He was introduced there in the field of Biomedical Data Quality and trained in sophisticated statistical tools for assessing data source variability and probabilistic change detection.

After this period, he was able to implement and utilize those variability metrics: the Global probabilistic deviation (GPD) and the Source probabilistic outlyingness (SPO). His initial work was centered on the evaluation of these metrics, developed in BDSLab, in real multi-source biomedical repositories, respect to few parameters. However, the first results revealed that the study of such metrics would require multiple deeper analyzes that would make it possible, when assembled them all in an orderly way, an adequate understanding of GPD, SPO and their usage. Hence, this is one of the main motivations of this final degree project: to face those emerging challenges, providing an experimental but also theoretical approach, explaining the causes of the observed phenomena.

At the same time, he was made aware of the importance of providing comprehensible data source variability analysis results, especially when users were not familiar with those complex statistical metrics (e.g. doctors, hospital managers). This is another essential motivation which derived in the development of this work.

1.2. Objectives

The main objectives addressed in this project can be summarized in the following list:

1. Study the influence of different preprocessing techniques when implementing GPD and SPO metrics, and define criteria for its election.
2. Evaluate the existence of undesired biases in the values of the metrics introduced by pre-set factors (number of data, number of sources, number of attributes, etc).
3. Develop new clustering algorithms based on the theoretical framework of the metrics, and assess it in real multi-source biomedical repositories.

1.3. Contributions

This final degree project presents relevant findings, along with novel methods and guidelines, which will be mentioned beneath:

Firstly, it has been detected in real multi-source biomedical repositories two main behaviors respect to the supports taken for the discretization of estimated probability distributions in different dimensions of reduction. Besides, it has been put forward a new algorithm with applicability in the univariate case to overcome undesired effects, as well as principles in order to develop a similar algorithm in the multivariate case.

Secondly, the effect of the curse of dimensionality over probability distribution functions has been studied, proposing a metric for fixing the maximum dimension of reduction allowed in order to avoid misleading estimations. This metric has been evaluated over different sets of distributions, exhibiting applicability.

Moreover, different dimensionality reduction techniques have been assessed, from linear to non-linear techniques, over simulated datasets and real multi-source biomedical repositories. The results of this study have allowed the definition of guidelines to select dimensionality reduction methods when implementing GPD and SPO, as well as recommendations about principal treats to consider when testing new ones.

Furthermore, it has been conducted an analysis in order to survey possible dependences among GPD and SPO values respect to intrinsic factors of the repositories (i.e. number of data, number of sources, number of variables, etc.), over a set of real multi-source biomedical repositories. The approach taken has been multivariate, comparing different models while using well-established techniques.

Finally, a novel clustering algorithm based on the previous metrics characterization has been proposed and evaluated over real multi-source biomedical repositories, revealing the presence of different data source stability patterns in biomedical repositories. Results of this finding were presented as oral presentation in the 30th IEEE International Symposium on Computer-Based Medical Systems (Ferri-Borredà, Sáez, & García-Gómez, 2017).

1.4. Outline

This final degree project is organized in chapters, each one of them facing specific objectives, since its inherent complexity requires it. However, they are highly related; the findings of a given section provide procedures in order to initiate the approach of the next one, and so on.

Chapter 1 has described the motivations, objectives and contributions of this work. Chapter 2 offers a deeper presentation of concepts related with biomedical data, data source variability and its assessment, the theoretical framework for GPD and SPO metrics and the main issues when implementing these metrics. Then, Chapter 3 presents the materials used in this work. Chapter 4 initiates the characterization task of these metrics, regarding to the issues presented in Chapter 2. After that, Chapter 5 analysis potential dependences among factors related with treats of repositories (e.g. number of data, number of sources) and GPD and SPO values. Chapter 6 describes a novel clustering algorithm for data source variability pattern discovery, its evaluation and findings. Finally, Chapter 8 concludes this report remarking the main contributions of this work and its implications in data source variability assessment.

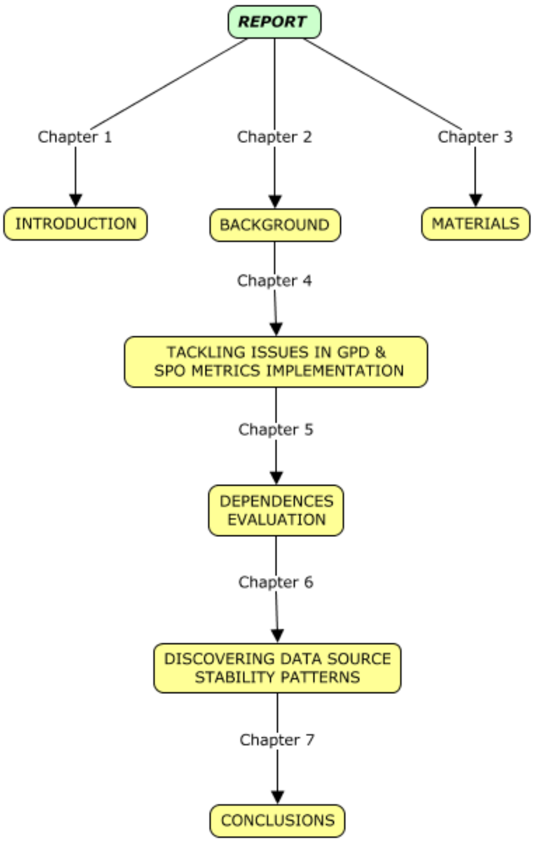


Figure 1.1. Outline.

Chapter 2

Background

2.1. Biomedical data quality

Data is a key resource in a wide range of fields. From the physicist who aims to determine the trajectory of an asteroid to the physician who must decide which is the best treatment for his patient, data based inference is unavoidable.

However, either calculating simple statistical parameters (e.g. mean, median, standard deviation, etc.) or by means of more sophisticated techniques such as clustering, extract information from data is absolutely dependent of its quality; if data do not offer enough quality, the subjacent inference process will be completely misleading (McMurry, 2013).

In the context of biomedical sciences, this lack of data quality not only involves deceiving findings, but can be potentially hazardous. For example, in primary data use (patient care), low data quality may lead physicians to a set of direct errors, such as inappropriate or outmoded therapy, technical surgical errors, inappropriate medication, error in dose or use of medications; and indirect errors, such as failure to take precautions, failure to use indicated tests, avoidable delay in diagnosis or failure to act on results of tests (Aspden, 2004).

Therefore, detect and characterize this data quality deficiency is a priority objective in any clinical decision, research study or any biomedical related task which outcome is highly influenced by information extracted from data.

However, dealing with this challenge is not easy. On the one hand, lack of data quality could be produced by many causes; on the other hand, biomedical data presents increased levels of complexity when used at population level. It comprises huge amounts of data, generated from multiple sources, concerning to diverse data types, with distributions which often reveal the presence of subpopulations and whose interoperability is a complicated task to achieve (Sáez, 2016).

Hence, it seems reasonably in the detection of data quality issues to focus on separate specific data quality aspects for a proper assessment. These aspects are known as Data Quality Dimensions (Wang & Strong, 1996). Once understood, it makes sense to review the available tools for their detection, asses the pros and cons of each one and evaluated them (if they have not been assessed yet). This would lead to the final step of usage criteria definition, allowing a user not familiarized with these techniques to use them properly. Then, another data quality loss cause will be tackled, and so on.

2.1.1. Data source variability

As have been mentioned before, biomedical data often comes from multiple sources (e.g. several hospitals). When dealing with data from different sites, it could be taken various approaches: for instance, if data distributions are rather similar among sources they may be treated as a whole; however, if they are quite disparate maybe it should be better to consider each one individually. Data source variability exercises, therefore, an influence over the data quality and its characterization allows a confident data usage.

There exist some statistical tools which can be used for the detection of these inhomogeneities, from univariate and single type techniques to multivariate and multi-type ones, but most of them are restricted by its strong assumptions, which not always verify in biomedical data (e.g. Gaussian data, homoscedasticity, unimodality, etc.) (Sáez, 2016). However, metrics such as Global probabilistic deviation (GPD) and Source probabilistic outlyingness (SPO), which were proposed in (Sáez, Robles, & García-Gómez, 2017) does not suppose any underlying distribution and allow dealing with multitype data.

Although GPD and SPO have been shown promising in the detection of data source variabilities, their implementation requires to select from a broad range of methods and once implemented its behavior needs to be understood so as to ensure the extraction of right conclusions.

2.2. Theoretical background

2.2.1. Global Probabilistic Deviation (GPD) and Source Probabilistic Outlyingness (SPO)

GPD and SPO metrics for data source variability will be explained in this section, according to its presentation in (Sáez, Robles, & García-Gómez, 2017):

The global probabilistic deviation metric Ω among a set of data sources $S = (S_1, \dots, S_N\ )$ is defined as follows:

$$\Omega(S_1, \dots, S_N\ ) = \frac{Std(P_1, \dots, P_N)}{d_{1R}(D)} \tag{2.1}$$

expression in which denominator refers to a normalization factor, concerning to the distance between any vertex of a regular simplex of dimension D and its centroid:

$$d_{1R}(D) = \frac{1}{2sin\left(\frac{\gamma(D)}{2}\right)} \tag{2.2}$$

here γ denotes the angle between any pair of segments defined from two given simplex vertices to the simplex centroid, in a regular simplex:

$$\gamma(D) = arccos\left(\frac{-1}{D}\right) \tag{2.3}$$

The dimension of the simplex corresponds to:

$$D = N - 1 \tag{2.4}$$

being N the number of data sources.

Back to formula (2.1), numerator can be written this way:

$$Std(P_1, \dots, P_N) = \frac{\sum_{i=1}^{N} d(V_i, C)}{N} \tag{2.5}$$

where N is the number of data sources and $d(V_i, C)$ is Euclidean distance between the simplex vertex V_i and the simplex centroid C. The latter is calculated as the arithmetic mean of simplex vertices:

$$C = \sum_{i=1}^{N} \frac{V_i}{N} \tag{2.6}$$

Each one of these simplex vertices V_i is obtained after performing full-dimensional scaling (De Leeuw, 1993) over a dissimilarity matrix $J = (J_{11}, \dots, J_{NN})$. Each one of these matrix entries J_{ij} represents the Jensen-Shannon distance (JSD) (Endres & Schindelin, 2003) between the probability distribution function (PDF) of data in source i and the PDF of data in source j. The Jensen-Shannon distance is defined as the square root of the Jensen-Shannon divergence (JS) (Lin, 1991):

$$JSD_{ij} = JS(P_i||P_j)^{\frac{1}{2}} = \left(\frac{1}{2}KL(P_i||M) + \frac{1}{2}KL(P_j||M)\right)^{\frac{1}{2}} \tag{2.7}$$

where $M = \frac{1}{2}(P_i + P_j)$ and $KL(P||Q)$ is the Kullback-Leibler divergence (Kullback & Leibler, 1951) between distributions P and Q, whose discrete expression using the base 2 logarithm is:

$$KL(P||Q) = \sum_i log_2\left(\frac{P(i)}{Q(i)}\right)P(i) \tag{2.8}$$

Likewise, the source probabilistic outlyingness metric Θ of a data source S_i, respect to the central tendency among the set of data sources ($S_1, \dots, S_N$) is defined as:

$$\Theta(S_i) = \frac{d(V_i, C)}{d_{max}(D)} \tag{2.9}$$

where numerator is the Euclidean distance between the simplex vertex V_i and the simplex centroid C, and $d_{max}(D)$ refers to a normalization factor:

$$d_{max}(D) = 1 - \frac{1}{D-1} \tag{2.10}$$

being D the dimension of the simplex.

Thus, starting from raw data of each source, PDFs are estimated if the relation among number of available data and attributes allows it. If not, some preprocessing techniques must be taken in order to avoid curse of dimensionality. After that, the Jensen-Shannon distance among pairs of PDFs is computed. Prior to its calculation, PDF should be discretized if they are continuous. Once the dissimilarity matrix is constructed, Euclidean embedding via full-dimensional scaling is performed, resulting in the construction of the simplex. Afterwards, the calculus of the metrics from its definition is straightforward.

As can be inferred from the former explanation, GPD and SPO metrics involve some steps which are chosen by the user (e.g. the way PDFs are estimated). Furthermore, these stages in metrics calculation may present a high influence over GPD and SPO values. Its characterization, thereupon, is compulsory in order to carry out a suitable implementation of the metrics which lets data source variability assessment.

2.2.2. Issues in GPD and SPO metrics implementation

2.2.2.1. Discretizing estimates of probability distributions

Estimation of probability distribution functions (PDFs) is a key issue in statistics. Also known as density estimation, its main objective is to infer how data distributes in a population from a sample taken from that population.

The calculus of Global probabilistic deviation (GPD) and Source probabilistic outlyingness (SPO), as has been mentioned in the previous section, involves a process of density estimation. Concretely, nonparametric density estimation, whose performance will affect metrics values.

However, GPD and SPO are not computed directly from estimated probability distribution functions. They require an evaluation process (also called discretization) which will allow its calculation in a computer. Even an excellent density estimation could be exploited improperly if the supports taken for its evaluation are not apposite.

There are two main approaches in probability distribution function estimation. The first one is the parametric approach while the second one is the nonparametric.

Parametric density estimation is based on the assumption that sample data comes from a distribution which shape is known although its parameters are unknown. These parameters are estimated from sample data. There exist two main types of parameter estimation: the first one is the Maximum likelihood estimation (Edgeworth, 1908), while the other is Bayesian estimation (Lord, 1984).

Otherwise, the nonparametric approach does not make the prior assumptions. Instead, it is based on sample data itself for estimating the distribution of data in population, but without supposing any underlying distribution.

There are several ways of performing nonparametric density estimation. One of the most popular are the histogram, which bin width accomplishes the function of smoothing parameter. If its selection is precise, the representation of the PDF may reflect accurately the PDF of the population. Nevertheless, resort to histograms for PDF estimation could be rather problematic, as can be inferred from equation (2.8), mainly because discrete Kullback-Leibler divergence requires that the evaluated PDFs present the same supports. The logical choice for the supports of an individual histogram would be its bin-width, so when the optimal bin-width of two histograms is different in any of their dimensions, we would be comparing not optimal PDFs estimations. Furthermore, this problem increases as the number of sources grows and the dimensionality of the histogram rises. Hence, histograms will not be considered for metrics implementation.

Another well-known method in the field of nonparametric density estimation is Kernel density estimation (KDE) (Parzen, 1962). This technique constructs an estimation of the PDF of a population from a finite data sample based on the following approach in the univariate case:

$$\hat{p}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right),\ x \in \mathbb{R},\ \mathrm{h} > 0 \tag{2.11}$$

where $X_1,\ldots,X_i,\ldots,X_n$ represent the observations, h is the smoothing parameter (also called window width or bandwidth) and K is a kernel function.

Likewise, it is possible to extend KDE for the multivariate case:

$$\hat{p}_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{n|\boldsymbol{H}|}\sum_{i=1}^{n} K\left(\boldsymbol{H}^{-1}(\boldsymbol{x}-\boldsymbol{X}_i)\right),\ \boldsymbol{x} \in \mathbb{R}^r \tag{2.12}$$

expression in which H is a nonsingular matrix which generalizes the univariate bandwidth h.

Kernel functions verify two main properties:

- Non-negativeness.

$$K(\boldsymbol{x}) \geq 0, \forall\, \boldsymbol{x},\ \boldsymbol{x} \in \mathbb{R}^r \tag{2.13}$$

- Normalization.

$$\int_{-\infty}^{+\infty} K(x)dx = 1, x \in \mathbb{R}^r \tag{2.14}$$

Therefore, there are two main choices in Kernel density estimation: the kernel and the bandwidth. Referring to the kernel, the most popular types are the Gaussian kernel and the Epanechnikov kernel, although exist others such as the biweight or the cosine. However, the choice of the bandwidth is much more critical than the kernel (Izenman, 2009).

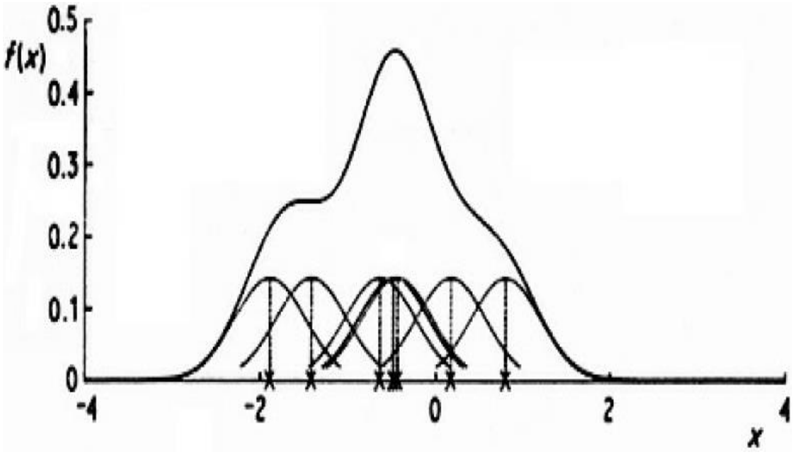


Figure 2.1. Example of a distribution estimated by means of Kernel density estimation (KDE). Found in (Silverman, 1986).

It can be mentioned different techniques in order to perform automated optimal bandwidth selection. Most well-known methods are the Rule-of-thumb (Silverman, 1986), Cross-validation (Rudemo, 1982) and Plug-in methods (Sheather & Jones, 1991).

As can be inferred, KDE implies the construction of a continuous function. However, back to equation (2.8), Kullback-Leibler divergence is computed from a discrete approach. Thus, discretization is required, and supports for performing it must be defined. As with histograms, the supports for the estimation of the PDF of a given data source could not be appropriate for others. However, unlike histograms, this situation can be overcome performing an adequate sampling of the estimated continuous PDF.

2.2.2.2. Measuring the curse of dimensionality

The term "curse of dimensionality" was introduced by Bellman in 1961 when considering problems in the field of dynamic optimization (Bellman, 1961). It refers to the difficulty of confronting statistical issues in high dimensional spaces, adversities which do not appear in lower dimensional spaces.

Curse of dimensionality affects many fields, many of them related with data mining tasks. Density estimation, which has been presented in the previous section, is also affected by this phenomenon, deriving in over adjusted estimations which do not constitute an accurate representation of the population.

The main implication of the curse of dimensionality in density estimation is the fact that the number of data required to perform the estimation of a multivariate probability distribution function increases exponentially with number of dimensions (i.e. variables considered), quantity of data which is not available in most cases (Izenman, 2009), (Lee & Verleysen, 2007).

In the context of nonparametric density estimation, there no exist a metric for indicating how "cursed" is a certain high dimensional space, mainly because it depends on many factors, such as the method used for the estimation, if there is any underlying distribution of the data, etc. However, some authors have found theoretical expressions referring to the rate of convergence (RC) when performing KDE estimation (Clarke, Fokoue, & Helen-Zhang, 2009):

$$RC_{KDE} = n^{\frac{-4}{4+d}} \tag{2.15}$$

expression in which n refers to the number of data samples available, while d represents the dimensionality of the considered space in which KDE is conducted. It can be appreciated that as the number of data increases, the rate of convergence tends to decrease; on the other hand, when the dimensionality increases the rate of convergence is reduced consequently.

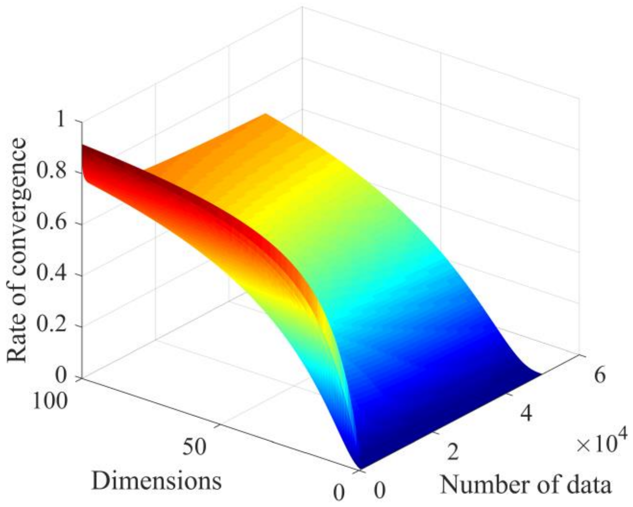


Figure 2.2. Behavior of rate of convergence respect the number of data and dimensionality of the estimated PDF, by means of Kernel density estimation (1).

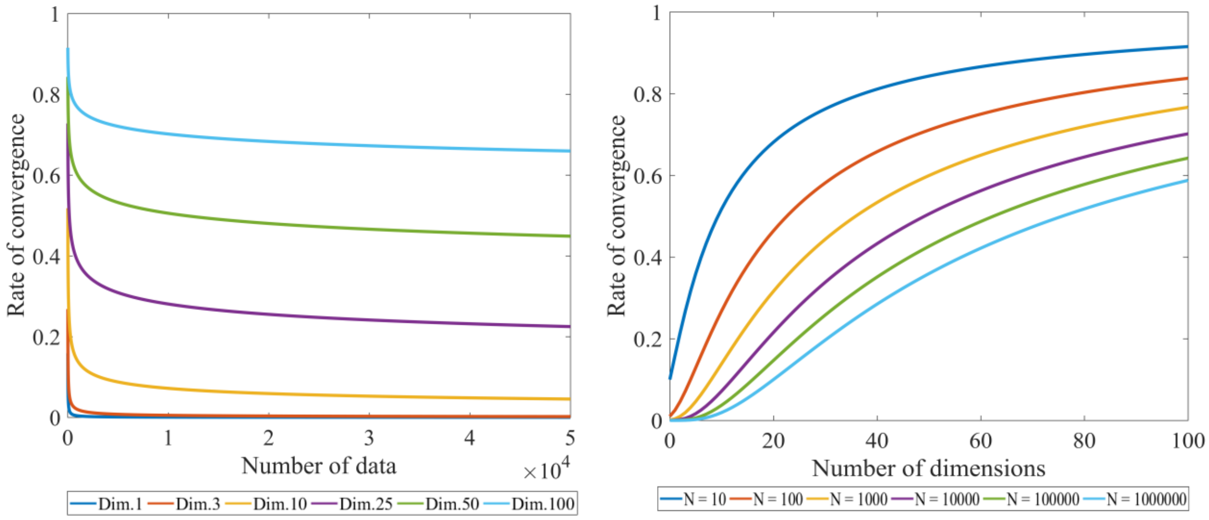


Figure 2.3. Behavior of rate of convergence respect the number of data and dimensionality of the estimated PDF, by means of Kernel density estimation (1).

It has to be recalled here that given a sequence $x_1, x_2, \dots, x_n$, it is stated that converges to a value r with a certain rate of convergence $RC \geq 1$, if there exist a real number $\lambda \geq 1$ such that:

$$\lim_{n\to\infty} \frac{|x_{n+1}-r|}{|x_n-r|^{RC}} = \lambda \tag{2.16}$$

Thus, it is though RC_{KDE} could be taken in consideration so as to define a metric of curse of dimensionality when carrying out Kernel density estimation.

2.1.2.3. Facing the curse of dimensionality

There exist two main approaches when dealing with the curse of dimensionality: the first one is feature selection, while the other one is feature extraction (Alpaydin, 2010). Even though both try to preserve as much relevant information as possible, they accomplish that in different ways. On the one hand, feature selection techniques are based on the reduction of the dimensionality through the selection of a subset of variables. On the other hand, feature extraction creates a new subset of features from the original ones so as to maintain the original information hold in data.

The principal advantages of feature selection process is that its results are easy interpretable, detecting redundant variables which can be omitted. However, this procedure is sometimes dependent of the definition of a cost function which may not be suitable for our specific task, and other implies high computational costs so as to detect relevant variables and its interactions (Alpaydin, 2010).

Referring to feature extraction methods, they present the advantage of considering all the variables in the data, pondering more those variables which are more relevant (from an informative or a discriminative view). Nevertheless, some of them may fail due to the assumptions of some subjacent models in data (Izenman, 2009).

In this work, feature extraction methods will be considered, with the aim of taking into account each one of the different variables relatively. However, it may be interesting to consider in future approaches some feature selection techniques.

There exist a wide range of feature extraction procedures. Formally, they can be classified into supervised methods, which consider a priori information, and unsupervised methods, which do not take into account a priori information. Similarly, they can be divided into linear methods, if they perform a linear transformation over the multidimensional starting space, and nonlinear methods, which deal with the dimensionality reduction process assuming that data lies in low dimensional nonlinear manifolds (Lee & Verleysen, 2007).

Concerning to the evaluation of data source variability, supervised feature extraction methods will not be taking into account, since they could potentiate differences among data distributions which are currently small. Instead of that, unsupervised feature extraction procedures, focused on the maintenance of the variability and information hold in data will be considered.

There are several ways for evaluating the performance of different dimensionality reduction techniques. For example, it can be studied the level of variance retained in the lower dimensional space from the original space, or by means of nonparametric measures, such as the reconstruction error, based on the comparison between the space reconstructed from the reduced data and the original one.

However, despite the wide range of unsupervised methods available, only a subset of them allow back-projection into the data space (i.e. calculate its reconstruction error), such as linear techniques, autoencoders and Gaussian process latent variables models (GPLVM). Other methods like ISOMAP, Laplacian Eigenmaps (LE) or Locally Linear Embedding (LLE) cannot be assessed this way; instead they are generally evaluated comparing the distances among data points in the original space and the distances among data points in the reduced space (Van der Maaten, Postma, & Van den Herik, 2009).

Due to the high memory costs of the assessment of feature extraction techniques which cannot not be back-projected, in this work will just be considered two different unsupervised feature extraction procedures: the first one will be Principal Component Analysis (PCA), one of the most famous feature extraction process, while the other one will be Autoencoders, concretely Sparse autoencoders, which are based on Artificial Neural Networks (ANN) so as to find lower dimensional representations of the original multivariate data. Next it will be presented a brief explanation of each one of these methods:

PRINCIPAL COMPONENTS ANALYSIS (PCA)

Principal components analysis (PCA) is an unsupervised linear feature extraction technique developed by Pearson in (Pearson, 1901). It tries to find a set of orthogonal linear projections of a single collection of variables $\boldsymbol{X} = (X_1, \dots, X_r)^\tau$which may present correlation among them. The projections are ordered so as to the first component explain most of the variance in the original high dimensional space, and so on.

Dissertation provided by (Izenman, 2009) will be considered for the mathematical presentation of Principal component analysis:

Assume that the random r-vector:

$$\boldsymbol{X} = (X_1, \dots, X_r)^T \tag{2.17}$$

presents mean μ_X and covariance matrix Σ_{XX}. PCA aims to move from the r-dimensional space defined by these unordered and correlated input variables to a lower dimensional space described by a set t ordered and uncorrelated variables $(\gamma_1, \dots, \gamma_t)$, $(t \leq r)$ via a linear transformation.

$$\gamma_i = b_i^T X = b_{i1}X_1 + \cdots + b_{ir}X_r, i = 1,2, \dots, t \tag{2.18}$$

where it is tried to minimize the loss of information due to replacement.

PCA attempts to preserve the information hold in the high dimensional space, by interpreting this information as the variance of the original input values:

$$\sum_{i=1}^{r} var(X_i) = tr(\Sigma_{XX}) \tag{2.19}$$

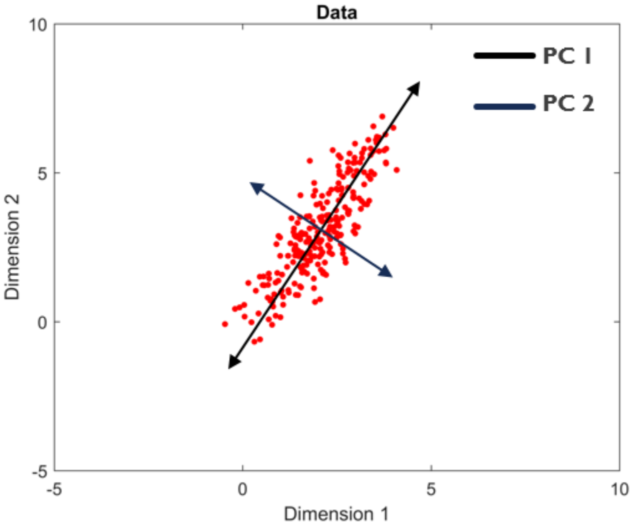


Figure 2.4. Example of principal components obtained after performing Principal component analysis.

It can be derived the following result if we take into account the spectral decomposition theorem:

$$\Sigma_{XX} = UDU^T, U^T U = I_r \tag{2.20}$$

D is a diagonal matrix whose diagonal elements correspond to the eigenvalues $\{\lambda_i\}$ of Σ_{XX}. Furthermore, the columns of U are the eigenvectors of Σ_{XX}, defining an orthogonal basis in $\mathbb{R}^t$.

Therefore, the total variation (i.e. variance) is $tr(\Sigma_{XX}) = tr(D) = \sum_{i=1}^{r} \lambda_i$

The ith coefficient vector, $b_i = (b_{1i}, \dots, b_{ri})^T$ is selected so as to:

- The first t linear projections of X are ranked in importance through their variances, which are listed in decreasing order of magnitude.
- γ_i is uncorrelated with all γ_j, $i \neq j$

The linear projections are also known as the first t principal components of X.

PCA can be derived using a least-squares optimality criterion, or it can be derived as a variance-maximizing technique.

SPARSE AUTOENCODERS

Autoencoders (AC) are a type of Artificial Neural Networks which perform unsupervised nonlinear dimensionality reduction, that is, unlike other ANNs, AC learn informative features from unlabeled data (Liou, Huang, & Yang, 2008). Specifically, AC are ANN which are trained so as to replicate its input at its output, based on the optimization of a cost function (such as other ANNs). This cost function measures the error between the input and its reconstruction at the output.

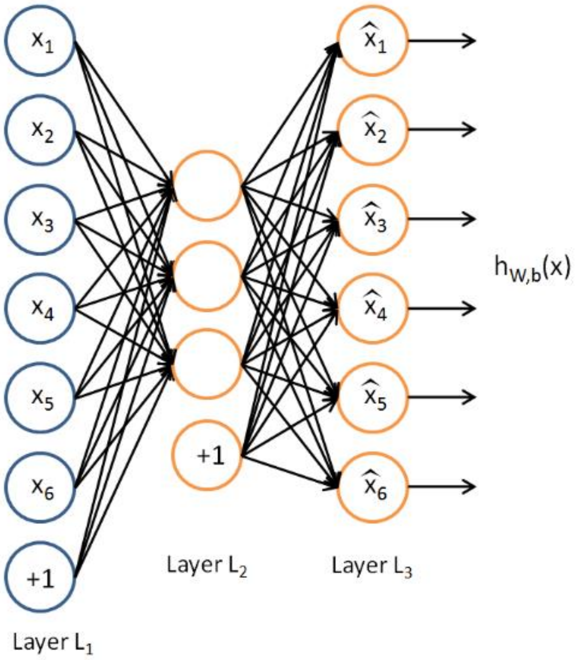


Figure 2.5. Structure of an Autoencoder (AC).

An autoencoder is comprised of an encoder and a decoder, and, as has been mentioned before, tries to learn an approximation to the identity function. However, they attempt to learn this representation subject to certain constraints imposed on the network. One of these restrictions consists on limiting the number of hidden units in the network, which derives in the discovery of a compressed representation of the data taking into account its main features. Of course, if the different variables distribute randomly respect the others, construct this reduced version of the original data will be tough, but if there is an underlying structure in the data (i.e. relation among features) this structure will be learned by the autoencoder.

Furthermore, it is possible to encourage that each neuron in the hidden layer focuses on small number of training examples; in other words, it is feasible that each neuron specializes only in certain features that are present just in a small subset of training examples, by adding what it is called a sparsity regularizer. Autoencoders which present this sparsity regularizer are known as Sparse Autoencoders (SAC), and they present application in many pattern recognition applications (Ng, 2015).

Therefore, the cost function used for training a SAC presents this expression:

$$C_f = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{P}(x_{kn} - \hat{x}_{kn})^2 + \lambda \cdot \omega_w + \alpha \cdot \omega_s \qquad (2.21)$$

here the first summand represents the Mean Squared Error (MSE), ω_w is a L_2 regularization term which is weighted by λ and ω_s is the sparsity regularization term which is modulated by α.

Chapter 3

Materials

In this section, the materials used in this work will be presented. They are divided into two main groups; those which are artificial (i.e. simulated) biomedical data sources and those which are real multi-source biomedical repositories (i.e. each one comprising multiple real biomedical data sources). Both have its own advantages and considerations, which will condition its election. It has to be highlighted that exploiting these particular traits will allow the evaluation of the variability metrics explained in the previous chapter, over many different configurations (i.e. under different preprocessing techniques which may alter its value). Furthermore, they will be crucial to carry out the assessment of the novel algorithms, metrics and procedures which are proposed in this work.

3.1. Simulated biomedical data sources

Simulated biomedical data sources are data distributions artificially generated which try to emulate real biomedical data distributions. Its main advantages are flexibility in its number of data, since it can be established how many data samples are going to be generated, and dimensionality, due to the fact that they can be multivariate with a number of variables specified by the user.

There were considered a total of five main families of simulated distributions, designed according to the inherent peculiarities of biomedical data which were exposed in Chapter 2, as well as author's experience after dealing for a long period with distributions obtained from real biomedical data sources. Next, more details about these distributions are exposed.

3.1.1. Multivariate normal distribution

Given the Central Limit Theorem (CLT), the multivariate normal (also named gaussian) distribution is one common distribution in many fields, also the biomedical field. It presents the following probability density function:

$$f(x,\mu,\Sigma) = \frac{1}{\sqrt{|\Sigma|\cdot(2\pi)^d}} \cdot e^{\frac{-(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}} \tag{3.1}$$

here x represents a multivariate random variable, concretely a vector of size 1-by-d. μ is the vector of the mean in each dimension of the distribution, also which size 1-by-d. $|\Sigma|$ is the determinant of the d-by-d covariance matrix Σ, which is a symmetric positive definite matrix. Finally, d is the dimensionality of the multivariate distribution.

Simulated biomedical data sources coming from this distribution presented the following structure:

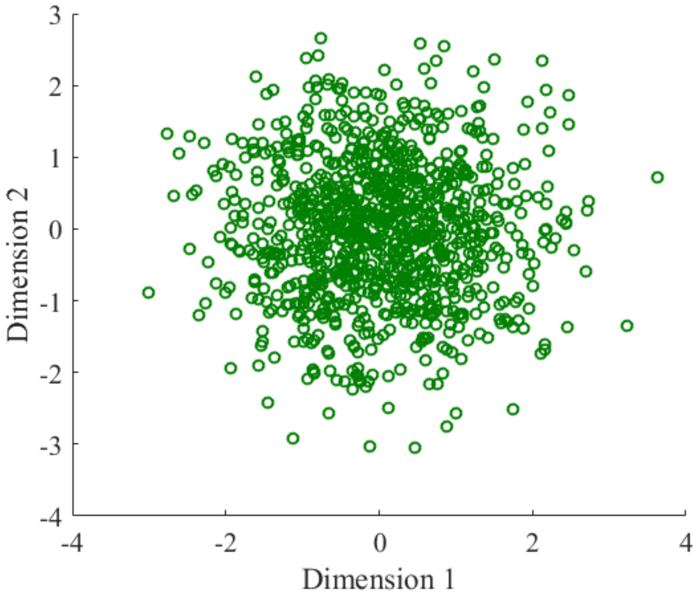
Parameters	2D Visualization
$\mu = (0, \dots ,0)_{1\times d}$ $\Sigma = \mathrm{diag}(1)_{d\times d}$	

Figure 3.1. Multivariate normal distribution.

3.1.2. Multivariate bimodal distribution

As has been mentioned in previous sections of this work, the presence of subpopulations in biomedical data is frequent. Hence, consider just unimodal distributions may result in an accuracy loss.

Bimodality will be emulated in this work by means of the combination of two different multivariate normal distributions. In the case considered in this subsection, these differences will be generated defining a specific mean vector for each one. Furthermore, these subpopulations will be composed by different number of data samples in order to reflect an unbalanced presence in the population of *N* individuals.

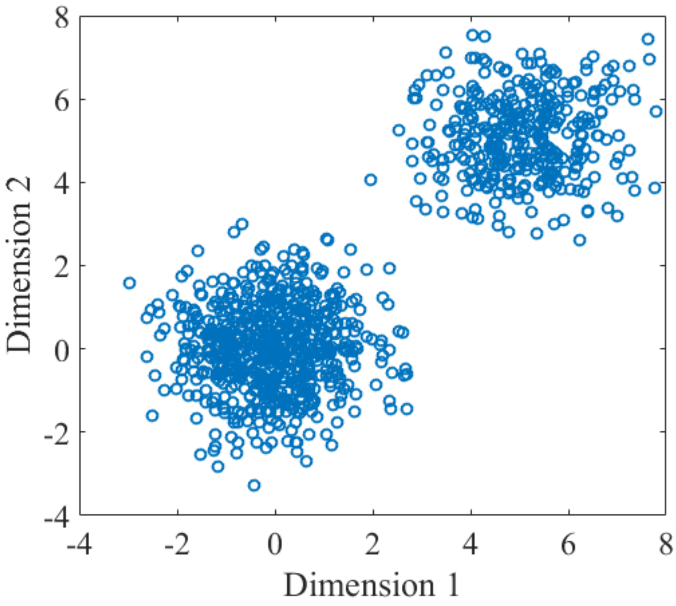
Parameters	2D Visualization
$\mu_1 = (0, \dots ,0)_{1\times d}$ $\Sigma_1 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_1 = 0{,}65 \cdot \mathrm{N}$ $\mu_2 = (5, \dots ,5)_{1\times d}$ $\Sigma_2 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_2 = 0{,}35 \cdot \mathrm{N}$	

Figure 3.2. Multivariate bimodal distribution.

3.1.3. Multivariate weighted-bimodal distribution

Multivariate weighted-bimodal distribution is a variant of the multivariate bimodal distribution, where differences between subpopulations are not originated just by its position, but also by different dispersion.

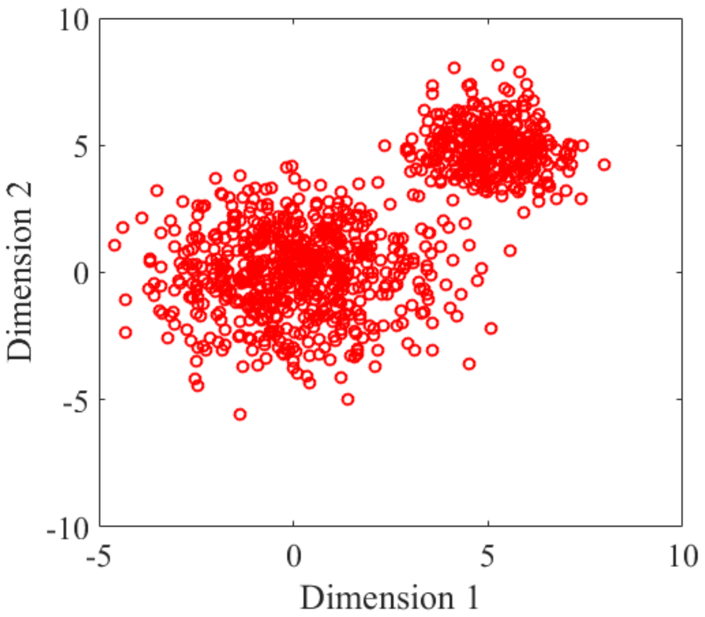
Parameters	2D Visualization
$\mu_1 = (0, \ldots ,0)_{1\times d}$ $\Sigma_1 = \mathrm{diag}(3)_{d\times d}$ $\mathrm{N}_1 = 0{,}65 \cdot \mathrm{N}$ $\mu_2 = (5, \ldots ,5)_{1\times d}$ $\Sigma_2 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_2 = 0{,}35 \cdot \mathrm{N}$	

Figure 3.3. Multivariate weighted-bimodal distribution.

3.1.4. Multivariate multimodal distribution

As exposed in Chapter 2, it is common to deal with biomedical data distributions where may coexist different underlying subpopulations. Previous simulated distributions have assumed the presence of two subpopulations; here more subpopulations will be added, each one of different size, so as to reflect distributions where exist different groups of data (e.g. different clusters of patients). As in the bimodal case, only differences in means will be considered in this type of this distributions (next distributions will encompass these traits).

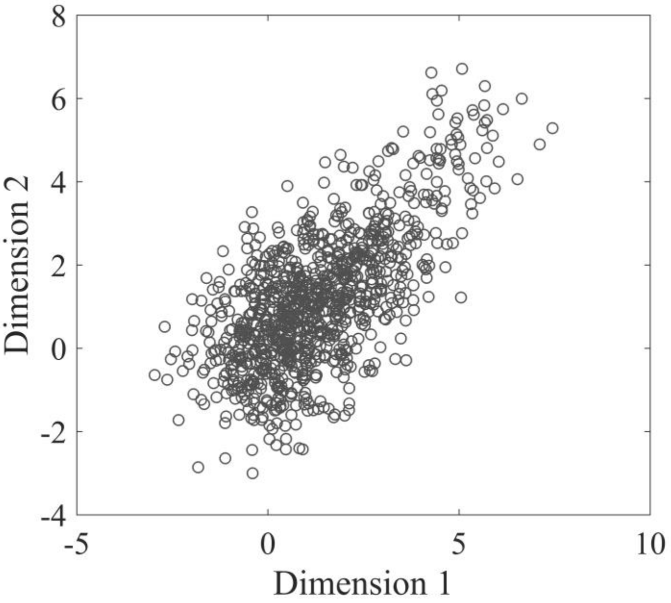
Parameters	2D Visualization
$\mu_1 = (0, \dots ,0)_{1\times d}$ $\Sigma_1 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_1 = 0{,}4 \cdot \mathrm{N}$ $\mu_2 = (1, \dots ,1)_{1\times d}$ $\Sigma_2 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_2 = 0{,}25 \cdot \mathrm{N}$ $\mu_3 = (2, \dots ,2)_{1\times d}$ $\Sigma_3 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_3 = 0{,}2 \cdot \mathrm{N}$ $\mu_4 = (3, \dots ,3)_{1\times d}$ $\Sigma_4 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_4 = 0{,}1 \cdot \mathrm{N}$ $\mu_5 = (5, \dots ,5)_{1\times d}$ $\Sigma_5 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_5 = 0{,}05 \cdot \mathrm{N}$	

Figure 3.4. Multivariate multimodal distribution.

3.1.5. Multivariate weighted-multimodal distribution

Finally, multivariate weighted-multimodal distribution is a variant of the multivariate multimodal distribution, where differences among subpopulations are not originated only by its position, but also by differences in their dispersion.

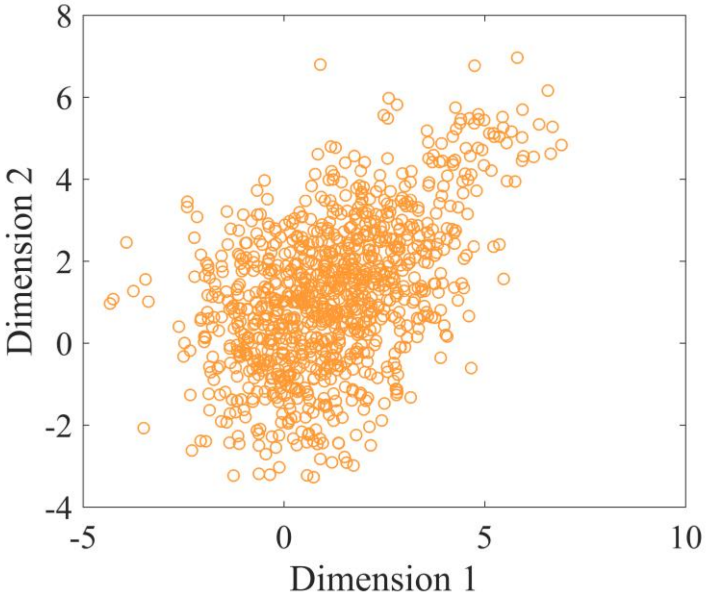
Parameters	2D Visualization
$\mu_1 = (0, \ldots, 0)_{1\times d}$ $\Sigma_1 = \mathrm{diag}(2)_{d\times d}$ $\mathrm{N}_1 = 0{,}4 \cdot \mathrm{N}$ $\mu_2 = (1, \ldots, 1)_{1\times d}$ $\Sigma_2 = \mathrm{diag}(1{,}5)_{d\times d}$ $\mathrm{N}_2 = 0{,}25 \cdot \mathrm{N}$ $\mu_3 = (2, \ldots, 2)_{1\times d}$ $\Sigma_3 = \mathrm{diag}(1{,}25)_{d\times d}$ $\mathrm{N}_3 = 0{,}2 \cdot \mathrm{N}$ $\mu_4 = (3, \ldots, 3)_{1\times d}$ $\Sigma_4 = \mathrm{diag}(1)_{d\times d}$ $\mathrm{N}_4 = 0{,}1 \cdot \mathrm{N}$ $\mu_5 = (5, \ldots, 5)_{1\times d}$ $\Sigma_5 = \mathrm{diag}(0{,}75)_{d\times d}$ $\mathrm{N}_5 = 0{,}05 \cdot \mathrm{N}$	

Figure 3.5. Multivariate weighted-multimodal distribution.

3.2. Real multi-source biomedical repositories

Real multi-source biomedical repositories are comprised each of them by multiple real biomedical data sources. Working with this type of repositories may lead to the discovery of some phenomena which are hard to model in simulated data repositories.

In this work, a total of 42 real-multi source biomedical repositories were collected. They were obtained from different sets of open data portals, published by different institutions. It has to be mentioned that there were considered inclusion criteria referring to the acceptance of a certain repository:

- The repository was real and multi-source (i.e. contains data from different sources).
- It contained more than two data sources.
- Each source contained at least fifty individuals, in order to ensure acceptable PDF estimations.

Furthermore, there were considered exclusion criteria for variables within a repository. Concretely, a repository variable was excluded of the analysis if it was:

- An identifier (e.g. a unique key).
- A codification of another categorical variable present in the repository.
- A variable related to geographic information and not chosen for source specification.

- A categorical variable whose classes were common for all sources (i.e. it was not a real variable in the sense that was previously fixed).

Next Table 3.1 presents some information about real multi-source biomedical repositories collected in this work:

Table 3.1. Real-multi source biomedical repositories.

ID	Num.Data	Num.Sources	NDSS	Reference
1	920	4	123	(UC Irvine Machine Learning Repository, 1988)
2	11257	39	56	(State of California, 2017)
3	163065	51	231	(Department of Health & Human Services , 2016)
4	16350	55	50	(State of California, 2017)
5	23469	200	56	(State of New York, 2016)
6	149851	10	8785	(State of New York, 2017)
7	8207	61	57	(State of New York, 2016)
8	3596	59	60	(State of California, 2017)
9	15194	52	53	(State of New York, 2017)
10	11713	46	50	(State of New York, 2017)
11	28626	57	63	(State of California, 2016)
12	41268	3	9937	(Generalitat Valenciana, 2016)
13	947	4	184	(Generalitat de Catalunya, 2012)
14	405	3	84	(Eusko Jaurlaritza, 2017)
15	11629	3	3232	(Gobierno de Aragón, 2016)
16	6698	6	1110	(Ajuntament de València, 2017)
17	297	3	99	(Eusko Jaurlaritza, 2017)
18	38714	27	664	(NHS Digital, 2017)
19	63403	6	188	(NHS Digital, 2014)
20	11482	64	50	(NHS Digital, 2015)
21	831	4	128	(NHS Digital, 2016)
22	7531	4	1373	(NHS Digital, 2017)
23	7531	4	1373	(NHS Digital, 2017)
24	39877	14	1370	(NHS Digital, 2017)
25	2287	6	168	(Gene Expression Omnibus, 2017)
26	13486	3	1964	(Generalitat Valenciana, 2016)
27	166142	16	834	(Junta de Castilla y León, 2012)
28	596	3	140	(Comunidad Autónoma de País Vasco, 2011)
29	596	3	140	(Comunidad Autónoma de País Vasco, 2011)
30	596	3	140	(Comunidad Autónoma de País Vasco, 2011)
31	596	3	140	(Comunidad Autónoma de País Vasco, 2011)
32	596	3	140	(Comunidad Autónoma de País Vasco, 2011)
33	281	4	57	(Utah Department of Health, 2015)
34	218637	51	111	(Centers for Disease Control and Prevention, 2017)
35	126465	52	2431	(Centers for Disease Control and Prevention , 2017)
36	346344	9	22855	(Centers for Disease Control and Prevention, 2017)
37	6021	34	53	(Centers for Medicare and Medicaid Services, 2017)
38	6019	32	53	(Centers for Medicare and Medicaid Services, 2017)
39	6019	34	53	(Centers for Medicare and Medicaid Services, 2017)
40	19669	46	66	(Centers for Medicare and Medicaid Services, 2017)
41	2741	30	54	(Centers for Disease Control and Prevention, 2017)
42	71779	52	323	(Centers for Disease Control and Prevention, 2017)

Chapter 4

Tackling issues in GPD and SPO metrics implementation

4.1. Introduction

This chapter deals with the main issues in GDP and SPO implementation, regarding to the conceptual framework provided in Chapter 2.

Firstly, it is carried out an analysis of the effect of considering different support points over GPD and SPO values, describing them accurately. After that, it will be offered an explanation of the observed phenomena, and it will be provided and evaluated a method to overcome undesirable effects.

Secondly, the task of studying the behavior of multivariate kernel density estimation respect the number of data, dimensions and shape of the probability distributions functions will be tackled. Besides, guidelines and a metric for limiting the number of dimensions taken in multivariate kernel density estimation so as to avoid curse of dimensionality will be proposed and evaluated.

Finally, different dimensionality reduction methods for the calculation of GPD and SPO metrics will be discussed. Besides, selection criteria for establishing which one of these methods are more appropriate will be defined.

4.2. Discretizing estimates of probability distributions

4.2.1. Effect characterization

4.2.1.1. Methods

Multivariate Kernel Density estimation was conducted, over a subset of the real multi-source biomedical repositories. Dimensionality of the different repositories was reduced, by means of Principal Component Analysis (PCA) with mixed data (using dummy-coding for categorical data), taking one, two and three principal components. Prior to this procedure, data was normalized via z-score, that is, data was standardized so as to have mean 0 and standard deviation of 1.

Then PDFs were obtained using Kernel Density Estimation (KDE) with optimum bandwidth selection provided by (Silverman, 1986):

$$b_w = \left(\frac{4\hat{\sigma}}{3N}\right)^{\frac{1}{5}} \qquad (4.1)$$

being N the number of data samples and $\hat{\sigma}$ the standard deviation of the samples.

The kernel chosen was the Gaussian kernel due to Silverman rule is derived from Gaussian kernels:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} \tag{4.2}$$

After that, continuous estimated PDFs were evaluated using the Dual tree algorithm (Gray, 2003).

The supports for the evaluation were defined in the projected PCA space. First of all, a region of interest was defined, taking into account the optimal bandwidth found and the fact that the 99’7 % of the probability of a univariate Normal distribution is encountered in the area defined by 3 times its standard deviation, which in kernel density estimation coincides with the bandwidth. Thus, the defined region was rectangular:

$$x_{low}^{(k)} = \min\left(X^{(k)}\right) - 3 \cdot b_w \tag{4.3}$$

$$x_{up}^{(k)} = \max\left(X^{(k)}\right) + 3 \cdot b_w \tag{4.4}$$

here $x_{low}^{(k)}$ and $x_{up}^{(k)}$ represent the lower bound and the upper bound respectively of the region for the variable k, which in this case is a latent variable obtained after performing principal component analysis. $X^{(k)}$ is the set of all observations relative to variable k.

Once the region of evaluation has been defined, a decision concerning to the situation of the support points must be made, that is, how to construct the N-dimensional mesh for the evaluation. For instance, the followed criteria could be based on an equally spaced meshing where every consecutive support point is always at the same distance. Otherwise, an adaptive sampling could be considered, concentrating more support points in those regions where the PDFs are more shifting.

In addition to the election of the meshing procedure, the number of support points must also be chosen. Due to the fact that each underlying distribution (i.e. the PDFs of each source) presents a shape which is unknown, the determination of the minimum number of support points required to obtain an accurate representation of the continuous estimated PDF is tricky. Consequence of this unawareness, Nyquist criteria cannot be directly applied. In addition, although the maximum frequency of the multivariate distribution was known for each one of their variables, it will remain the fact of calculating analytically the Fourier transform of an expression which present as gaussians as number of points, calculus which is not straightforward.

Hence, it seems reasonable at this point not to put the cart before the horse and begin with a more intuitive and feasible approach. If the conclusions extracted from this first analysis are informative enough, perhaps more complex studies could be omitted. If not, then it is justified to conduct these tough analyses.

Therefore, in this work it will be considered an equally spaced meshing for each one the dimensions in the projected space, given the bounds obtained in (4.3) and (4.4).

Referring to the number of support points taken within each dimension, there were considered different sets of them for each one of the dimensions. It has to be remarked that as the dimensionality of the estimated PDFs increases, if the number of support points per dimension wants to be preserved, the total number of support points increases exponentially. Thus, this fact conditions the number of points taken as well as the number of PCA projections considered in this study.

4.2.1.2. Results

Figure 4.1 shows the evolution of the Global Probabilistic Deviation (GPD) respect to the number of support points taken for the evaluation of the continuous probability density function (PDF) estimated via Kernel Density Estimation (KDE). So as to facilitate visualization, only a subset of repositories has been represented.

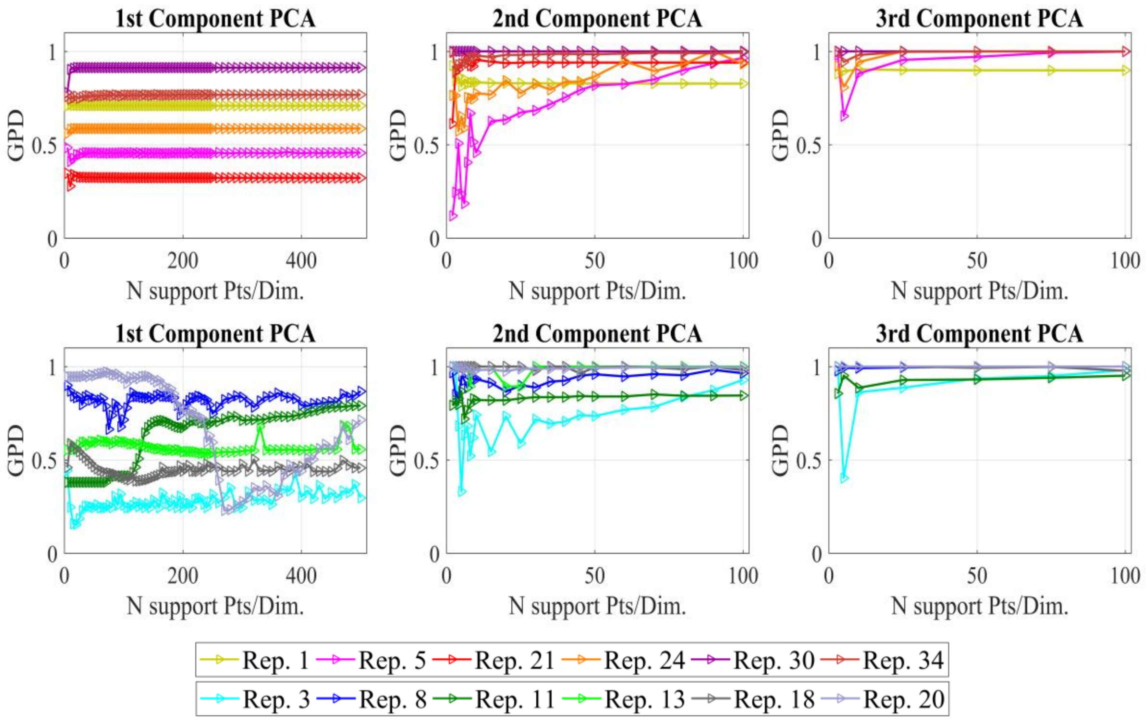


Figure 4.1 Behavior of GPD over different number of support points and dimensionality.

The figure is divided into three columns, each one of them relative to a different dimension of reduction (i.e. to a different PCA components in this case), as well as two rows, with the objective of illustrating two main behaviors: in the first row it can be appreciated that GPD values present a fast convergence in the first PCA component, behavior which is not shown in the GPD values of the second row referring to the same component; in this case, it can observed an oscillatory behavior, which lead to convergence only when the number of support points is high enough.

Furthermore, it can be seen that repositories which offered fast convergence of GPD values in the first PCA component became oscillatory in higher components, where its GPD value tends to increase, sometimes until saturation in 1. Conversely, some repositories which presented an oscillatory behavior in the first PCA component turn into convergent repositories when its dimension is increased.

In addition, although it has been stated that there coexist two tendencies in GPD values, one associated with a convergent behavior while other with an oscillatory trend, both are in fact convergent, as can it will be justified later when the causes of this phenomenon are identified. Hence, it should be better to speak about a fast convergence behavior and a late convergence behavior, controlled mainly by the number of support points, but also by the dimensionality of the estimated continuous PDF.

Figure 4.2 represents the evolution of SPO metric respect the number of support points taken for the evaluation of the continuous PDFs of each source, also for different dimensions of reduction. First row shows an example of a repository which offers fast convergence of the SPO values, while second row presents an example of a repository whose tendency is closer to a late convergence of SPO values.

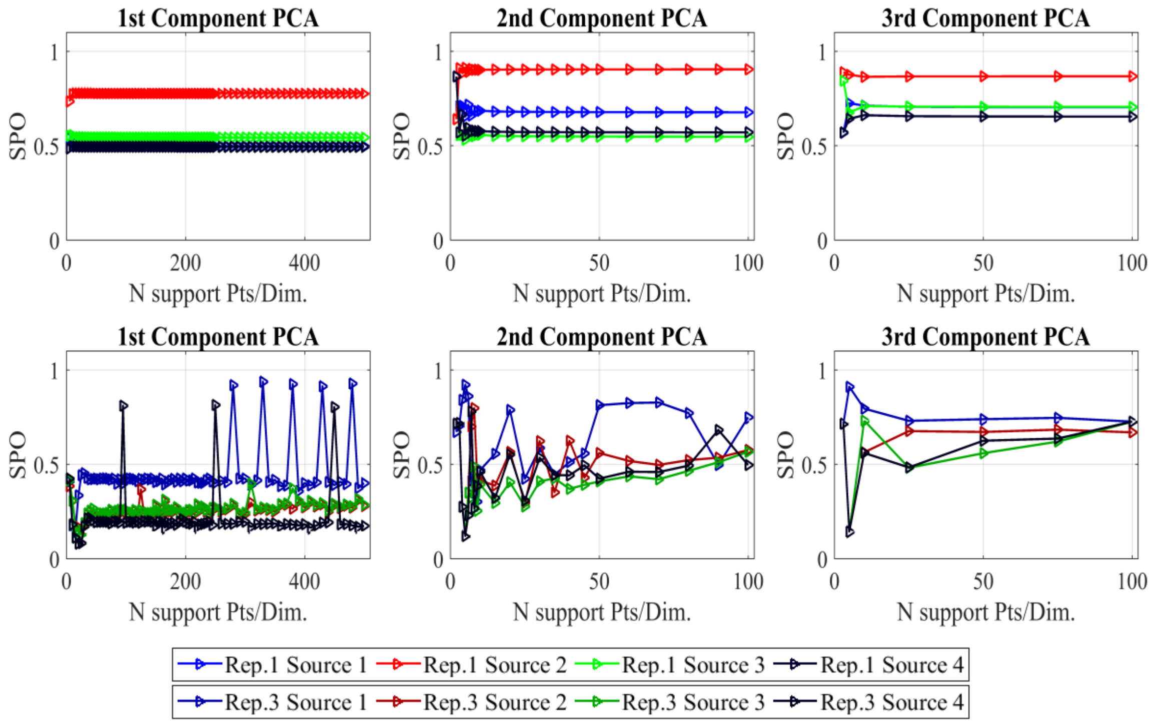


Figure 4.2. Behavior of SPO over different number of support points and dimensionality.

In Figure 4.3 it can be appreciated the cause of the two behaviors. When the distance between consecutive supports points is reduced (i.e. when the number of support points is increased in a certain region) the evaluation of the continuous Probability Distribution Functions (PDFs) becomes more accurate.

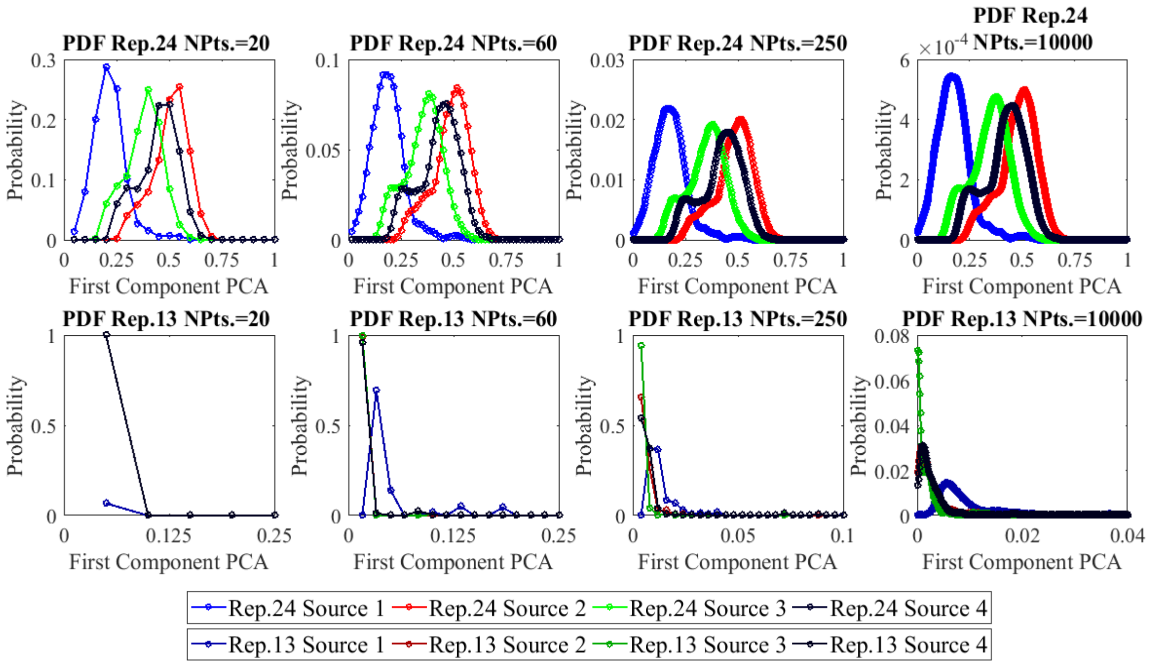


Figure 4.3. Effects of supports taken over probability distribution evaluation.

It can be observed in the first row of this figure an example of PDFs corresponding to a repository which presents a fast convergence behavior in its first principal component of PCA respect to GPD and SPO values. When the distance among support points is high, the shape of the underlying PDFs is roughly represented; however, its main traits are captured. As the space among support points is decreased, the discretization of the continuous estimated PDFs turns into finer representations, but its principal characteristics had already been captured previously with a lower number of points of evaluation. This is the reason that justifies the fast convergence behavior, respect to GPD and SPO, of these types of repositories.

In contrast, the second row of figure 4.3 reflects an example of PDFs relative to a repository which offers a late convergence behavior in its first principal component of PCA respect to GPD and SPO values. It can be inferred from the graphic that when the distance among support points is high, PDFs are not captured; in fact, they could not be even detected because support points may lie in regions where the continuous PDFs shows probabilities near to zero. As the number of evaluation points for a given region is increased, the probability of detection of these sharp PDFs rises; however, its main features are not still summarized in the discretized version, resulting in oscillating GPD and SPO values. Finally, when distances among support points is small enough, continuous PDFs are sampled accurately, leading to the stabilization of GPD and SPO values.

If both rows are compared (and more PDFs, which graphical representations will not be provided in this work due to the limited extent of this document) it is clear that the origin of this phenomenon is the existence of values which great difference in range (even when data is normalized as in all the analyses conducted in this work), resulting in PCA projections where PDFs are concentrated in small regions of the space. Moreover, these extreme values may not constitute outliers, since they could be derived from a multimodal population which huge differences in one or more features among the individuals of different subpopulations; before discarding them, data has to be studied in relation to its context.

Finally, it has to be mentioned that although this justification has been provided with examples taking the first principal component, the dissertation above presented is extensible to any dimension of reduction.

4.2.2. Proposed solution

Given the results obtained in the previous section, it seems clear that establishing a number of support points which offer convergence in GPD and SPO values is critical so as to avoid misleading interpretation of the metrics. Then, the user has to take into account this phenomenon when implementing them.

One way of assessing convergence of these metrics could be carrying out the same kind of analyses that have been conducted in the previous section. However, this kind of evaluation implies high computational costs, especially when the dimensionality of the continuous estimated PDFs is considerable, as well as the number of sources.

In this section, an algorithm for facing this tricky situation will be provided, which applicability in the univariate case, but whose principles can be extended to the multivariate case. Furthermore, it will be evaluated over those repositories which presented an oscillatory behavior, checking its effectivity.

4.2.2.1. Hill finder algorithm

The algorithm is based on the detection of the different crests associated each one with regions where the probability of the PDF concentrates. It has to be remarked that these regions could be or not coinciding each one with a mode of the distribution, since if the modes are near each other, the region detected will comprise them. Therefore, it has been decided to name it as “Hill finder algorithm (HFA)”, since the similarity between the shape of the PDF at the found regions and hills.

HFA requires just two parameters for its running: the number of support points for the evaluation of the less variance hill (i.e. region which concentrates an amount of probability which is not depictable) and the value of a weighting factor k. It can be summarized in the following steps:

1. Sorting of the data used for carrying out Kernel Density Estimation (KDE).
2. Calculation of border points. A point is considered border point if the difference between the consecutive point greater and this point is higher than $2k$ times (if the kernel used is Gaussian, a value of $k = 3$ would offer a great performance) the optimal bandwidth chosen for the estimation of the univariate PDF with KDE.
3. Sorting of the border points.

4. Obtaining non-depictable probability intervals. This is achieved by calculating the difference between consecutive border points, where minuends are those border points relative to even positions, and subtrahends are those border points relative to odd numbers.
5. Identification of the hill with less extension, that is, the smallest non-depictable probability interval.
6. Calculation of the distance between two consecutive support points in that region given the number of support points relative to the input parameter of the algorithm.
7. Discretize the PDF taking as many support points as necessary to maintain the distance between points obtained in the previous step.

It has to highlighted that although the HFA has been exposed for just a PDF, when calculating GPD and SPO metrics it has to be applied to the PDF of every source and the final discretization mesh would be that which presents the smallest interpoint distance, since all the PDFs must have the same support points.

4.2.2.2. Evaluation

It was conducted a first evaluation of the Hill finder algorithm over two repositories: repository 1, which offers a fast convergence behavior respect GPD and SPO values and repository 8, which presents a late convergence behavior. The number of support points for the evaluation of the less variance hill was set in 10, while the value of the k parameter was established in 3.

Table 1.1 shows the performance of the HFA in both repositories considered, respect to GPD values, where GPD_{ref} is the value of the metric when it converges (i.e. the reference value) and GPD_{hfa} is the GPD value obtained using the support points provided by the Hill finder algorithm. It can be appreciated a depictable error in the calculation of GPD metric in both repositories.

Table 1.1. Performance of the HFA respect to GPD values.

Repository ID	GPD_{hfa}	GPD_{ref}	Relative error (%)
1	0,710685	0,708733	0,275501
8	0,920669	0,920957	0,031274

Table 1.2 illustrates the performance of the HFA in repository 1 (fast convergence), respect to SPO values, where SPO_{ref} is the value of the metric when it converges (i.e. the reference value) and SPO_{hfa} is the SPO value obtained using the support points provided by the Hill finder algorithm. It can be observed the relative error in the SPO calculation offered by the algorithm is truly small.

Table 1.2. Performance of the HFA respect to SPO values in repository 1 (fast convergence).

Repository 1	SPO$_{hfa}$	SPO$_{ref}$	Relative error (%)
Source 1	0,499838	0,497826	0,404236
Source 2	0,780362	0,778869	0,191667
Source 3	0,545579	0,543535	0,375943
Source 4	0,495310	0,494482	0,167538

Finally, Table 1.3 provides information about the performance of the HFA in repository 8 (late convergence), respect to SPO values, with the same nomenclature considerations respect the previous table. Also in this table, it can be seen that the estimation provided by the HFA algorithm presents a minimum error respect to the reference value.

Table 1.3. Performance of the HFA respect to SPO values in repository 8 (late convergence).

Repository 8	SPO$_{hfa}$	SPO$_{ref}$	Relative error (%)
Source 1	0,379793	0,379122	0,176861
Source 2	0,744412	0,744145	0,035919
Source 3	0,756755	0,756489	0,035156
Source 4	0,683014	0,682740	0,040173
Source 5	0,756413	0,756146	0,035229
Source 6	0,564633	0,564268	0,064674
Source 7	0,756557	0,756291	0,035199
Source 8	0,621003	0,620736	0,043031
Source 9	0,756609	0,756343	0,035187
Source 10	0,748539	0,748270	0,036065

4.3. Measuring the curse of dimensionality

4.3.1. Methods

The proposed procedure to determine which is the minimum number of data required so as to carry out multivariate kernel density estimation is based in the expression (2.15). Firstly, the minimum number of samples required to perform univariate kernel density estimation has to be defined. This step could be tricky; however, it can be taken into account the graphics presented in Figure 2.3 in order to set this value (i.e. when the expression of the RC_{KDE} for $d = 1$ begins to stabilize). A good practice if one wishes to reduce subjectivity would be to define two values, a lower and an upper value, constructing something similar to an interval of acceptance. After that, the following equation has to be solved, for a given value of D:

$$n^{\frac{-4}{4+D}} = RC_{KDE}(d = 1) \tag{4.5}$$

The rounded n value obtained will correspond to the minimum number of data required in dimension D so as to perform multivariate kernel density estimation in that dimension.

So as to assess the evolution of values of this metric, there were fixed different minimum number of data samples for the estimation in the univariate case, and for each one of them, rounded n values were calculated, from $D = 2$ to $D = 10$.

Furthermore, a validation procedure was conducted, based on the simulated multivariate probability distribution functions explained in Chapter 3.

First of all, these distributions were created, each one containing 10 million data, and presenting dimensionality from 1 to 4. Then, its respective PDFs were estimated via multivariate kernel density estimation. After that, there were extracted several random samples from each family, with increasing size: 5, 10, 20, 50, 100, 250, 500, 1000, 2500, 5000, 10000, 25000, 100000, 500000 and 1000000.

In addition, with the aim of reducing possible undesired effects introduced by a unique extraction process, there were extracted 100 random samples for a given size and family.

Probability distribution functions were estimated for each one of the different random samples for a given size via multivariate kernel density estimation. Jensen-Shannon distances were calculated among each one of these distributions and the reference distributions (those with 10 million simulated data). The kernel used was Gaussian, combined with the automatic bandwidth selection method proposed by (Silverman, 1986). In addition, the number of bins used for the estimation was high enough in order to avoid instability problems.

4.3.2. Results

Figure 4.4 shows the minimum number of data that is required in each dimension so as to keep a constant value of the curse of dimensionality metric. It can be appreciated that the proposed measure shows an exponential behavior, which is in concordance with the exponential need of sample data, characteristic of the curse of the dimensionality. It has also to be mentioned that, as can be inferred from the graphics, the determination of the minimum number of data in the one-dimensional case (MND_{1D}) is critical: due to the exponential changes respect to dimensionality, small variations in this parameter derives in huge differences as dimensionality increases.

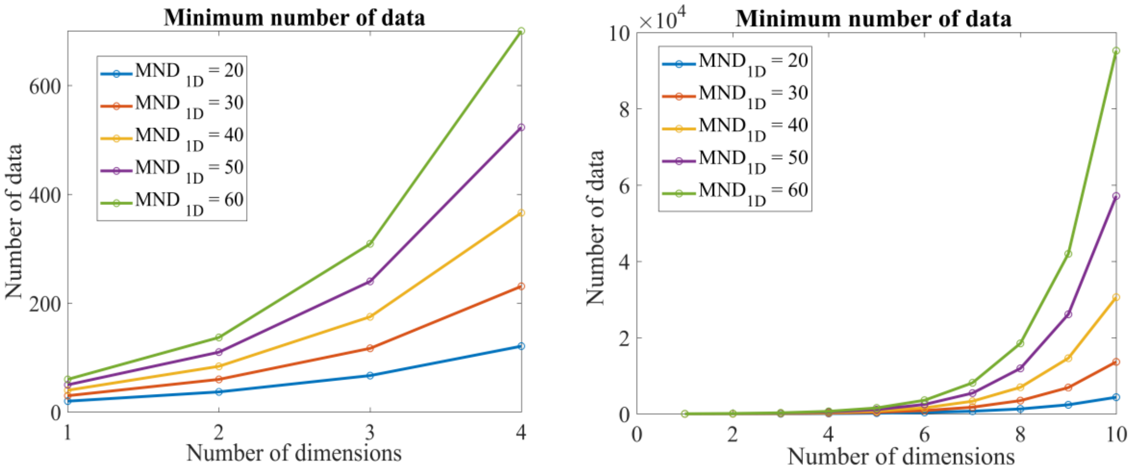


Figure 4.4. Metric for assessing curse of dimensionality.

Figures 4.5 and 4.6 represent the median of the Jensen-Shannon distances for each one of the extractions, respect the number of data, and respect to the different dimensions considered and multivariate distribution families.

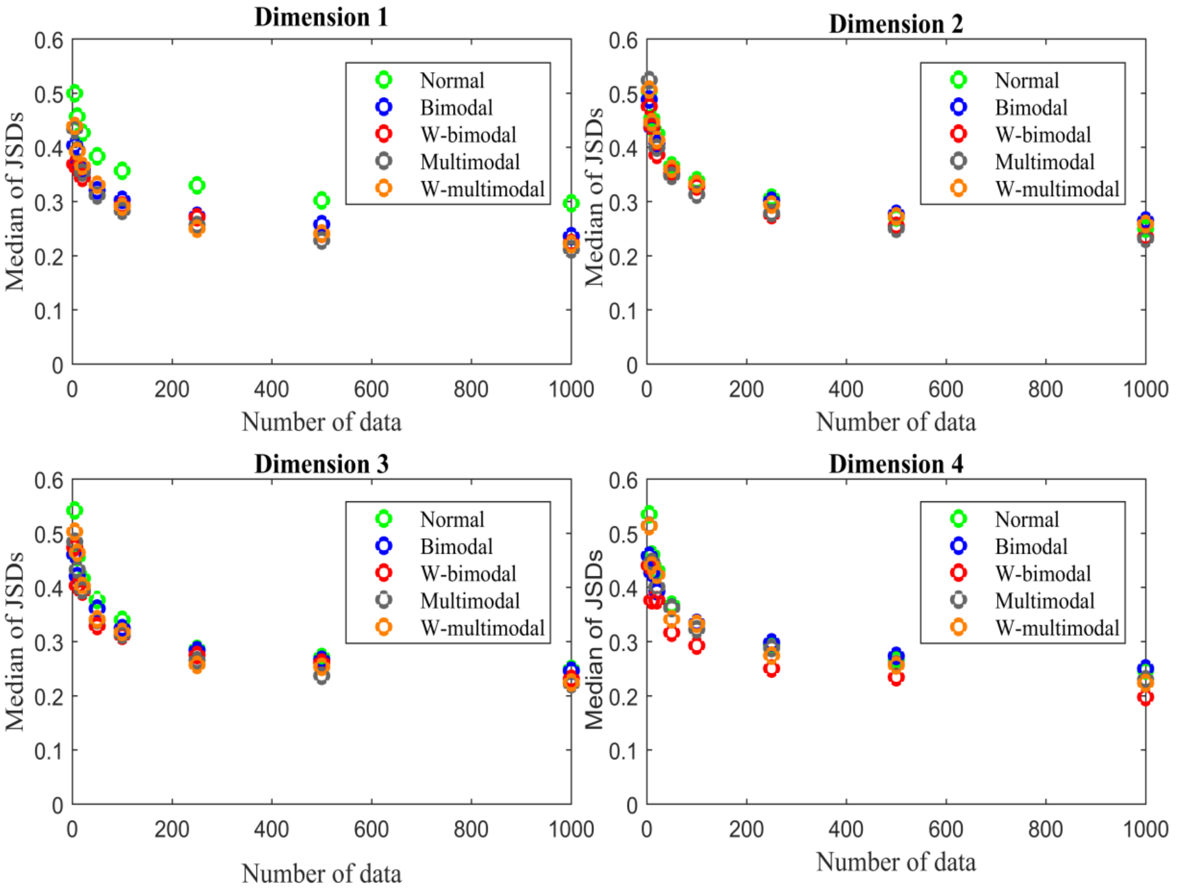


Figure 4.5. Evaluation of the effect introduced by the curse of dimensionality (1)

The first thing that has to be highlighted is the similarity among these graphics and Figure 4.4, supporting the initial though considered in this chapter which states that it is possible to establish a relation among the RC_{KDE} and the goodness of the nonparametric density estimation. It can also be seen that there is not a clear difference among the different graphics, even though the dimensionality of each one of the different multivariate distributions is different; what would be expected would be that the median of the Jensen-Shannon distances tends to increase as the dimensionality of the PDF does. A possible explanation to this phenomenon could be merely randomness; for a given family of distributions, its data it is not directly related with data of the same family but different dimension (i.e. values for common dimensions vary because of the random generation, although the family parameters are pre-set).

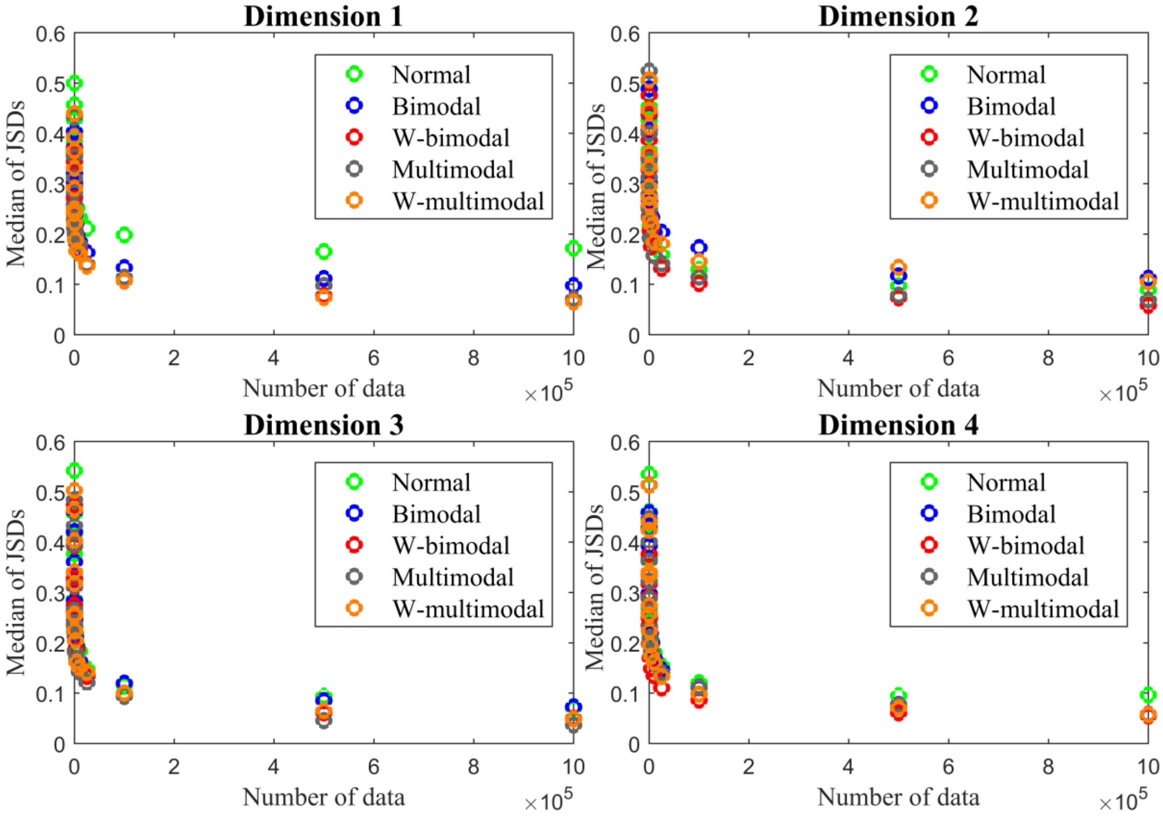


Figure 4.6. Evaluation of the effect introduced by the curse of dimensionality.

4.4. Facing the curse of dimensionality

4.4.1. Methods

Performance of PCA and SAC was assessed over simulated repositories and real multi-source biomedical repositories.

Referring to the simulated repositories, there were considered the same families as in the previous section, but the number of data and the dimensionality of each one of them was varied so as to adapt to the specific purposes of this chapter.

Concerning real multi-source biomedical repositories, it was used a small subset of them given the training time of the networks. Concretely, there were considered repositories 1, 3, 4, 7, 8, 9 and 10.

It was performed dimensionality reduction from the original data in the high dimensional space, using PCA and SAC. The dimensions of reduction considered were 1 to 4, in relation with the limitations defined in previous sections. The reconstruction error for 45 simulated repositories was calculated, both for PCA and SAC. This reconstruction error was also evaluated in the 7 real multi-source biomedical repositories considered.

The reconstruction error (ρ) is calculated as follows:

$$\rho = \frac{\|\hat{X}-X\|_2}{\|X\|_2} \tag{4.6}$$

expression in which X is a matrix representing the original data (rows for observations and columns for features, or vice versa) while $\hat{X}$ refers to the reconstructed data based on the low dimensional representation. Hence, the reconstruction error is a nonparametric indicator of the goodness of a certain feature extraction procedure, which is lower bounded by 0 when the reconstruction is perfect.

4.4.2. Results

It can be seen in figure 4.7 that the behavior of Principal Components Analysis (PCA) and Sparse Autoencoders (SAC) is very similar for each one of the configurations. It has to be highlighted that for the Multivariate Bimodal, Multivariate Weighted-Bimodal, Multivariate Multimodal and Multivariate Weighted-Multimodal, the performance of both feature extraction techniques is excellent, presenting values of the reconstruction error always below 0.5, even when the amount of data is low and the dimensionality of the original dataset is high.

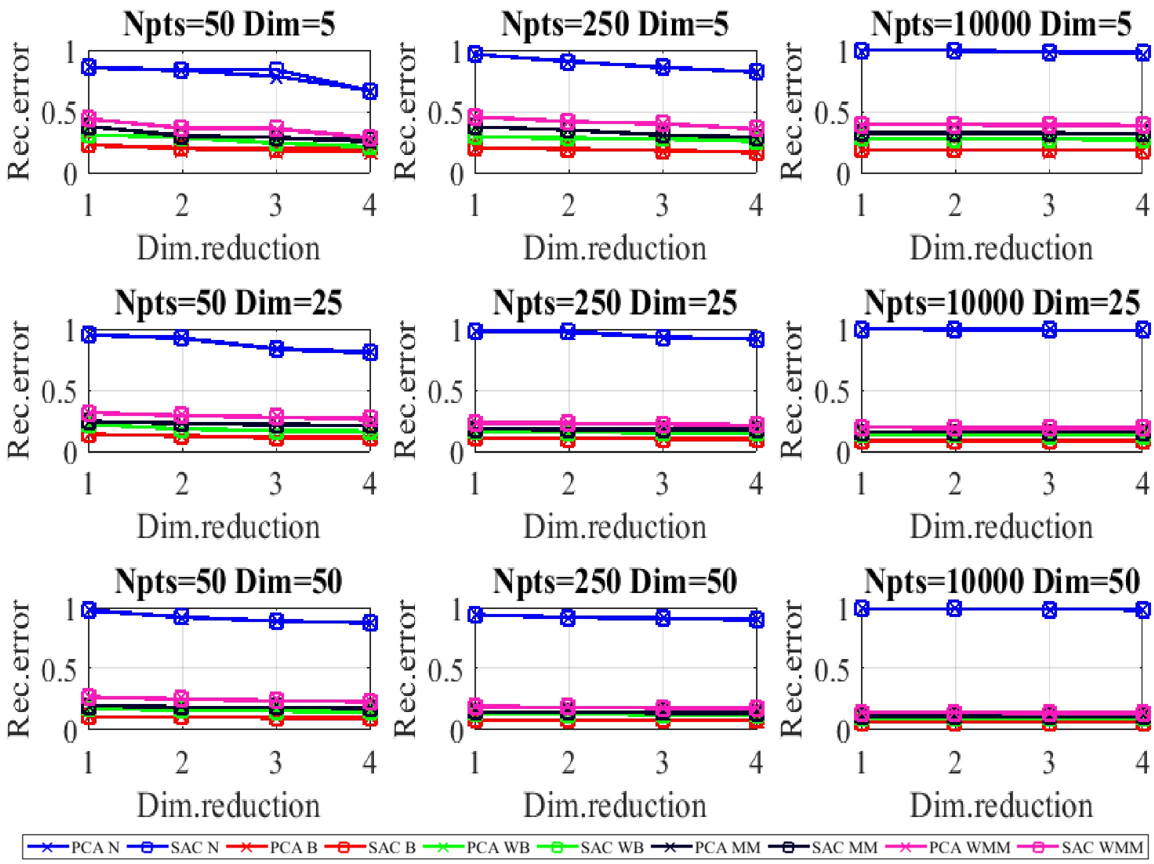


Figure 4.7. Performance of PCA and SAC in simulated repositories.

Referring to the dimensionality reduction of the Multivariate Normal Distribution, its reconstruction errors are much greater than other distributions. This fact is not surprising if is considered its N-dimensional symmetry (i.e. all the variables present the same relevance in order to represent the distribution, so selecting a subset of them is difficult). Furthermore, another interesting result is that these errors tend to increase with the number of data of the original space. A possible explanation of both phenomena could be that, unlike other simulated multivariate distributions, the Multivariate Normal Distribution presents a more diffuse structure in their data, that is, the presence of main directions for variability in other distributions facilitates data compression.

Figure 4.8 shows the performance of PCA and SAC over the subset of real multi-source biomedical repositories considered. The similarity between PCA and SAC reconstruction errors that was appreciated in Figure 4.7 when the repositories were simulated is maintained in this graphic, although some light differences appear. In general, Sparse Autoencoders offer a better reconstruction when are used for performing feature selection. However, this improvement over PCA is hardly ever considerable; just in certain repositories under concrete dimensions of reduction (repository 1 in the second dimension of reduction and repository 9 in its first dimension of reduction).

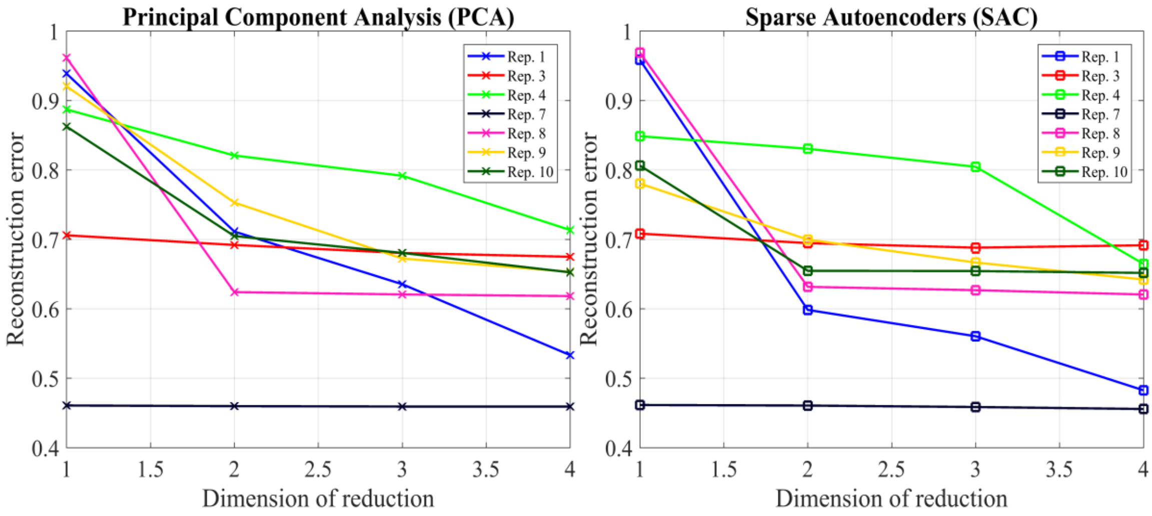


Figure 4.8. Performance of PCA and SAC in real repositories.

4.5. Discussion

4.5.1. Significance

Concerning discretization of estimated probability distributions, there have been detected two main behaviors in GPD and SPO values. As the number of points for performing kernel density estimation increases, it can be appreciated a group of repositories which offer fast convergence, but also other group with serious convergence problems. These problems may be caused by great differences in the range of the estimated PDFs within a repository (despite data normalization).

However, the Hill finder algorithm seems to tackle this effect in the univariate case, and although it requires the development of more formal criteria for the selection of its parameters, allows the obtaining of GPD and SPO values in the convergence zone.

Regarding the measurement of the curse of dimensionality, if Figure 4.4 is compared with Figure 4.5, it can be derived that for the minimum number of data in the first dimension considered in Figure 4.4, the obtained bounds seem optimistic. In spite of this, limits established by the proposed dimensionality reduction metric are close to the experimental values found. Hence, although this measure requires a deeper analysis, especially for the determination of the minimum number of data required to obtain an accurate PDF estimation in the univariate case (value which condition next values in higher dimensions) it has to be taken into consideration and can be applied, along with the experimental graphs obtained, so as to dispose of guidelines for limiting the dimensionality of the estimated PDFs.

Finally, referring to the election of a certain feature extraction method for GPD and SPO implementation, several things have to be taken into account. Focusing on the techniques analyzed in this chapter, each one them presents its advantages and disadvantages.

For example, although in the simulated repositories the performance of both methods was high similar, in the real case it seemed that SAC offer a slightly better accuracy in the representation of the high dimensional spaces. Furthermore, for certain repositories these light differences could turn into considerable improvements.

However, SAC depends on the initialization of weights, which uses to be random. Therefore, its reproducibility is not ensured, although similar results are achieved. Other drawback of SAC is its training time when compared to PCA; PCA is much faster. Finally, there is the fact that there are a lot of free parameters which have to be set for the running of the SAC and optimizing each one of them could result a tough task; PCA does not present this issue, since it is nondependent of the definition of any parameter for its running.

Thereupon, depending on application one option or other could be taken, and they would be acceptable in each one of those cases. For instance, if speed is important, PCA offers it, combined with a barely accuracy sacrifice respect to SAC (although in some case its accuracy could be even better than the obtained by the SAC). On the other hand, if time is not a priory, SAC could be tested (also along with PCA) so as to check if they can offer a better feature extraction process. In this situation, an optimization process of the parameters should be conducted, evaluating different parameters combinations, instead of falling into arbitrary choices.

4.5.2. Limitations

The main limitation in this chapter is the huge computational cost, which is present in each one of the analysis carried out. Despite the use of the Dual tree algorithm, execution times for the evaluation of the continuous estimated PDFs are huge, and memory requirements increase exponentially with the dimensionality of the distribution. Same considerations can be applied when simulating high dimensional distributions.

Another limitation of this work is the number of repositories considered. Although in the simulated repositories case, the amount of distributions which has been studied is acceptable, maybe more real multi-source biomedical repositories should be analyzed.

Finally, it remains a limitation regarding to the training time of the Sparse Autoencoders (SAC), which can be sometimes really high, deriving in a practical limitation of the number of configurations that can be tested. Regarding to this last aspect, in this chapter SAC have been used always with the same parameters combination. It has not been carried out any process for arriving to some network parameters combinations that enhance the feature extraction process, but this fact is justified by the huge computational cost required in order to accomplish it.

4.5.3. Future work

Future work will include the definition of new criteria in order to construct the mesh for the support points. For example, it could be used an Epanechnikov kernel instead of Gaussian kernel so as to dispose of regions with null probability whose discretization would not be necessary. Furthermore, computational times would be reduced.

In addition, it has to addressed the extension of the Hill finder algorithm (HFA) to the multivariate case, probably including those novel meshing criteria, due to the high computational cost of evaluating estimated multivariate probability distribution functions, even when it is used the Dual tree algorithm.

However, one of the most promising future lines is finding an analytical expression for the calculation of the Kullback-Leibler divergence. This expression would avoid the evaluation of the continuous estimated probability distribution, deriving in an exact accuracy in the calculated distances and depictable calculation times.

Further work will also include the development of an accurate rule so as to define the minimum number of data required to carry out univariate kernel density estimation and avoid overfitting. Furthermore, seek new approaches based on RC_{KDE} or other theoretical expressions which were more directly related with the tackled task could be interesting, since its calculation is fast and the information provided is useful.

In addition, future work may include the study of additional feature extraction techniques, starting from those which allow back-projection to the original high dimensional space, such as Gaussian process latent variables models (GPLVM).

Then, it would be interesting to consider other feature extraction techniques which cannot be back-projected, such as Sammon mapping (SM), ISOMAP, Locally Linear Embedding (LLE) or Laplacian Eigenmaps (LE). A possible way to evaluate this dimensionality reduction methods could be calculating distances among instances before carrying out the feature extraction process, and then comparing that distances with those in the low dimensional space. A more in-depth study about which distances measures are appropriate for multi-type data (e.g. distances which do not overweight categorical variables), among the different multivariate observations should be conducted, combined with techniques for optimizing memory resources so as to perform this type of evaluation.

Finally, it remains a last and crucial aspect related with feature extraction techniques, which has not been addressed in this chapter, and that it is related with the following question: what is the optimal number of reduction components? If most information is preserved, for instance, taking 3 components of reduction, it makes no sense to perform feature extraction processes with more than those components. Studying criteria for the determination of this optimal number of components will be also the next step of this work.

4.6. Conclusions

In this chapter, there have been carried out several analyses, comprising each one of them critical implementation issues:

Firstly, a study of the influence between support points taken for the discretization of estimated probability distribution functions and GPD and SPO values has been conducted, detecting two main metric tendencies: a fast convergence behavior and a late convergence behavior.

After that, a novel algorithm has been proposed in order to address late convergence behaviors in the univariate case. Its first evaluation over a subset of repositories shows a promising performance.

Then, the effect of the curse of dimensionality has been assessed, characterizing the overfitting phenomenon introduced in density estimation. In addition, a new metric for the measurement of the curse of dimensionality effects has been proposed and evaluated, showing applicability.

Finally, different dimensionality reduction methods are discussed. An experimental evaluation of some of them has been conducted, and guidelines for its election have been defined.

Chapter 5

Dependencies evaluation

5.1. Introduction

In the previous chapters it was examined the effect of different settings and preprocessing techniques over the metrics under study for the evaluation of data source variability. There were found several issues when implementing these metrics, and there were provided guidelines, algorithms and metrics to tackle them. However, the influence of pre-set factors (e.g. number of data, number of repository sources, number of variables, etc.) over the values of GPD and SPO was not analyzed.

Therefore, in this chapter it is carried out a statistical analysis so as to determine if there exist some biases introduced by factors which should not be correlated with GPD and SPO values.

5.2. Background

Determining the influence of certain pre-set factors respect the values of GPD metric implies carrying out some statistical analysis. If it is considered that it is aimed to study possible dependences of scalar values respect different continuous attributes (number of data, number of sources, number of variables in the repository, etc) it makes sense to conduct some Multiple Regression (MR) analyses.

There are two main approaches when performing MR: the first one is based on linear regression models, which assume there is a linear relationship among one variable (predicted variable) and other ones (predictors variables); the second tactic is to consider that there exist nonlinear relations among the predictors variables which are explicative for the predicted attribute.

In this work only linear techniques are considered, mainly because the number of repositories that are used could lead to overfitting when recurring to these nonlinear models (linear models use to be more robust in this sense).

The general expression of a multiple linear regression model could be written this way:

$$y = X\beta + \varepsilon \tag{5.1}$$

where y is a N-by-1 vector (being N the number of observations) containing the values of the dependent variable, X is the matrix of predictors whose size is $N{\times}p$ being p the number of explanatory variables (or $N{\times}(p+1)$) if a constant term (i.e. intercept) is considered in the model, β is the N-by-1 vector of coefficients (each one related with a different predictor) and ε is the vector of errors, whose size is N-by-1.

Obviating the assumption relative to the existence of linear relationship between the predictors and the dependent variable, there exist three main assumptions in multiple linear regression:

- **Homoscedasticity:** different response variables present the same variance in their errors (difference between the predicted value by the mode and the real value).
- **Independence of errors:** errors of the response variables are not correlated each other.
- **Lack of multicollinearity:** it is assumed that any explanatory variable can be expressed as a linear combination of other explanatory variables.

Failure to comply with these assumptions may lead to misleading results. For example, if multicollinearity is present, then exist problems related with the estimation of regression coefficients, since they are dependent of the inverse matrix of $(X^T X)$, which would be singular in this case.

Among the large number of different linear models for MR, a set of four families of models will be evaluated in this chapter, each one of them presenting particular traits which could make them useful for the defined task:

- **Stepwise regression (SR):** provides an automatic variable selection process involving a set of steps, in which variables are added or removed from the model according to a statistical criterion (such as an F-test) (Efroymson, 1960). There are two main types of stepwise procedures in regression; backwards elimination and forwards selection:
  - **Backwards elimination (BE):** begins with the full set of variables. At each step, variables whose do not verify the condition of significance given by the statistical criterion.
  - **Forwards selection (FS):** begins with an empty set of variables. At each step, we select from the variable list that variable which is most significant.
  - **Hybrid selection (HS):** third procedure which is a hybrid of both procedures incorporating ideas of each of them. Hybrid stepwise procedures alternates backwards and forwards in its model selection and stops when all variables have either been retained for inclusion or removed.

  However, stepwise methods are sensitive to collinearity among predictors. Furthermore, there is no guarantee that the optimal solution is achieved (Izenman, 2009).

- **Ridge regression (RR):** this type of multiple regression model adds a small constant value k to the diagonal entries of the matrix $(X^T X)$ before taking its inverse. Then, multicollinearity problems are avoided in this type of regression, but certain bias in coefficient estimation is introduced by the addition of k (Hoerl & Kennard, 1970).

- **Lasso regression (LR):** lasso (least absolute shrinkage and selection operator) regression is a type of regularized regression. It combines the best properties of ridge regression and variable selection: subset selection, shrinkage to improve prediction accuracy, and stability in the face of data perturbations (Tibshirani, 1996).
- **Partial least-squares regression (PLSR):** partial least-squares regression (PLSR) deals with the determination of some latent variables which retain most of the information in the X variables so as to predict Y, while reducing the dimensionality of the regression. PLSR is usually obtained through an algorithm, rather than as result of an optimization procedure. PLSR allow tackling multicollinearity and works well even when the number of predictors is high related to the number of observations (Wold, 1966).

5.3. Methods

GPD metric was calculated in the 42 real multi-source biomedical repositories; then, there were 42 GPD values available to carry out the MR analyses. Given the relationship among GPD and SPO values (GPD is a weighted combination of SPO values), conclusions extracted from analyzing GPD could be extrapolated to SPO.

The dimensionality reduction technique used was PCA, based on the dissertations exposed in the previous chapter (speed, parameter free, acceptable accuracy). Given the findings in the same chapters as well as the characteristics of the repositories, the dimension of reduction was set in one (the first principal component). Likewise, the number of support points taken for the evaluation of the estimated continuous PDFs was high enough so as to avoid late convergence problems.

There were considered five different predictor variables for the evaluation of dependences in GPD metric:

- Number of data.
- Number of sources.
- Number of data of the smallest source.
- Number of variables.
- Categorical variables ratio.

This last predictor is defined this way:

$$c_R = \frac{n_c}{n_t} \tag{5.2}$$

being c_R the categorical variables ratio, n_c the number of categorical variables in the repository and n_t the total number of variables in the repository.

The values of these explanatory variables were normalized via z-score, previously to the initiation of any analysis. Furthermore, it has to be mentioned that it was introduced an intercept term in all the models.

Quality of the data was assessed so as to check if the assumptions of each one of the MR models considered were verified, and also in order to avoid influences introduced by extreme values.

Primarily, outliers (i.e. extreme values in the predicted variable) were detected. The range of acceptance was set in three median absolute deviations (MAD) away from the median. MAD is calculated from this expression:

$$MAD = median\left(\left|Y_i - \tilde{Y}\right|\right), for\ i = 1,2,\dots,N \tag{6.3}$$

where Y refers to a vector made up of N scalar observations, and $\tilde{Y}$ its median respect its components.

After that, influential observations (i.e. extreme values in the predictors variables) were assessed. The rage of acceptance was set in three times the mean Cook's distance (CD). CD measures the normalized variation in the vector of coefficients consequence of the deletion of an observation. It is obtained this way:

$$CD_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)}{\rho \cdot MSE} \tag{6.4}$$

Expression in which $\hat{Y}_j$ is the jth fitted response value, $\hat{Y}_{j(i)}$ is the jth fitted response value when the fit does not include observation i, MSE is the mean squared error and ρ is the number of coefficients in the regression model.

After performing this initial quality data assessment, those values which were categorized as outliers or influential observations were removed from the analysis.

Concerning to multicollinearity, its presence was studied by means of the calculation of the condition number (k) of the predictors matrix (taking as p-norm the spectral norm), combined with the extraction of the Variance Inflation Factor (VIF), which quantifies the increased degree of dispersion (measured in variance terms) of an estimated regression coefficient due to the presence of collinearity. The threshold established in k value for multicollinearity (i.e. if k is higher than the threshold then it will be considered the presence of collinearity among predictors) was set in 15 (Williams, 2015). Referring to the VIF of each one of the coefficients, the threshold was set in 10, according to (Williams, 2015).

Then, there were carried out a total of 1000 bootstrap samplings for the GPD values in order to obtain a considerable number of estimations which allow us to determine confidence intervals. It has to be recalled here that bootstrapping is just a statistical technique consisting on random sampling with replacement, and whose size is constant and equal to the original data set. Its main advantage is its simplicity. Although for most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality.

Afterwards, the results of each one of these random extractions with repetition was used to estimate multiple regression models, concretely, those ones explained in the previous section: Stepwise regression (SW), Ridge regression (RR), Lasso regression (LR) and Partial least-squares regression (PLS).

After performing each one of the MR analyses, the remaining assumptions in multiple liner regression were verified. In order to assess the normality of residuals there were conducted Saphiro-Wilk tests. Referring to the homoscedasticity of the residuals, it was checked recurring to a Breusch-Pagan test. Finally, the independence of errors was studied via a Durbin-Watson test. P-values of each one of these validation analyses were extracted, along with its confidence intervals provided by bootstrap.

Moreover, it has to be remarked that the Mean Squared Error (MSE), as well as the Mean Absolute Error (MAE), was obtained, for each one of the bootstrap extractions and model. MSE and MAE were calculated following these expressions:

$$MSE = \frac{1}{N}\sum_{i=1}^{n} r_i^2 \tag{5.5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n} |r_i| \tag{5.6}$$

where N refers to the number of observations while r_i denotes the residuals, which are defined this way:

$$r_i = \hat{Y}_i - Y_i \tag{5.7}$$

formula in which $\hat{Y}_i$ represents the ith predicted value of the model, while Y_i is the real value corresponding to observation i.

Then, it was extracted the adjusted coefficient of determination (R^2_{adj}), which can be calculated from this expression:

$$R^2_{adj} = 1 - \left(\frac{N-1}{N-\rho}\right) \cdot \frac{SS_{res}}{SS_{tot}} \tag{5.8}$$

here N is the number of instance, ρ the number of coefficients of the regression model, SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares.

After that, the F ratio was calculated, and the associated F test was carried out in order to determine if each of the models were or not explicative. F ratio was calculated following this expression:

$$F_r = \frac{SS_{exp} \cdot df_{res}}{SS_{res} \cdot df_{exp}} \tag{5.9}$$

where SS_{exp} refers to the sum of squares explained by the multiple regression model, SS_{res} is the sum of squares of residuals, df_{res} are the residual degrees of freedom and df_{res} are the degrees of freedom associated with the regression model.

Finally, it was conducted a t-test so as to determine the significance of each factor in each regression model, given the different sets of coefficients obtained via bootstrapping. The t-value was calculated from this quotient, which served for the extraction of its inherent p-value:

$$t_{k\mu} = \frac{b_k}{s_e} \tag{5.10}$$

expression where k is the index related with the kth coefficient, μ the corresponding degrees of freedom, b_k is the estimated regression coefficient for predictor k and s_e is the standard error in the estimation of b_k.

5.4. Results

Table 5.1 offer information about data analyses results. Any outlier was detected in GPD values; however, there were some influential observations, concretely one referring to repository 23, which were influential for the construction of all the models (as was mentioned in Methods section, influential observations were deleted before performing bootstrap).

The value of the condition number of the predictors matrix was lower respect the threshold established for collinearity. Furthermore, VIF was lower than the threshold of collinearity defined for this parameter in all the coefficients. Therefore, it is concluded that there were no problems related with multicollinearity in this work.

Table 5.1. Prior data quality assessment.

Outliers	Influential observations	Condition number (k)	Pre-set factors	Variance inflation factor (VIF)
No	Yes	4,465326	**Number of data**	2,113145
			Number of sources	1,113166
			Size of the smallest source	2,183788
			Number of variables	1,179721
			Categorical variables ratio	1,220176

Tables 5.2 and 5.3 show information about the performance of the different models. Confidence intervals to 90 per cent are provided, for the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) in Table 5.2, while those respective to the adjusted coefficient of determination (Radj) and statistical significance of the model (p-value) are offered in table 5.3.

Table 5.2. Mean squared error (MSE) and Mean absolute error (MAE) for each one of the models.

		MSE			MAE	
	LB	**Mean**	**UB**	**LB**	**Mean**	**UB**
Stepwise regression (SR)	0,022584	0,038414	0,055806	0,108501	0,149990	0,189559
Ridge regression (RR)	0,021143	0,036097	0,050768	0,106663	0,146701	0,185953
Lasso regression (LR)	0,021135	0,036085	0,050763	0,106563	0,146569	0,185922
Partial least-squares regression (PLSR)	0,021135	0,036085	0,050763	0,106564	0,146567	0,185921
LB: lower bound (5%), UB: upper bound (95%).						

Table 5.3. Adjusted coefficient of determination (Radj) and significance (p-value) for each one of the models.

		Radj			p-value	
	LB	**Mean**	**UB**	**LB**	**Mean**	**UB**
Stepwise regression (SR)	-0,030223	0,245019	0,509761	0,000002	0,011613	0,064332
Ridge regression (RR)	0,070603	0,293208	0,521389	0,000004	0,029986	0,170986
Lasso regression (LR)	0,070694	0,293417	0,521604	0,000003	0,028474	0,158976
Partial least-squares regression (PLSR)	0,070694	0,293417	0,521604	0,000003	0,028448	0,158812
LB: Lower bound (5%), UB: Upper bound (95%).						

It can be observed in these tables that the model with best performance is Partial Least-Squares (PLS), although differences between this model and Lasso model are minimum. However, Stepwise regression is the model which offers greater certainty concerning the rejection of the null hypothesis (i.e. refusing that all the coefficients in the model are equal to zero). Furthermore, it has to be highlighted that in general confidence intervals are wide; it can be inferred from this circumstance that the prediction is highly influenced by the bootstrap extraction considered. Hence, high predictive values may be derived from an overfitting phenomenon, which supports the inexistence of any underlying structure in the population of GPD values which could be detected using these models.

Table 5.4 represents the behavior of the remaining assumptions. It is suggested that homoscedasticity and independence of errors are verified in the experiments. Nevertheless, the assumption of normality of residuals seems to be rejected.

Table 5.4. Posterior data quality assessment (p-values).

		NR			HC			IE	
	LB	Mean	UB	LB	Mean	UB	LB	Mean	UB
Stepwise regression (SR)	0,0001	0,0391	0,2266	0,0082	0,4455	0,9464	0,0324	0,4813	0,9497
Ridge regression (RR)	0	0,0271	0,1357	0,0108	0,4155	0,9122	0,0382	0,4901	0,9472
Lasso regression (LR)	0	0,0287	0,1559	0,0107	0,4139	0,9096	0,0403	0,4901	0,9491
Partial least-squares regression (PLSR)	0	0,0287	0,1561	0,0107	0,4138	0,9095	0,0403	0,4901	0,9490
NR: Normality of residuals, HC: homoscedasticity, IE: independence of errors									

Tables 5.5 and 5.6 provide the results concerning to the analyses of significance of each one of the pre-set factors considered. That variable which seems to present higher influence is the intercept, and its value may indicate the mean value of GPD in the population of repositories, since other coefficient values show lack of statistical significance. Only the number of variables has lower p-values respect other pre-set factors; however, its confidence intervals are wide.

Table 5.5. Regression coefficients significance (p-values) for Stepwise regression (SR) and Ridge Regression (RR).

		SR			RR	
Predictor	**LB**	**Mean**	**UB**	**LB**	**Mean**	**UB**
Intercept	0,037302	0,061959	0,111049	0,026685	0,045448	0,085636
Number of data	0,399912	0,468328	0,498238	0,367213	0,449136	0,497106
Number of sources	0,367988	0,428959	0,483090	0,366244	0,431920	0,484506
Size of the smallest source	0,375686	0,460952	0,498198	0,354549	0,441409	0,496445
Number of variables	0,001159	0,144064	0,370080	0,001836	0,163953	0,382426
Categorical variables ratio	0,354614	0,435742	0,493759	0,360624	0,438579	0,494722

Table 5.6. Regression coefficients significance (p-values) for Stepwise regression (SR) and Ridge Regression (RR).

		LR			PLSR	
Predictor	**LB**	**Mean**	**UB**	**LB**	**Mean**	**UB**
Intercept	0,026972	0,046183	0,087160	0,026977	0,046195	0,087181
Number of data	0,358104	0,445888	0,496976	0,359537	0,446301	0,496996
Number of sources	0,363519	0,431344	0,484858	0,363486	0,431328	0,484834
Size of the smallest source	0,335121	0,437572	0,495849	0,336962	0,437980	0,495917
Number of variables	0,001743	0,162485	0,382364	0,001743	0,162466	0,382295
Categorical variables ratio	0,358658	0,438416	0,494467	0,358623	0,438405	0,494439

Table 5.7 and 5.8 represent regressors coefficients obtained in this work. It can be observed that they are small in general. Only those related with intercept are higher, in concordance to the significance values commented in the previous table.

Table 5.7. Regression coefficients values for Stepwise regression (SR) and Ridge Regression (RR).

		SR			RR	
Predictor	**LB**	**Mean**	**UB**	**LB**	**Mean**	**UB**
Intercept	0,51166	0,66027	0,75641	0,51453	0,65335	0,73653
Number of data	-0,04942	0,00513	0,07131	-0,09182	0,00278	0,09329
Number of sources	-0,12364	-0,06426	-0,00550	-0,12659	-0,06105	0,00389
Size of the smallest source	-0,04410	0,00696	0,08384	-0,09744	-0,00208	0,08544
Number of variables	-1,06184	-0,43160	-0,10796	-0,97798	-0,38800	-0,09449
Categorical variables ratio	-0,00688	0,06129	0,14607	-0,01760	0,05684	0,13975

Table 5.8. Regression coefficients values for Lasso regression (LR) and Partial least squares regression (PLSR).

		LR			PLSR	
Predictor	**LB**	**Mean**	**UB**	**LB**	**Mean**	**UB**
Intercept	0,51228	0,65246	0,73681	0,51226	0,65245	0,73682
Number of data	-0,10056	0,00211	0,09720	-0,10064	0,00209	0,09725
Number of sources	-0,12899	-0,06179	0,00600	-0,12903	-0,06180	0,00604
Size of the smallest source	-0,10263	-0,00185	0,09057	-0,10271	-0,00184	0,09069
Number of variables	-0,98534	-0,39238	-0,09372	-0,98538	-0,39242	-0,09376
Categorical variables ratio	-0,01897	0,05721	0,14311	-0,01899	0,05721	0,14315

5.5. Discussion

5.5.1. Significance

Results of this work suggest the inexistence of influences of pre-set repository factors over GPD values for any of the models considered, taking into account the considerable number of analysis conducted thanks to bootstrap. Only the intercept of the models may present some influence over metrics values. However, its interpretation could be that GPD values use not to be 0 in the population of GPD values (which, in fact, could be plausible), since remaining terms are not significant.

Referring to SPO, similar conclusions can be extracted. Since GPD is a weighted combination of the multiple SPOs within a repository, it is also suggested in this work the absence of influence of the pre-set factors considered over SPO values, with the same considerations as GPD.

Implications of these findings support the use of GPD and SPO metrics for assessing multi-source variability, since they do not present biases related to characteristics of the repositories which should influence the comparison of similarity among data distributions.

5.5.2. Limitations

There are several limitations in this work. On the one hand, it has to be mentioned the number of repositories considered; although a representative sample combined with bootstrap processes could lead to general conclusions, if the size of the available sample is greater, conclusions would be more reliable.

On the other hand, a group of multiple regression models have been evaluated so as to determine undesired effects in data source variability metrics introduced by pre-set factors. Despite the justification of its choice, there is no guarantee that other models will not find relationships, although these dependences would require attention in order to determine if there is some kind of causality behind them, or are just spurious.

Finally, with the aim of avoiding possible influences introduced by the curse of dimensionality, the number of factors considered has had to be limited. However, is not ensured that other factors, not taken into account in this work, do not present influences over the values of GPD and SPO.

5.5.3. Future work

Further work in this area may include considering other multiple regression linear models (e.g. Elastic Net Regression (ENR), Least-Angle Regression (LARS)), as well as nonlinear models.

In addition, collect more repositories will constitute a priority so as to dispose of more samples to carry out these regression analyses. A greater number of repositories would also allow the introduction of interactions among variables in the regressors matrix without falling in incorrectness due to the curse of dimensionality, as well as the evaluation of additional pre-set factors.

5.6. Conclusions

A formal multivariate statistical analysis over GPD (and consequently, over SPO) has been conducted in this chapter. There have been considered four multiple regression models: Stepwise regression (SR), Ridge regression (RR), Lasso regression (LR) and Partial least-squares regression (PLSR), combined with different sets of bootstrap extractions. Global statistical tests and tests over individual regressors have also been performed. Data quality and assumptions of homoscedasticity, independence and normality were also evaluated. Results extracted from this chapter suggest the inexistence of undesired biases in GPD and SPO metrics introduced by pre-set factors of repositories.

Chapter 6

Discovering data source stability patterns

6.1. Introduction

Assessing data source variability taking into account the peculiarities of biomedical repositories justified the definition of a novel theoretical framework by (Sáez, Robles, & García-Gómez, 2017), which was presented in Chapter 2. However, as explained in the same chapter, preprocessing techniques must be applied so as to implement that theoretical framework. Luckily, the study conducted in Chapter 4 offered a dissertation about possible undesired phenomena, while providing guidelines and procedures to overcome them. After that, dependences between a set of factors and GPD values (which conclusions can be extrapolated to SPO values) were analyzed, showing no clear influence of any of them. Therefore, it is now possible to carry out a variability analysis free of undesired effects whose conclusions are reliable.

This chapter deals with the assessment of data source variability in real multi-source biomedical repositories. In order to face this task, it will be developed a new clustering algorithm, which will gather data distributions according to its similarity. Furthermore, the evaluation of this novel procedure, over the 42 collected real multi-source biomedical repositories, will lead to the discover of four main data source stability patterns in biomedical repositories: the Global stability pattern (GSP), the Local stability pattern (LSP), the Sparse stability pattern (SSP) and the Instability pattern (IP). This clustering algorithm, as well as the patterns found, were presented as oral presentation in the 30th IEEE International Symposium on Computer-Based Medical Systems (Ferri-Borredà, Sáez, & García-Gómez, 2017).

6.2. Methods

Evaluating the multi-source stability of a repository implies the comparison of the information shared among its data sources. The more information they share, the more similar they are (in terms of data distribution). Based on this approach it will be introduced a set of concepts that will allow to justify the later steps followed in this work:

- **Similarity (S):** information shared between two different data sources:

$$S = I_A \cap I_B \tag{6.1}$$

here I_A represents the information associated with source A, while I_B refers to the information offered by source B. Notice it is not provided a definition about what information is for us. Similarity is a relative concept, which always needs at least two elements, so it seems reasonable not to define an information function for evaluating it; depending on application it will be suitable to select one function or other, as when comparing two physical objects we can choose among several properties (color, dimensions, density, rugosity, etc) that gives us information about how similar these objects are respect to each other.

Moreover, it will be considered that similarity is bounded between 0 and 1, that its, a similarity of 0 means that the information provided by two data sources is completely different, while a similarity of 1 means that the information offered by two data sources is the same.

- **Disparity (D):** information not shared between two different data sources:

$$D = 1 - S \tag{6.2}$$

Although S and D have been defined for two sources, they can be extended to more sources:

$$S_N = I_{A1} \cap I_{A2} \cap I_{A3} \ldots \cap I_{A4} \tag{6.3}$$

$$D_N = 1 - S_N \tag{6.4}$$

Here N represents the number of sources, and I_{Ai} the information associated with source i.

- **Minimum pairwise similarity (δ):** minimum similarity between two different data sources so that they can be considered similar to each other. Its value ranges from 0 (null similarity) to 1 (absolute similarity).
- **Maximum pairwise disparity (Δ):** maximum disparity allowed between two different data sources before that they can be considered disparate to each other. Its value ranges from 0 (null disparity allowed) to 1 (high disparity allowed).
- **Minimum global similarity (ε):** minimum similarity among the distributions of a group of sources so that they can be considered similar among them. Its value ranges from 0 (null similarity) to 1 (absolute similarity).

Therefore, these last three parameters can be understood as thresholds defined by analysts performing a comparison among the similarity of different data sources.

Back to the objectives of this chapter, it is pursued the assessment of the degree of homogeneity of statistical distributions among data sources (i.e. its multi-source stability). Hence, it be will taken as information function for evaluating S (or D) the probability distribution function of each source. Taking into account that probability distribution functions can be represented as points of a statistical manifold it will be defined some parameters in order to translate the previous concepts in such statistical manifold:

- **Maximum similarity distance (r_δ):** maximum distance between two points associated each one to a data source so that they can be considered similar to each other.
- **Disparity distance (r_Δ):** maximum distance between two points associated each one to a data source before they can be considered as disparate to each other.
- **Maximum coherence distance (r_ε):** maximum distance between two points within a group of points so that all data sources belonging to the group could be considered similar among them.

The concepts and parameters defined above lead to the description of the proposed procedure to characterize patterns of source stability in a IDR. It comprises the following steps:

1) PARAMETER SELECTION

The user chooses the value of δ, Δ and ε according to the similarity/disparity bounds he wants to establish. Higher values of δ and ε with lower values of Δ provide more restrictive similarity analyses, while higher values of Δ with lower values of δ and ε allow more permissive analyses.

In this work, with the aim of disposing a range of results with reasonable parameters selections, there were taken different combinations of δ (from 0.85 to 0.975 in steps of 0.025), Δ (from 0.4 to 0.8 in steps of 0.05) and ε (from 0.6 to 0.9 in steps of 0.05).

2) ESTIMATION OF THE PDFs FOR EACH SOURCE IN THE IDR

Due to the high dimensionality of some repositories, and taking into account the findings of Chapter 4, dimensionality of the repositories was reduced. Given the dissertation hold in the same chapter, Principal Component Analysis (PCA) was chosen as dimensionality reduction selection method. It has to be remarked that the number of principal components taken was set in one, with the aim of obtaining accurate PDFs estimations (see Chapter 4) but also so as to provide a clear visualization of the results. Concerning to PDFs estimation, they were obtained using KDE with optimum bandwidth selection (Silverman, 1986). It has to be highlighted that the number of support points taken for the evaluation of the continuous estimated probability distribution functions was high enough so as to avoid deceiving samplings, considering the findings and procedures in Chapter 4.

3) CALCULUS OF THE PAIRWISE JSD DISTANCES AMONG ESTIMATED DATA SOURCE DISTRIBUTIONS

It has to be recalled that given the bounds of the Jensen-Shannon distance, the maximum distance between two sources is one. Hence, the following relationships are verified in this work:

$$r_{\delta} = 1 - \delta \tag{6.5}$$

$$r_{\Delta} = \Delta \tag{6.6}$$

$$r_{\varepsilon} = 1 - \varepsilon \tag{6.7}$$

4) CALCULUS OF THE SIMPLEX COORDINATES USING FULL-MULTIDIMENSIONAL SCALING

5) CALCULUS OF SPO METRIC

6) PROJECTION OF THE FIRST TWO SIMPLEX COORDINATES (THE TWO MOST IMPORTANT COORDINATES IN TERMS OF PRESERVING THE REAL DISTANCES AMONG PDFS) FOR EACH SOURCE IN ORDER TO ALLOW 2D VISUALIZATION

7) DBSCAN CLUSTER ANALYSIS TAKING THE PROJECTED SIMPLEX COORDINATES

DBSCAN (Density-based spatial clustering of applications with noise) (Ester, Kriegel, Sander, & Xu, 1996) is a clustering algorithm which classifies points into three categories: core points, reachable points and outliers, given the values of the two parameters required for its running: the minimum number of points required to form a cluster and the radius of evaluation.

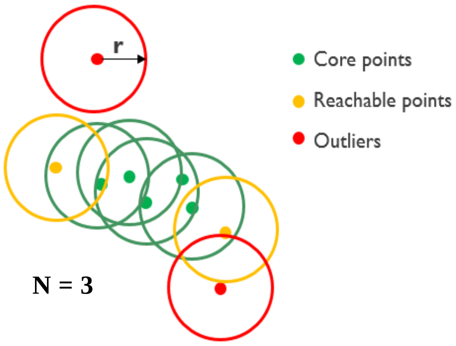


Figure 6.1. Example of a cluster analysis performed by DBSCAN.

A certain point will constitute a core point if it is possible to find within the area of evaluation defined by its radius of evaluation (r) at least N points (being N the minimum number of points required to form a cluster). Reachable points are points that are not core points but the can be found within the area defined by a core point, while outliers are those points which are not reachable from any core point. Therefore, core points form clusters with those points that are reachable from them, being those points also core points or reachable, while outliers do not form part of any cluster.

Based on the theoretical framework in this section, along with DBSCAN cluster algorithm, it has been developed a new clustering algorithm for discovering data source stability patterns in biomedical repositories. This algorithm relies on carrying out three different DBSCAN analyses:

- **1st DBSCAN analysis:** takes as radius of evaluation the Maximum similarity distance (r_δ) and as minim number of points required to form a cluster 2. It is focused on the detection of local similarities among between different data sources.
- **2nd DBSCAN analysis:** uses as radius of evaluation the Maximum coherence distance (r_ε) and the number of repository data sources as minimum number of points required to form a cluster. It is focused on the detection of global similarities among different data sources.
- **3rd DBSCAN analysis:** takes as radius of evaluation the Disparity distance (r_Δ) and the number of repository data sources as minimum number of points required to form a cluster. It is focused on the detection of global disparities among different data sources.

6.3. Results

The systematic application of the methods explained in the previous section to the 42 repositories, with the different parameter combinations, has led to the discovery of four main data source stability patterns. Next, each of these patterns is described. The descriptions are supported with the example shown in Figure 6.2, where an example of each pattern is shown (one per row). Note that with the aim of facilitating the interpretation, we have selected repositories with few sources, or using a small subset of sources within a repository:

- **Global stability pattern (GSP):** repositories offer great multi-source stability among its sources, letting data to be treated as a whole.

  Given a source belonging to a repository showing GSP, it is impossible to find other source in the repository whose disparity is higher than $1 - \varepsilon$.

  An example of this type of repository, concretely corresponding to repository 23, is showed in the first row of Figure 6.2. It can be appreciated a high similarity among its PDFs (a). Furthermore, even there exist some differences, SPO metric is low for all data sources (b). The distribution of PDFs distances concentrates in values that are lower respect r_Δ and r_ε (c). Likewise, GSP can be identified performing the 2nd cluster analysis with DBSCAN; if the repository shows GSP then all sources constitute core points in this analysis (d).

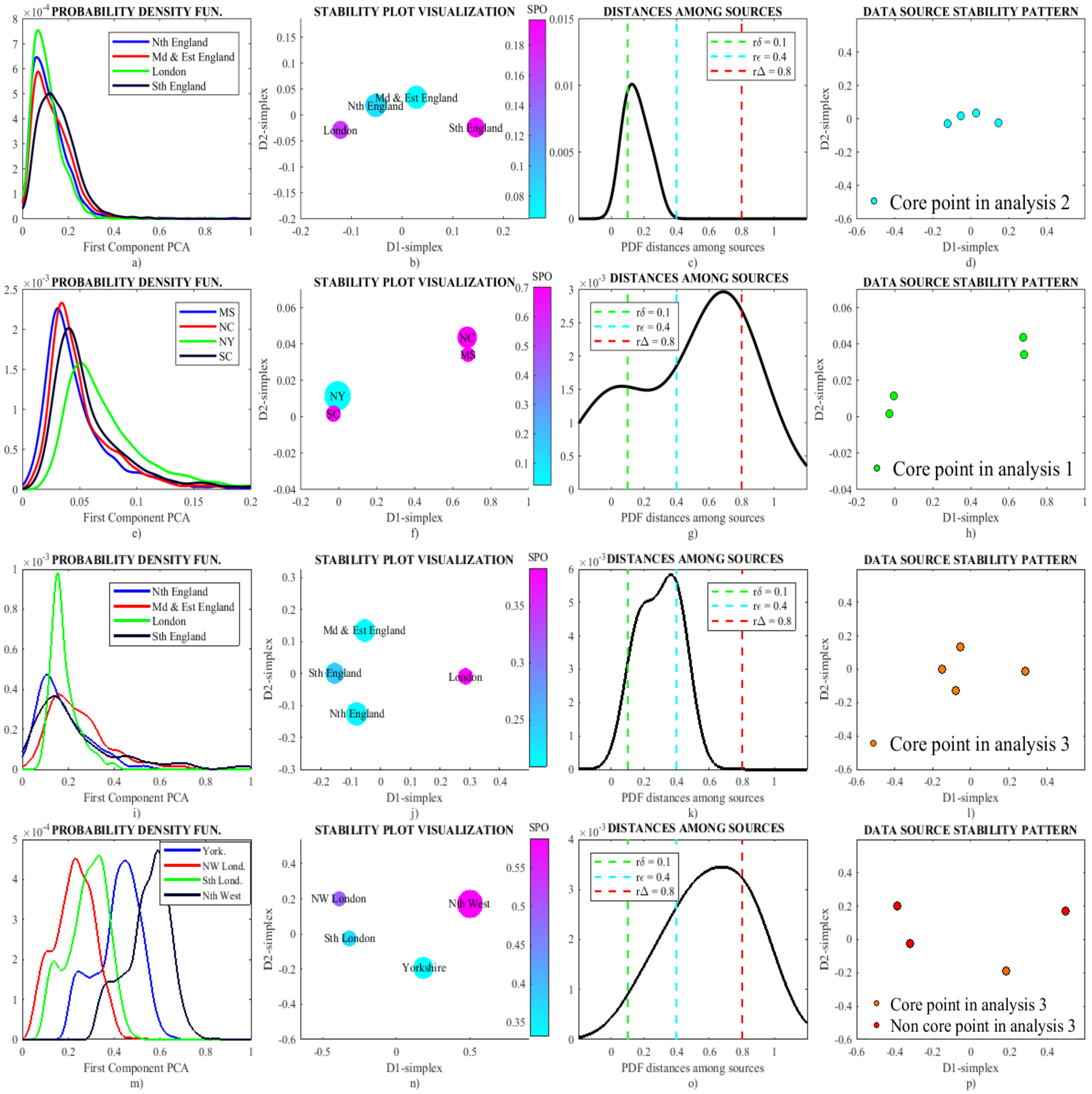


Figure 6.2. Patterns found in this study, showed through four examples of biomedical repositories presenting them: Global stability pattern (GSP), Local stability pattern (LSP), Sparse stability pattern (SSP) and Instability pattern (IP). The values of parameters in this experiment were δ = 0.9, Δ = 0.8 and ε = 0.6.

- **Local stability pattern (LSP):** repositories present one or more groups of sources whose data distributions are very similar, so data can be treated as a whole within each group, but it may not be appropriate to combine data of different groups.

  We state that a certain repository shows LSP if we can find at least two sources whose similarity is equal to or lower than δ, at least two sources whose disparity is greater than $1 - \varepsilon$ and impossible to find two sources whose disparity is higher than Δ.

  In the second row of Figure 6.2 it can be appreciated an example of repository showing LSP, concretely corresponding to repository 3. We can see there are some overlaying data sources (e), defining a cluster of sources whose data may be treated as a whole i.e. a group of sources which high data source stability among them. This fact is reflected on the stability plot, showing the respective points at close distances (f). The distribution of PDFs distances shows two peaks in this example, the first associated with intra cluster distances and the second with inter cluster distances, but this distribution may vary in other repositories presenting LSP (g). Repositories offering LSP present at least one core point in analysis 1 and one non-core point in analysis 2, but any non-core point in analysis 3 (h).
- **Sparse stability pattern (SSP):** repositories do not offer great multi-source stability among any of its sources, but neither global instability, letting data to be treated as a whole depending on application.

  Given a source belonging to a repository showing SSP, it is not possible to find other source whose similarity is at least δ, but also impossible to find other source in the repository whose disparity is higher than Δ. However, it is possible to find at least two sources whose disparity is greater than $1 - \varepsilon$.

  An example of this type of repository is showed in the third row of Figure 6.2 (repository 21). It can be seen some differences among its PDF, but they are not critical (i). It can be inferred from the stability plot that any SPO is really low, but neither high (j). The distribution of PDFs distances concentrates in values between r_δ and r_Δ (k). Repositories offering SSP do not present any non-core point in the 3rd analysis, at least one non-core point in the 2nd and any core point in analysis 1 (l).
- **Instability pattern (IP):** repositories present poor multi-source stability among its sources, so it is not recommended in this type of repositories to treat data as a whole.

  Given a repository showing IP, it will always be possible to find at least two sources within the dataset whose disparity is higher than Δ.

  An example of this type of repository is illustrated in the last row of Figure 6.2 (repository 24). It can be observed that at least one source shows high disparity respect to other PDFs (m). SPO metric is high for these unstable sources (n). The distribution of PDFs distances concentrates in values that are higher than r_δ, r_Δ

and r_ε in this example, but this distribution may vary in other repositories presenting IP (o). Likewise, it can be recognized performing the 3rd DBSCAN analysis; if the repository shows IP then some sources do not constitute core points in this analysis (p).

Table 6.1 shows the results of the classification task, taking δ=0.9, Δ=0.8 and ε =0.6, as well as some suggestions so as to ensure an appropriate data reuse. It can be appreciated that under these assumptions, IP is the most common data source stability pattern among our biomedical repositories.

Table 6.1. Data source stability patterns (DSSP) of each repository (δ=0.9, Δ=0.8 and ε=0.6).

DSSP	Repositories	Frequency (%)	Suggestions for data reuse
GSP	10,22,23	7,143	Treat data as a whole.
LSP	3,35,42	7,143	Possible to treat data as a whole, but performance may be improved dividing the multi-source repositories into subsets of local source similarity.
SSP	21,26,33	7,143	Possible to treat data as a whole, but performance may be improved if working individually with each source.
IP	1,2,3,4,5,6,7,8,9,11, 12,13,14,15,16, 17,18,19,20,24, 25,27,28,29,30, 31,32,34,36,37, 38,39,40,41	78,571	Do not treat data as a whole. Divide your repository into subsets according to similarity among sources and work with them individually.

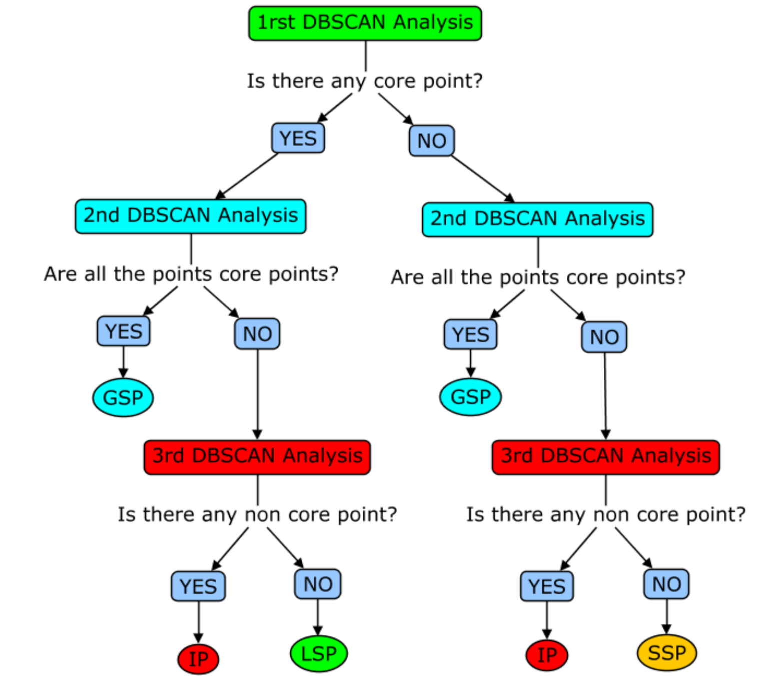


Figure 6.3. Schematization of the clustering algorithm developed.

6.4. Discussion

6.4.1. Significance

Patterns found in this work show the existence of four types of source stability in the evaluated multi-source biomedical repositories, each one requiring different managements, so treating data as a whole in these repositories could be potentially misleading. Prior to any analysis in a multi-source repository, its data source stability pattern must be identified in order to ensure a right use of data.

Regarding to the method, cluster analysis performed with DBSCAN based on points obtained by means of simplicial projections from probability distribution distances and the parameters exposed in 6.2 has been shown as a useful tool in the task of discovering data source stability patterns in this study.

First, it has been observed that simplicial projections from PDF distances keep the similarities and disparities among PDFs of each source in a repository, even when its source PDFs are complex (e.g. multimodal, multi-type), allowing reliable analysis of any repository without considering any underlying data distribution.

Second, concepts defined in 6.2 enable clustering analysis with DBSCAN based on thresholds chosen by user depending on the study or application. It has been provided some intuitive general concepts (S, D) from which it can be derived general parameters (δ, Δ and ε) whose interpretation is also intuitive, and specific parameters (r_δ, r_Δ, r_ε) designed to tackle our task with DBSCAN. However, it has to be mentioned here that it could be possible to take other clustering approaches just adapting the specific parameters and preserving the others.

6.4.2. Limitations

A limitation of this study is the number of repositories. Although a representative sample can enable an acceptable generalization, the more repositories the more credibility of the conclusions.

Other limitation of this study is the selection of only one principal component to summarize the repository variables. Although the maximum variance is conserved, dissimilarities of smaller sources, or those with less variance, may be hidden in higher components. This may be even a problem using the first component on other non-linear reduction methods.

6.4.3. Future work

The developed methodology linked the multi-source statistical manifold with a specific DBSCAN parametrization. This opens the study of generalizing this approach to other clustering algorithms.

Furthermore, evaluate the suggestions offered referring to data reuse tasks would remark the importance of detecting these data source variability patterns in biomedical repositories.

6.5. Conclusions

A method for discovering data source stability patterns in multi-source biomedical repositories is proposed in this chapter. This method has been evaluated over a group of 42 real multi-source biomedical repositories, suggesting the existence of four main data source stability patterns in biomedical repositories: the Global stability pattern (GSP), the Local stability pattern (LSP), the Sparse stability pattern (SSP) and the Instability pattern (IP). Each one of this type of repositories may require specific considerations when dealing with its data in order to maximize the knowledge extraction process and avoid misleading results.

Chapter 7

Conclusions

Reusing data of repositories which integrate data from different sources could be potentially misleading if the variability among different data distributions is not evaluated. This fact is critical in the biomedical field, where classical statistical tools may fail.

This final year project has analyzed multiple aspects regarding to GPD and SPO, novel metrics for data source variability assessment:

Firstly, it has been studied the influence of preprocessing techniques over GPD and SPO values. It has been shown that users must be careful when implementing these metrics, since sometimes adverse phenomena introduced by these techniques is present. Causes of these problematic effects have been identified and procedures to overcome them have been proposed and evaluated, showing promising results.

Secondly, once those influences were understood and tackled, a formal multivariate statistical analysis over GPD and SPO metrics was carried out. Results from this analysis suggest absence of dependence between metrics values and repository pre-set factors, showing a robust behavior.

Finally, based on the previous findings, a novel clustering algorithm to discover data source stability patterns in biomedical repositories was proposed, and its evaluation over real multi-source biomedical repositories lead to the discover of four main data source stability patterns: the Global stability patterns (GSP), the Local stability pattern (LSP), the Sparse stability pattern (SSP) and the Instability pattern (IP). This new procedure and its findings were presented as oral presentation in the 30th IEEE International Symposium on Computer-Based Medical Systems.

References

Ajuntament de València. (2017, June 27). *Datos de estaciones de contaminación atmosférica*. Retrieved from http://gobiernoabierto.valencia.es/va/data/?groups=medio-ambiente

Alpaydin, E. (2010). *Introduction to Machine Learning*. London: The MIT Press.

Aspden, P. (2004). Patient safety: achieving a new standard for care.

Bellman, R. (1961). *Adaptative Control Processes: A Guided Tour*. NJ: Princeton University Press.

Centers for Disease Control and Prevention . (2017, June 24). *Behavioral Risk Factor Data: Health-Related Quality of Life (HRQOL)* . Retrieved from https://catalog.data.gov/dataset/behavioral-risk-factor-data-health-related-quality-of-life-hrqol-76ea6

Centers for Disease Control and Prevention. (2017, February 18). *Air Quality Measures on the National Environmental Health Tracking Network*. Retrieved from https://catalog.data.gov/dataset/air-quality-measures-on-the-national-environmental-health-tracking-network

Centers for Disease Control and Prevention. (2017, June 24). *Deaths in 122 U.S. cities - 1962-2016. 122 Cities Mortality Reporting System*. Retrieved from https://catalog.data.gov/dataset/deaths-in-122-u-s-cities-1962-2016-122-cities-mortality-reporting-system

Centers for Disease Control and Prevention. (2017, June 24). *LymeDisease_9211_county*. Retrieved from https://catalog.data.gov/dataset/lymedisease-9211-county

Centers for Disease Control and Prevention. (2017, June 24). *Quitline Service Utilization - 2010 To Present*. Retrieved from https://catalog.data.gov/dataset/quitline-a-service-utilization-2010-to-present

Centers for Medicare and Medicaid Services. (2017, June 24). *ESRD QIP - Anemia Management Reporting - Payment Year 2015*. Retrieved from https://catalog.data.gov/dataset/esrd-qip-anemia-management-reporting-payment-year-2015

Centers for Medicare and Medicaid Services. (2017, June 24). *ESRD QIP - Mineral Metabolism Reporting - Payment Year 2017*. Retrieved from https://catalog.data.gov/dataset/esrd-qip-mineral-metabolism-reporting-payment-year-2016

Centers for Medicare and Medicaid Services. (2017, June 24). *ESRD QIP - Vascular Access - Payment Year 2015*. Retrieved from https://catalog.data.gov/dataset/esrd-qip-vascular-access-payment-year-2015

Centers for Medicare and Medicaid Services. (2017, June 24). *Hospital Readmissions Reduction Program*. Retrieved from https://catalog.data.gov/dataset/hospital-readmissions-reduction-program

Clarke, B., Fokoue, E., & Helen-Zhang, H. (2009). *Principles and Theory for Data Mining and Machine Learning.* Springer series in statistics.

Comunidad Autónoma de País Vasco. (2011, June 21). *Mortalidad por cáncer de colón en hombres en Euskadi (1996-2003).* Retrieved from http://datos.gob.es/es/catalogo/a16003011-mortalidad-por-cancer-de-colon-en-hombres-en-euskadi-1996-20032

Comunidad Autónoma de País Vasco. (2011, June 21). *Mortalidad por cáncer de estómago en mujeres en Euskadi (1996-2003).* Retrieved from http://datos.gob.es/es/catalogo/a16003011-mortalidad-por-cancer-de-estomago-en-mujeres-en-euskadi-1996-20032

Comunidad Autónoma de País Vasco. (2011, May 31). *Mortalidad por diabetes en mujeres en Euskadi (1996-2003).* Retrieved from http://datos.gob.es/es/catalogo/a16003011-mortalidad-por-diabetes-en-mujeres-en-euskadi-1996-20032

Comunidad Autónoma de País Vasco. (2011, May 31). *Mortalidad por enfermedad cerebrovascular en hombres en Euskadi (1996-2003).* Retrieved from http://datos.gob.es/es/catalogo/a16003011-mortalidad-por-enfermedad-cerebrovascular-en-hombres-en-euskadi-1996-20032

Comunidad Autónoma de País Vasco. (2011, June 21). *Mortalidad por suicidio en mujeres en Euskadi (1996-2003).* Retrieved from http://datos.gob.es/es/catalogo/a16003011-mortalidad-por-suicidio-en-mujeres-en-euskadi-1996-20032

De Leeuw, J. (1993). *Fitting distances by least squares.* Tech Rep No 130, Interdivisional Program in Statistics, UCLA.

Department of Health & Human Services . (2016, April 5). *Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis-Related Groups (DRG).* Retrieved from https://www.healthdata.gov/dataset/inpatient-prospective-payment-system-ipps-provider-summary-top-100-diagnosis-related-groups

Edgeworth, F. (1908). On the probable errors of frequency-constants. *Journal of the Royal Statistical Society,* 71(2), 381-397.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers,* 191-203.

Endres, D., & Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory,* 49(7), 1858-1860.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, Vol. 96, No. 34, pp. 226-231.

Eusko Jaurlaritza. (2017, June 28). *Calidad de aguas de consumo de Euskadi durante el 2016*. Retrieved from http://opendata.euskadi.eus/catalogo/-/calidad-de-aguas-de-consumo-de-euskadi-durante-el-2016/

Eusko Jaurlaritza. (2017, July 4). *Mediciones del polen en Euskadi durante el 2017*. Retrieved from http://opendata.euskadi.eus/catalogo/-/mediciones-del-polen-en-euskadi-durante-el-2017/

Ferri-Borredà, P., Sáez, C., & García-Gómez, J. M. (2017). Discovering data source stability patterns in biomedical repositories based on simplicial projections from probability distribution distances. *30th IEEE International Symposium on Computer-Based Medical Systems*.

Gene Expression Omnibus. (2017, May 15). *Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia.* Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76312

Generalitat de Catalunya. (2012, March 20). *Número de donacions per província, regió sanitària, comarca i municipi.* Retrieved from http://observatorisalut.gencat.cat/ca/indicadors_i_publicacions/publicacions/sistema_sanitari/qualitat_i_satisfaccio/seguretat/controls_de_qualitat_dels_processos_de_transfusio_/Dades_obertes/

Generalitat Valenciana. (2016, April 15). *Datos de Mortalidad 2013*. Retrieved from http://www.dadesobertes.gva.es/es/dataset/isp-mortalidad-2013

Generalitat Valenciana. (2016, November 4). *Registro de centros, servicios y establecimientos sanitarios 2016*. Retrieved from http://datos.gob.es/es/catalogo/a10002983-registro-de-centros-servicios-y-establecimientos-sanitarios-2016

Gobierno de Aragón. (2016). *Población usuaria del SAS por sexo. Municipios.* Retrieved from http://opendata.aragon.es/datos/poblacion-usuaria-del-sas-por-sexo

Gray, A. (2003). Very Fast Multivariate Kernel Density Estimation using via Computational Geometry. *Proceedings Joint Statatistics*.

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Izenman, A. (2009). *Modern multivariate statistical techniques: regression, classification, and manifold learning.* Springer Science & Business Media.

Junta de Castilla y León. (2012, March 27). *Información Polínica Histórica.* Retrieved from http://datos.gob.es/es/catalogo/a07002862-informacion-polinica-historica1

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The annals of mathematical statistics, 22(1),* 79-86.

Lee, J., & Verleysen, M. (2007). *Nonlinear dimensionality reduction.* Springer Science & Business Media.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information 37,* 145{151.

Liou, C., Huang, J., & Yang, W. (2008). Modeling word perception using the Elman network. *Neurocomputing,* Volume 71, 3150–3157 (2008).

Lord, F. (1984). Maximum likelihood and Bayesian parameter estimation in item response theory. *ETS Research Report Series.*

McMurry, A. (2013). Shrine: Enabling nationally scalable multi-site disease studies. *PLoS ONE.*

Ng, A. (2015). Sparse autoencoder. *Lecture notes.*

NHS Digital. (2014, March). *GP Earnings and Expenses.* Retrieved from https://data.gov.uk/dataset/gp_earnings_and_expenses

NHS Digital. (2015, November). *Mental Health and Learning Disabilities Statistics Data.* Retrieved from https://data.gov.uk/dataset/monthly-mental-health-minimum-dataset-reports

NHS Digital. (2016, December). *NHS Continuing Healthcare Activity.* Retrieved from https://data.gov.uk/dataset/nhs_continuing_healthcare_activity

NHS Digital. (2017, March). *CCG Prescribing Data.* Retrieved from https://data.gov.uk/dataset/ccg_prescribing_data

NHS Digital. (2017, April 1). *Numbers of Patients Registered at a GP Practice .* Retrieved from https://data.gov.uk/dataset/numbers_of_patients_registered_at_a_gp_practice

NHS Digital. (2017, May 1). *Numbers of Patients Registered at a GP Practice: by Single Year of Age.* Retrieved from https://data.gov.uk/dataset/numbers-of-patients-registered-at-a-gp-practice-single-year-of-age

NHS Digital. (2017, January). *Sickness Absence Rates in the NHS.* Retrieved from https://data.gov.uk/dataset/sickness-absence-rates-in-the-nhs

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics,* 1065-1076.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11),* 559-572.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics,* 65-78.

Sáez, C. (2016). Probabilistic methods for multi-source and temporal biomedical data quality assessment. *(Doctoral dissertation).*

Sáez, C., Robles, M., & García-Gómez, J. (2017). Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Statistical methods in medical research*.

Sheather, S., & Jones, M. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, Series B (Methodological), 683-690.

Silverman, B. (1986). Density estimation for statistics and data analysis . *CRC press*, Vol.26.

State of California. (2016, August 4). *Hospital Profitability, 2009-2013*. Retrieved from https://www.healthdata.gov/dataset/hospital-profitability-2009-2013

State of California. (2017, January 17). *Case Mix Index-1996-2015*. Retrieved from https://www.healthdata.gov/dataset/case-mix-index-1996-2015

State of California. (2017, May 21). *Emergency Department Data By Expected Payer Source*. Retrieved from https://www.healthdata.gov/dataset/emergency-department-data-expected-payer-source

State of California. (2017, May 20). *Infant Mortality, Deaths Per 1,000 Live Births (LGHC Indicator 01)*. Retrieved from https://www.healthdata.gov/dataset/infant-mortality-deaths-1000-live-births-lghc-indicator-01

State of New York. (2016, August 19). *All Payer Inpatient Potentially Preventable Complication (PPC) Individual Rates by Hospital (SPARCS): Beginning 2013*. Retrieved from https://www.healthdata.gov/dataset/all-payer-inpatient-potentially-preventable-complication-ppc-individual-rates-hospital

State of New York. (2016, December 14). *Medicaid Chronic Conditions, Inpatient Admissions and Emergency Room Visits by County: Beginning 2012*. Retrieved from https://www.healthdata.gov/dataset/medicaid-chronic-conditions-inpatient-admissions-and-emergency-room-visits-county-beginning

State of New York. (2017, June 6). *All Payer Inpatient Quality Indicators (IQI) by Hospital (SPARCS): Beginning 2009*. Retrieved from https://www.healthdata.gov/dataset/all-payer-inpatient-quality-indicators-iqi-hospital-sparcs-beginning-2009

State of New York. (2017, June 4). *All Payer Patient Safety Indicators (PSI) by Hospital: Beginning 2009*. Retrieved from https://www.healthdata.gov/dataset/all-payer-patient-safety-indicators-psi-hospital-beginning-2009

State of New York. (2017, June 22). *Child Health Plus Program Enrollment: Beginning 2009*. Retrieved from https://www.healthdata.gov/dataset/child-health-plus-program-enrollment-beginning-2009

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

UC Irvine Machine Learning Repository. (1988, July 1). *Heart Disease Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Utah Department of Health. (2015). *Vaccinations By School District And School Utah 2014*. Retrieved from https://opendata.utah.gov/Health/Vaccinations-By-School-District-And-School-Utah-20/3nnk-8ku2

Van der Maaten, L., Postma, E., & Van den Herik, H. (2009). Dimensionality Reduction: A Comparative Review. *Tilburg University Technical Report*.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems,12(4)*, 5-33.

Williams, R. (2015). Review of Multiple Regression. *University of Notre Dame*.

Wold, H. (1966). Estimation of principal components and related methods by iterative least squares. *Multivariate analysis*.

Budget

Budget contents

Chapter 8

1. Introduction

In this section, the budget required to conduct this final degree project will be presented. It is divided into three main blocks: the first one is associated with hardware costs, the second one related with software cost, while the last one refers to expenses in staff.

2. Disaggregated budget

2.1. Hardware costs

Units	Description	Details	Provider	Quantity	Unit price (€)	Total price (€)
u	HP ProBook 470 G4	Intel® Core™ i7 processor (2.7 GHz), 8 GB DDR4-2133 SDRAM, 1 TB HDD storage, Windows 10 Pro 64	HP	1	865	865
Total:						865

2.2. Software costs

Units	Description	Provider	Quantity	Duration (Years)	Unit price (€)	Total price (€)
License	Matlab R2017a	MathWorks	1	1	900	900
License	Microsoft office professional 2016	Microsoft	1	1	399,99	399,99
License	Microsoft Windows 10 professional	Microsoft	1	1	279	279
Total:						1578,99

2.3. Staff costs

Description	Tasks	Rank	Quantity (h)	Unit price (€)	Social security spending (€)	Salary (€)	Total costs (€)
Biomedical engineer	Carry out the project	Junior	600	15	2124	6876	9000
Statistician	Supervise the project	Senior	120	30	828	2772	3600
Data scientist	Supervise the project	Senior	96	30	662,4	2217,6	2880
Total:							15480

3. Total budget

Description	Cost (€)
Hardware costs	865
Software costs	1578,99
Staff costs	15480
Total:	17923,99