Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances

Carlos Sáez[1,2], Montserrat Robles[1], and Juan Miguel García-Gómez[1,3,4]

[1]Grupo de Informática Biomédica (IBIME), Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 València, Spain

[2]Center for Health Technology and Services Research (CINTESIS), Faculdade de Medicina da Universidade do Porto, Rua Dr. Plácido da Costa s/n, 4200-450, Porto, Portugal

[3]Grupo de Investigación Biomédica en Imagen (GIBI230), Instituto de Investigación Sanitaria (IIS), Hospital la Fe, Spain

[4]Unidad Mixta de Investigación en TICs aplicadas a la Reingeniería de Procesos Sociosanitarios (eRPSS), Instituto de Investigación Sanitaria del Hospital Universitario y Politécnico La Fe, Bulevar Sur S/N, Valencia 46026, Spain

Abstract

Biomedical data may be composed of individuals generated from distinct, meaningful sources. Due to possible contextual biases in the processes that generate data, there may exist an undesirable and unexpected variability among the probability distribution functions (PDFs) of the source subsamples, which, when uncontrolled, may lead to inaccurate or unreproducible research results. Classical statistical methods may have difficulties to undercover such variabilities when dealing with multi-modal, multi-type, multi-variate data. This work proposes two metrics for the analysis of stability among multiple data sources, robust to the aforementioned conditions, and defined in the context of data quality assessment. Specifically, a global probabilistic deviation (GPD) and a source probabilistic outlyingness (SPO) metrics are proposed. The first provides a bounded degree of the global multi-source variability, designed as an estimator equivalent to the notion of normalized standard deviation of PDFs. The second provides a bounded degree of the dissimilarity of each source to a latent central distribution. The metrics are based on the projection of a simplex geometrical structure constructed from the Jensen-Shannon distances among the sources PDFs. The metrics have been evaluated and demonstrated their correct behaviour on a simulated benchmark and with real multi-source biomedical data using the UCI Heart Disease dataset. The biomedical data quality assessment based on the proposed stability metrics may improve the efficiency and effectiveness of biomedical data exploitation and research.

Keywords: data variability, data quality, data reuse, probability distribution distances, information geometry

Author version of manuscript published in Statistical Methods in Medical Research (`http://smm.sagepub.com/cgi/doi/10.1177/0962280214545122`)

1 Introduction

Biomedical data may be generated from different sources. Multi-centre data repositories are a well-known example. Other examples include data generated from different users, or groups of data at different levels of granularity through a sensible hierarchy, e.g., a geographical location. Hereafter multi-source data is defined as data comprising individuals generated from distinct, meaningful, originating sources, belonging each individual to a single, clearly identified, source.

Compiling data from multiple sources may ensure a good sample representation from a broader and more representative population. In fact, obtaining a representative and significant sample is usually the objective of multi-centre studies [1].

However, due to possible contextual biases in the processes that generate data, multi-source data may also entail an unexpected or undesired variability among its sources, which can lead to contradictory or unreproducible results [1]. As a consequence, two situations may arise: 1) data consumers do not consider such variability, leading their results to poor hypotheses, models, or wrong decisions; 2) data consumers are aware about the possible variability but the complexity of data either hinders such discovery or they do not have the proper discovery methods. Regarding to 2), Sáez et al. [2] showed that classical statistical tests may have difficulties or be not suitable at all when dealing with specific data features, such as in multivariate, multi-type and multi-modal data. In any of the cases, one could perfectly draw hypothesis or obtain acceptable models from data assuming that data is stable among sources —i.e., modelling and evaluation made with data from all sources. However, it may not be assured that these results will either maintain the same effectiveness when used or evaluated at a single source or be generalisable at all to other sources.

This variability among sources is in fact a variability among their data probability distribution functions (PDFs)[1]. Ideally, biomedical research studies, such as clinical trials or population studies, would expect PDFs to be stable among the different sources in order to draw generalizable conclusions. However, if this stability is not achieved, data fail to meet user expectations what, by definition [3], results in a lack of data quality. The variability among sources has been addressed by some authors in the biomedical data quality domain [4] [5]. Nevertheless, it has mainly been related to semantic, structural or element agreement among sources. In this work, the variability among sources' PDFs is studied as a *spatial stability* data quality dimension [6]. The study of the stability of data sources may help data consumers understand their data, detect problematic or biased sources, detect patterns among the sources or, more generally, take better decisions in the research process.

In this work, a method for obtaining representative measurements of the data source stability is presented. It contributes to the state-of-the art with two metrics of a spatial stability data quality dimension, designed as a descriptive statistical method to assess multi-source variability, and being robust to the aforementioned features where classical statistical tests may not be suitable. The first metric measures the degree of global multi-source variability —i.e. global probabilistic deviation (GPD)— and the second the

[1] Note that semantic or structural consistence among data sources is not discussed here, which is out of the scope of this work.

degree of outlyingness of single sources —i.e., source probabilistic outlyingness (SPO), being both designed to be comparable among different domains or datasets. The calculus of metrics is based on the projection of a D-dimensional simplex constructed from the pairwise PDF distances among sources. Additionally, this method provides the basis for a source clustering and for the spatial visualization of data source variability. The method is evaluated with simulated and real data using the UCI Heart Disease dataset [7] [8].

The rest of the paper is organized as follows. Section 2 reviews different stability problems found in biomedical studies, the statistical methods usually employed to detect such variabilities, and settles the work in the context of data quality. Section 3 describes the simplex geometrical structure and some of their properties. Section 4 describes the spatial stability methods presented in this work. The experiments to evaluate the method and the results are described in Section 5. Finally, Sections 6 and 7 describe the discussion and conclusions of the work.

2 Background

2.1 Variability in biomedical data

The outcomes of biomedical research and healthcare practice depend on taking decisions based on the available information [4]. The data behind such information is registered by humans or devices based on observations of facts, at any stage of the healthcare process, and under an environment or context. As a consequence, the interpretation of such observations may be different according to different contexts [4]. In addition, latent contexts (e.g., the socio-economic profile of a geographical location) can have a direct influence on the original facts, independent on its interpretation. In other words, contextual biases in the processes that generate data may have associated an undesired or unexpected variability among the data-generating sources.

Many examples in the literature can be used to illustrate these types of variabilities. Markus et al. [9] found differences in the interpretation of a common dataset of Doppler embolic signals among different centres, even using the same equipments. Verwey et al. [10] and Mattson et al. [11] found diagnostic variabilities among centres in several multi-centre studies evaluating the use of cerebrospinal fluid biomarkers for Alzheimer's disease. Verwey et al. recommended the standardization of procedures and homogenization of assays to reduce such variability. Such reduction was proved by Dargaud et al. [12] in the use of a thrombin generation test in clinical trials. However, Pagiani et al. [13] encountered that even using a common acquisition protocol, differences were still found among centres in diffusion tensor magnetic resonance imaging findings. On the other hand, as a single but relevant example of how the context can cause such variations, Jarman et al. [14] showed that some hospital characteristics have a direct interaction with the ratio of hospital death rates.

According to the type or purpose of the study, detecting and measuring multi-source variabilities are generally addressed by means of classical statistical methods. In clinical trials, the coefficient of variation or, its non-parametric equivalent, the quartile coefficient of dispersion are generally used to measure variabilities among some numerical indicators obtained from each source. These methods have some possible drawbacks. Summarizing in one scalar indicator the original distribution of what is measured on each source may in some cases entail an information loss. Whilst the coefficient of variation may be affected by the scale or type of the analysed variable (e.g., a mean near 0 on a non-ratio scale), the quartile coefficient of dispersion may miss additional information about the shape of variable PDF. One advantage of the quartile coefficient of dispersion is that is unit-free, and so is comparable among different problems.

Classical statistical tests used to contrast differences among two or more univariate data samples include One-way Analysis of Variance (ANOVA) for Gaussian data, Kruskal-Wallis test for non-Gaussian data, and χ^2 test for categorical. These tests are not designed to deal with multivariate or multi-type data. In addition, both the two-sample equivalent of One-way ANOVA, the Student's t-test, and the Kruskal-Wallis test have problems with multi-modal data [2]. Though, it is also expected in ANOVA, which is suited to unimodal and homoscedastic Gaussian data.

Another method to test differences on samples composed by numerical and categorical data is the N-way ANOVA. It evaluates the effect of multiple factors, the categorical variables, on a dependent numerical variable. Hence, it is not suited to measure the variability in the joint distributions of numerical and categorical variables.

Finally, the Multivariate ANOVA (MANOVA) test is suited when having more than one dependent variable. Analogous to One-way ANOVA, variables must be numerical, Gaussian and homoscedastic. While MANOVA may be useful under these assumptions, the contrast is made on linear combinations of the variables, where such a collinearity may not exist among these.

The stability metrics developed in this work are based on information-theoretic methods to measure PDF distances. As an alternative to classical statistical tests, information-theoretic methods are able provide more information about the variability between data distributions where the assumptions of the classical tests are not met (see Section 2.3).

The method presented in this work does not intend to replace the aforementioned tests for their specific use scenarios. Its purpose is to provide a metric for the stability among different sources of data and the degree of outlyingness of single sources, being 1) suitable to multivariate, multi-type and multi-modal data, and 2) bounded and therefore comparable among different problems. Additionally, it intends to 3) pose an alternative to the classical statistical tests for those cases where the aforementioned conditions of data hinder or impede their use.

2.2 Data source stability in the context of Data Quality

The variability among sources has been addressed by several authors as a data quality problem from different perspectives. Cruz-Correia et al. [4] reviewed different issues associated to data integration and sharing among different health information systems or organizations. They found structural and semantic interoperability as the major problems. Weiskopf et al. [5] carried out a systematic review on the methods and dimensions of data quality assessment in the context of reuse of electronic health records (EHRs) for research. From a set of 95 articles they derived five high-level dimensions and seven assessment methods. From these, the *concordance* dimension, and the *data source agreement* and *distribution comparison* methods can be related to our problem. They defined concordance as *Is there agreement between elements in the EHR, or between the EHR and another data source?*. Hence, concordance can refer to the agreement among observations of a patient EHR, agreement among the same observation of a patient on different information systems, or agreement among a set of EHRs with respect to a gold standard with the same information. Whilst the last two are related to the variability among sources, only the last is related to the problem of comparing data probability distributions. Though, they identified the method of comparison with a gold-standard distribution as a method to assess the concordance dimension. However, any of the articles comprising the systematic review neither intend to provide a stability metric among a set of sources nor put attention on the heterogeneous features of biomedical data.

2.3 Dissimilarities between biomedical data distributions

Biomedical data usually show heterogeneous conditions. Concretely, biomedical data are generally 1) multivariate (i.e., data have more than one variable), 2) multi-type (i.e., simultaneously continuous, discrete ordinal and non-ordinal variables), and 3) multi-modal (i.e., data distributions are generated by more than one mode). In a previous work [2], the authors studied the behaviour of different PDF dissimilarity metrics envisaging these data features. The results of such study are summarized in Table 1.

Feature	**T**	**KW**	**KS**	**JF**	**JS**	**EMD**
Multivariate	-	-	-	Yes	Yes	Yes
Multi-Type	-	-	-	Yes	Yes	Yes
Multi-Modal	-	-	Yes	Yes	Yes	Yes
Bounded	No	No	Yes	No	Yes	Yes

Table 1: Ability of PDF distances or test statistics (columns) for dealing with specific features of data (rows 1-3) and whether the distance is bounded (row 4). T: t-test statistic, KW: Kruskal-Wallis statistic, KS: Kolmogorov-Smirnov statistic, JF: Jeffrey (Symmetric Kullback-Leibler divergence), JS: Jensen-Shannon$^{-1/2}$, EMD: Earth Mover's Distance. The '-' means that the corresponding distance is not designed for the corresponding feature.

The results showed that the aforementioned data features may complicate the application of classical statistical or data analysis methods for the assessment of differences among data samples. Specifically, the results confirmed that classical statistical tests may have difficulties on multi-modal data, or may not be not suitable at all on multivariate or multi-type data. Information-theoretic distances, including the Jeffrey and Jensen-Shannon distances, and the Earth Mover's Distance (EMD) [15] resulted the most suitable distances to all conditions. Information-theoretic are distances which derive from the Shannon's entropy theory, while EMD derives from the digital imaging field as a measure to calculate the minimum cost of transforming one histogram into another.

Regarding to the information-theoretic distances, when the probability mass in any region of the support in any of the compared distributions tends to zero, the Jeffrey distance (symmetrized version of Kullback-Leibler divergence) tends to infinite. In contrast, the Jensen-Shannon distance (JSD), square root of the Jensen-Shannon divergence [16] [17], is a metric bounded between zero and one, and it was smoothly convergent to one on that situation. In fact, such bounds facilitate the distance comparison on different problems.

On the other hand, the EMD allows setting specific costs to the flow of probability density between regions (or bins) of the support. Based on these costs, the EMD can be bounded too, however, it requires knowing a priori the bounds of the probability support of all the involved variables. In contrast, the JSD bounds are approached based on the degree of overlapping between the compared distributions (as the overlap decreases the distance tends to its upper bound), being the distance defined only by what is measured, avoiding external configurations. As a consequence, and although both EMD and JSD could be suitable for the purpose of the spatial stability method, the JSD was chosen for its direct generalization for comparability.

3 Simplices and properties

Generally speaking, a simplex is the generalization of a triangle to D dimensions, $D \in \mathbb{N}$. A D-simplex, Δ^D, is composed by $v_1, ..., v_n : n = D + 1$ vertices, which form the convex hull of the simplest polytope in R^D. Simplices can be regular or irregular. Some properties of these that will be required in the development of the stability metrics are described next.

A simplex is regular when the distances among their vertices are equal. Consequently, the length of the segment formed from the centroid of the simplex to each vertex is also equal. The angle γ between any pair of these segments depends on the number of dimensions and is [18]:

$$\gamma(D) = \arccos(\,^{-1}/_{D}) \tag{1}$$

The simplex when all the distances between its vertices are one will be defined further on as 1-regular (1R) simplex. In any D, any pair of vertices and the centroid of the simplex form a triangle. Thus, according to the *law of sines*, the distance $d(v, O) = d_{1R}(D)$ between any vertex and the centroid on 1-regular simplices in D dimensions is defined as:

$$d_{1R}(D) = \frac{1}{2\sin(^{\gamma(D)}/_{2})}, \tag{2}$$

where $d_{1R}(1) = {}^{1}/_{2}$ as a continuity convention in $D = 1$ (two vertices). See Section 1 of the Supplementary Material for details.

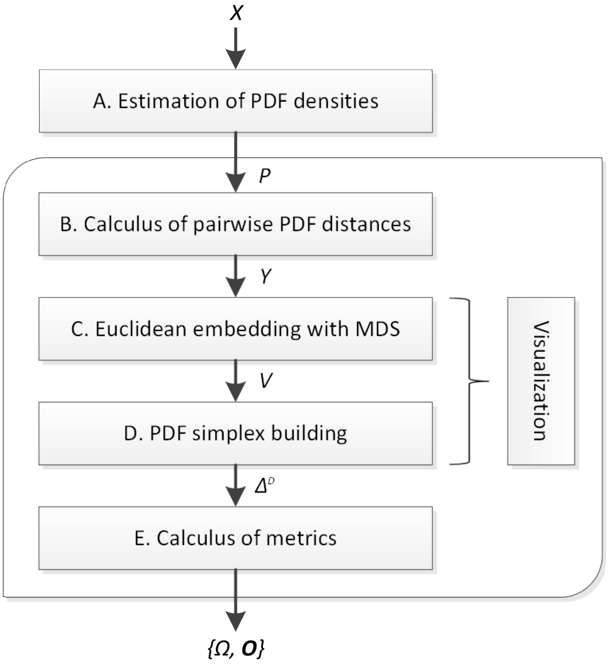
On the other hand, a simplex is irregular when at least one of its vertices is at a different distance from the centroid with respect to the others. Consequently, the distances between vertices do not have to be equal. In that case, if it is defined as a simplicial space upper-bounded by a 1-regular simplex —i.e., the simplicial space containing all the possible simplices where the maximum distance among vertices is one—, the distance of any vertex to the centroid of the irregular simplex will be bounded by:

$$d_{max}(D) = 1 - \frac{1}{D+1}, \tag{3}$$

which is larger than $d_{1R}(D)$ for the same D. See Section 2 of the Supplementary Material for details.

4 Methods

The spatial stability method provides two metrics of the data source stability: 1) the global probabilistic deviation (GPD — Ω), and 2) the source probabilistic outlyingness (SPO — $\mathbb{O}$). The GPD measures the degree of global multi-source variability. The SPO of a single source is understood as a measure of the distance of its PDF to a latent central distribution of all the sources. These metrics are obtained based on the simplex where each vertex represents a data source, and its edge lengths the pairwise PDF distance between the data of the sources represented by the adjacent vertices. A stability plot visualization of the data source stability can be derived as a by-product of the process. Figure 1 shows the procedure to obtain these outcomes. In the rest of the section, the different steps of the procedure are described. The procedure input is a multi-source dataset $X = (X_1, ..., X_S)$, where X_s is the sub-sample of data corresponding to source s and S is the total number of sources.



X
A. Estimation of PDF densities
P
B. Calculus of pairwise PDF distances
Y
C. Euclidean embedding with MDS
V
D. PDF simplex building
Δ^D
E. Calculus of metrics
Visualization
{Ω, O}

Figure 1: Steps of the method to obtain the stability metrics: global probabilistic deviation (GPD — Ω), and source probabilistic outlyingness (SPO — $\mathbb{O}$). Each step is described in its corresponding subsection in Section 4.

4.1 Estimation of PDF densities

The objective of this step is to obtain the set P of representative PDFs of the data of each data source as $P = (P_1, ..., P_S), P_s : p(X_s)$. Depending on the characteristics of the data or the problem, different preprocessing or density estimation methods may be chosen. In low dimensional problems, histograms or, a smoothing method for the numerical case, kernel density estimations [19, 20](KDE) may be used. In higher dimensional problems data can be embedded into a lower-dimensional representation using dimensionality reduction methods such as Principal Component Analysis (PCA) or non-linear manifold embeddings, such as ISOMAP [21]. As a consequence, the estimation of the probability distribution functions becomes easier, being as much of the original information conserved. Depending on the layout of data it is important to choose between linear and non-linear dimensionality reduction methods, as linear methods (such as PCA) may fail on projecting non-linear continuities of data points on the original higher-dimensional space. In addition, in the mixed multi-type case, i.e. when numerical and non-ordinal categorical variables coexist, a special density estimation may be required when histograms become noisy or sparse. In that case, a solution may be obtained using non-linear dimensionality reduction methods which allow defining a distance metric among the values of categorical data (e.g., using ISOMAP). In any case, this stage of the method is flexible to the use of different density estimation methods, thus, the selection of the proper density estimation method is out of the scope of this work.

The output of this step is the set P of PDFs with:

$$P = (P_1, ..., P_S), P_s : p(X_s) \tag{4}$$

4.2 Calculus of pairwise PDF distances

In this step the pairwise PDF distances among all sources are calculated. These distances correspond to the magnitude of the edges of the simplex under construction. Hence, being S the number of sources, the number of distances to be calculated and, therefore, the number of edges of the simplex, corresponds to the binomial coefficient $\binom{S}{2} = \frac{S!}{2!(S-2)!}$.

According to the results introduced in Section 2.3, the pairwise PDF distance $d(P_s, P_{s'})$ between PDFs P_s and $P_{s'}$ is calculated based on the square root of the Jensen-Shannon divergence as:

$$d(P_s, P_{s'}) = JSD(P_s||P_{s'})^{1/2} = \left(\frac{1}{2}KLD(P_s||M) + \frac{1}{2}KLD(P_{s'}||M)\right)^{1/2}, \tag{5}$$

where $M = \frac{1}{2}(P_s + P_{s'})$, and $KLD(P||Q)$ is the Kullback-Leibler divergence between distributions P and Q. The Jensen-Shannon divergence is defined in the $[0, 1]$ interval when using the base 2 logarithm to calculate the Kullback-Leibler divergence (Equation 6).

$$KLD(P||Q) = \sum_i \log_2 \left(\frac{P(i)}{Q(i)}\right) P(i) \tag{6}$$

The discrete Kullback-Leibler divergence in Equation 6 allows computing the non-parametric Jensen-Shannon divergence on D-dimensional histograms by computing for each bin (i) in the common support the corresponding discrete Kullback-Leibler summations. However, the Jensen-Shannon divergence can also be calculated analytically for some families of continuous distributions based on analytical forms of Kullback-Leibler divergence for d-dimensional Gaussians (Equation 7) or approximations for mixtures of Gaussians [22].

$$KLD(P||Q) = \frac{\frac{1}{2}\left(\mathrm{tr}\left(\Sigma_Q^{-1}\Sigma_P\right) + \left(\mu_Q - \mu_P\right)^\top \Sigma_Q^{-1}(\mu_Q - \mu_P) - d - \log_e\left(\frac{\det \Sigma_P}{\det \Sigma_Q}\right)\right)}{\log_e(2)} \tag{7}$$

The output of this step is the S-by-S symmetric dissimilarity matrix Y:

$$Y = (Y_{11}, ..., Y_{SS}), Y_{ss'} : d(P_s, P_{s'}) \tag{8}$$

4.3 Euclidean embedding using multidimensional scaling

The Information Geometry field states that probability distributions lie on a Riemannian manifold which inner product is given by the Fisher information metric corresponding to a specific family of distributions [23]. The geodesic distance between the points representing probability distributions in such a statistical manifold can be approximated by means of PDF distances, such as the Jensen-Shannon. In this work, only the distances among a set of distributions are known. They are not restricted to a specific family, hence, it can be considered that they lie on a statistical manifold of unknown configuration (i.e., inner product and thus dimensionality). To the purpose of this work, a simplex must be constructed from such probabilistic distances in a $\mathbb{R}^D$ space. To this end, multidimensional scaling (MDS) [24, 25] is used, which calculates an Euclidean embedding of a inter-point dissimilarity matrix.

Given a dissimilarity matrix $Y = (Y_{11}, ..., Y_{SS})$, the objective of MDS is to obtain the $V = (V_{11}, ..., V_{SD})$ coordinates of the set S of points in a $\mathbb{R}^D$ Euclidean space. This is done by finding the best approximation of $||V_s - V_{s'}|| \approx Y_{ss'}$, where $|| \cdot ||$ is the Euclidean norm between points V_s and $V_{s'}$. This approximation can be generally solved by the minimization of a loss function, such as Kruskal's Stress-1 (Equation 9) [26]. The special case of $D = S - 1$ is known as full-dimensional scaling, which solution is a D-simplex and it can be found in a unique global minima [27].

$$\text{Stress-1} = \sqrt{\frac{\sum_{ss'}(||V_s - V_{s'}|| - Y_{ss'})^2}{\sum_{ss'}(Y_{ss'})^2}} \tag{9}$$

Modern MDS methods can be classified into metric and non-metric [25]. In metric MDS the resultant inter-point distances are related to the input dissimilarities by a continuous function, while in non-metric the objective is to preserve the rank order among the dissimilarities. Both methods compute iteratively the best approximation minimizing the Stress functions, starting from an initial configuration of points. If such initialization is obtained using *classical scaling*, based on eigendecompositions, the resultant coordinates will likely be ordered monotonically by their significance with respect to the approximation.

To the purpose of the spatial stability metrics, the PDF dissimilarities should be approximated as better as possible, while maintaining the $[0, 1]$-bounds. Full-dimensional scaling provides a perfect embedding when the input dissimilarities are Euclidean, however, this is not ensured for all types of dissimilarities, such as the Jensen-Shannon distance. On the other hand, according to the Whitney Embedding theorem [28], any

M-dimensional smooth manifold can be isometrically embedded into an $\mathbb{R}^{2M}$ Euclidean space. Hence, an MDS embedding of the PDF distances in this work may lead to a perfect Euclidean embedding if the Whitney theorem holds, even if M is unknown. In fact, in all the embeddings carried out for the evaluation of this work (Section 5), a zero Stress was obtained. As a consequence, we can conclude that maintaining the $[0,1]$-bounds cannot be considered an issue in this work.

The output of this step is the S-by-D coordinates matrix V:

$$V = (V_{11}, ..., V_{SD}), \tag{10}$$

where V_{sd} is the d^{th} significant coordinate of source s.

4.4 PDF simplex building

Each of the points obtained in the previous step represents a source PDF, and the euclidean distances among them keep the corresponding pairwise PDF distance. These S points and the $\binom{S}{2}$ edges represent the vertices and edges of a D-dimensional simplex. This simplex and its centroid stand as the basis of the proposed method.

Given that the pairwise PDF distances are upper limited by one, the distances between vertices are so. It makes the corresponding simplicial projection meeting the next properties:

Property 1. *For a specific number of sources $S = D + 1$, whatever the PDF distances among them, the maximum possible simplicial projection (i.e., when the distances of all vertices to the centroid are maximum) is a D-dimensional 1R simplex.*

Property 2. *In the case of a D-dimensional 1R simplex, the maximum distance between any vertex and the simplex centroid is $d_{1R}(D)$ (Equation 2).*

Property 3. *In the general D-dimensional case (irregular simplices) the maximum distance between any vertex and the simplex centroid will be bounded by $d_{max}(D)$ (Equation 3).*

Properties 1 and 2 define the theoretical maximum inter-source dissimilarity state, thus defining an upper bound of global multi-source variability. It is straightforward that the lower bound occurs when all distributions are equal and thus all points are the same. On the other hand, in Property 3, $d_{max}(D)$ establishes the limit for the cases where $d(P_s, P_s') = 0 : s, s' \in \{1, ..., S-1\}$ and $d(P_s, P_S) = 1$ (the distance among all the sources except one is 0, and the distances between this one and the formers are 1).

The output of this step is the D-dimensional simplex Δ^D:

$$\Delta^D = (V, C), \tag{11}$$

where V correspond to the coordinates of the vertices and and C to the simplex centroid (Equation 12), both defined in $\mathbb{R}^D$.

$$C = \sum_{s=1}^{N} \frac{V_s}{N}, \tag{12}$$

4.5 Calculus of metrics

The purpose of this step is to calculate the GPD and the SPO metrics based on the simplex obtained in the previous step. This simplex represents a projection of the sources' PDFs keeping the dissimilarities among them. As a consequence, it can be affirmed that the simplex centroid may represent a latent central point with respect to all PDFs, and two definitions can be derived:

Definition 1. *The centroid C of Δ^D represents a latent central tendency of the original measured population.*

Definition 2. *The distance of a vertex V_s to the centroid C, $d(V_s, C)$, represents the deviation of a data source with respect to the central tendency of the population.*

As a consequence, the closer the PDFs vertices are to the centroid, the more stable the dataset is, while the larger the more unstable. The resultant simplex is bounded by an 1R simplex, as described in the previous step. Additionally, the larger the distance of a vertex from the centroid, the more outlying a source is with respect to the latent central tendency. The stability metrics proposed on this work are based on such definitions.

4.5.1 Global probabilistic deviation

The standard deviation is a measure of the variability of a sample with respect to its central tendency. If the sources' PDFs are considered as individuals of a population and the centroid as its central tendency, the notion of standard deviation can be directly applied to obtain a measure of the variability of the PDFs. In fact, as the PDF points are embedded in a $\mathbb{R}^D$ Euclidean space where the triangle inequality holds, their distances to the centroid can be considered as PDF distances to a latent central distribution. Hence, the derived standard deviation among S PDFs can be defined as:

$$Std(P_1, ..., P_s) = \frac{\sum_{s=1}^{S} d(V_s, C)}{S}, \tag{13}$$

where $d(V_s, C)$ is the Euclidean distance between the vertex V_s and the centroid C. Note that as distances are always positive, the resultant deviation is given in the original units.

However, despite the pairwise PDF distances are $[0, 1]$-bounded independently of the number of dimensions D, the distances between each vertex and the centroid are neither defined in the same space for different D nor $[0, 1]$-bounded. It causes the standard deviation measurement of Equation 13 neither to be comparable when the number of sources S (and therefore D) is different, nor $[0, 1]$-bounded. This situation would not permit the deviation to be comparable among different domains. Using property 2 of section 4.4, the solution comes by normalizing the standard deviation by the maximum deviation on D dimensions given the upper multi-source variability bound, i.e. $d_{1R}(D)$. In fact, that upper bound distance is the upper bound of the standard deviation on D (even if the simplex is irregular, the upper standard deviation tends to this value). That makes comparable and bounded the measurement, and leads to the definition of the GPD metric:

Definition 3. *The global probabilistic deviation metric* Ω *among a set of datasets* $X = (X_1, ..., X_S)$ *is defined as:*

$$\Omega(X_1, ..., X_S) = \frac{Std(P_1, ..., P_s)}{d_{1R}(D)} \tag{14}$$

4.5.2 Source probabilistic outlyingness

The distance $d(V_s, C)$ gives a degree of how far a source is from the central tendency of the population (Definition 2). However, as well as in the GPD metric, that distance is defined in different spaces according to D, thus making the distance neither comparable nor $[0,1]$-bounded. Analogously to the GPD metric, a normalization factor is required. In this case it is the distance between a single vertex and the centroid what must be normalized. Hence, using Property 3, the normalization factor is given by $d_{max}(D)$, leading to the definition of the outlyingness metric:

Definition 4. *The source probabilistic outlyingness metric* $\mathbb{O}$ *of a dataset* X_s *with respect to the central tendency among the datasets* $X_1, ..., X_S$ *is defined as:*

$$\mathbb{O}(X_s) = \frac{d(V_s, C)}{d_{max}(D)} \tag{15}$$

4.6 Stability plot visualization

Although the objective of this work is to provide metrics for the data spatial stability, it must be mentioned that this method also provides the means to visualize the variability or interdependences among data sources. In fact, the visualization of complex scientific datasets using aggregated data is of special research interest [29].

Concretely, the simplex coordinates calculated by MDS serve as a D-dimensional visualization of the data stability, where the d^{th} coordinate is the d^{st} important in terms of conserving the real distance. Due to the obvious restriction that visualizations can be provided up to three dimensions, the most accurate visualization is obtained taking the first two or three simplex coordinates. In the next sections some examples are provided.

5 Evaluation

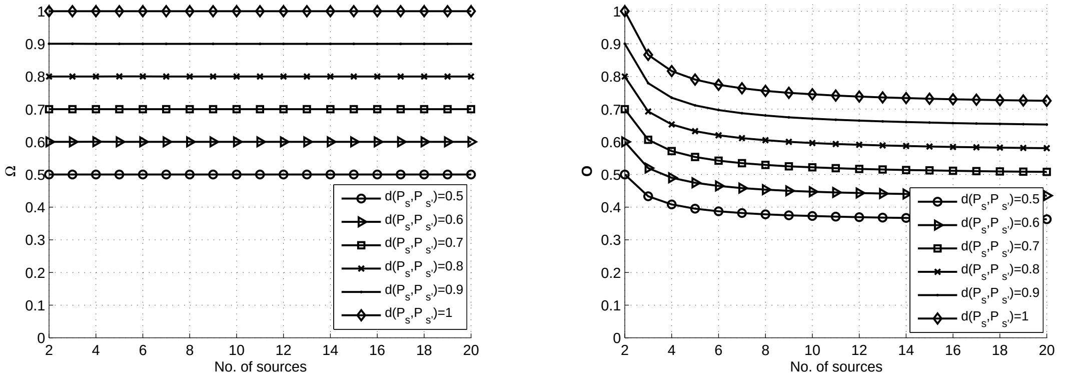
The stability metrics presented in this work have been first evaluated for scalability on different simulated conditions. Second, real multi-source biomedical data have been used with the purpose of completing the evaluation on real data variables and compare results with other classical statistical methods. In this section the evaluation experiments and their results are presented.

5.1 Evaluation of scalability

In this evaluation the GPD (Ω) and the SPO ($\mathbb{O}$) metrics were tested for scalability against variations in the number of sources, variables, and distributional dissimilarities. The GPD and SPO were measured and plotted at each iteration. Using the Jensen-Shannon distance in combination with non-parametric PDF estimations, the stability metrics are constructed to be robust against different variable types and multi-modality (as based in previous work [2]). As a consequence, to simplify the interpretation of these experiments unimodal Gaussian variables and analytical parametric Jensen-Shannon distances were used.

5.1.1 Different number of sources

New data sources were iteratively added at the same pairwise distance with respect to the previous sources. This leads to regular simplicial projections, thus, the SPO is the same for all sources at each iteration. Measurements were taken for different source pairwise distances. Results are shown in Figure 2.



(a) Global probabilistic deviation (Ω)

(b) Source probabilistic outlyingness ($\mathbb{O}$)

Figure 2: Results on different number of sources. Measurements were taken for different inter-source pairwise PDF distances, given by $d(P_s, P_{s'})$.

The GPD metric keeps stable as the number of sources increases. The effect of normalization can be observed, where the maximum GPD is one in the case all sources are at the maximum pairwise distance, i.e., one. In fact, it results as a very interesting property of the GPD metric that, due to the normalization by $d_{1R}(D)$, in the case all pairwise distances are the same the metric is equivalent to that distance.

On the other hand, the outlyingness metric shows a non-linear negative tendency which converges in all pairwise distances. As the number of sources at the same pairwise distance increases, the distance of vertices to the centroid does so until convergence. However, according to Property 3, in the case that pairwise distances are not the same among all sources, i.e. an irregular simplex, an independent source may be at a larger distance from the centroid than in the regular maximum case. Such irregular maximum corresponds to the normalization factor for outlyigness. Hence, as an expected property of the metric, when a source is at a large distance to a group of sources which are close among each other, the former will be more likely an outlier when the number of sources in the latter group increases.

5.1.2 Different number of variables

Given two multivariate Gaussian sources their number of variables is increased. The means of the first variable are at a fixed distance between the two sources, while the rest of the variables are equal (covariance matrices were diagonal with $\Sigma_{ij} = 1$). Hence, the purpose is to evaluate whether the variability caused by the first variable is maintained as new variables are included. Measurements were taken for different mean distances. Results are shown in Figure 3, where, as in the case of two sources the GPD and SPO are equivalent, only one plot series is shown representing both.

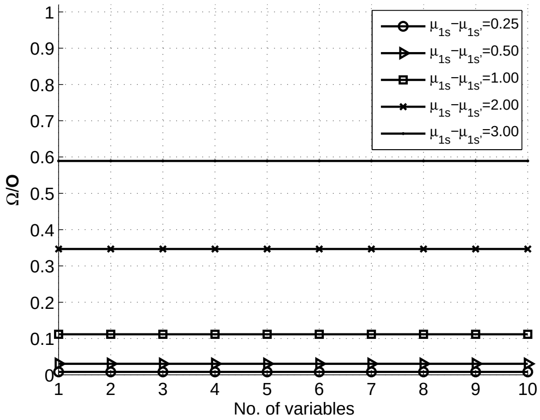


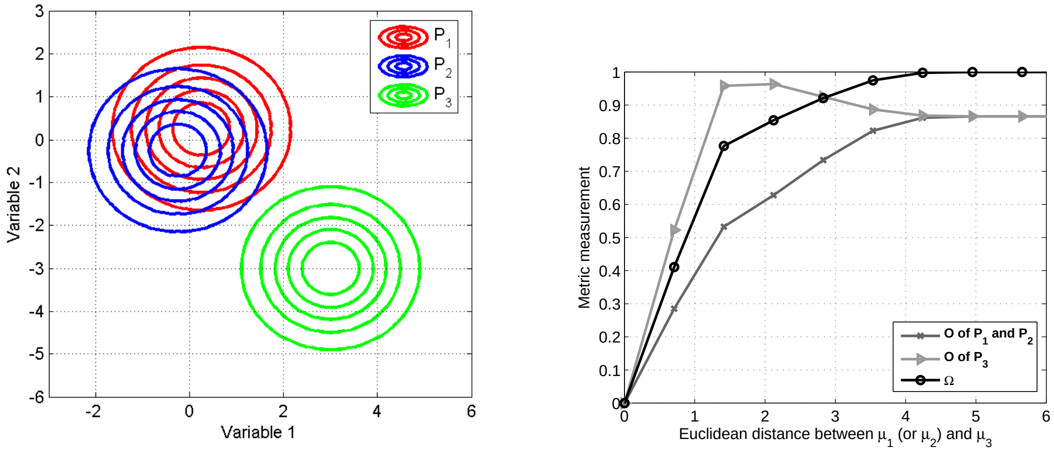
Figure 3: Results on different number of variables, where $\mu_{1s} - \mu_{1s'}$ indicates the Euclidean distance between means of the first variable of sources s and s'.

Results show the scalability of the metrics with the number of variables, as metrics keep stable as the number of variables increases. Hence, given a dissimilarity in a variable subspace, both GPD and SPO will theoretically be stable independently of the size of the full variable space.

5.1.3 Irregular source dissimilarities

In the general case differences among data sources will be irregular. That is, some sources may be close to each other, while others may show a higher outlyingness due, e.g., to sample biases. In this test this situation was evaluated. Using three bivariate Gaussian data sources with equal and diagonal covariance matrices, their means were iteratively and irregularly separated starting from an equal state until a convergence of the stability metrics. Concretely, sources 1 and 2 were smoothly separated from each other while source 3 equally separated from both with a larger velocity, expecting a larger outlyingness on it. Results are shown in Figure 4.

Figure 4(b) shows the stability metrics obtained during the iterative source separation, where figure 4(a) illustrates the PDFs in an intermediate state of the evaluation. It can be observed that as sources separate each other, the GPD does so until convergence, as well as the SPO metric of each source. Regarding to the source outlyingness, P_1 and P_2 are always at the same distance to the simplex centroid, hence showing the same



(a) Compared distributions in a intermediate iteration.

(b) GPD (Ω) and SPO ($\mathbb{O}$) of the distributions. The SPO is equivalent for distributions P_1 and P_2.

Figure 4: Results on a iterative irregular inter-source separation.

outlyingness. However, as P_3 is separated at a larger velocity it gets to large distance to the centroid which, once P_1 and P_2 have also achieved a larger probabilistic pairwise distance, is reduced. This is due to the repositioning of the simplex centroid, related to the increase of the edge length between P_1 and P_2, associated to their bounded PDF distance.

5.2 Evaluation on real data (UCI Heart Disease)

The UCI Heart Disease [7] [8] is a publicly available multi-source dataset concerning heart disease diagnosis. It contains 76 variables acquired at four different healthcare locations namely the Cleveland Clinic Foundation, OH; the Hungarian Institute of Cardiology, Budapest; the University Hospital, Zurich, Switzerland; and the V.A. Medical Center, Long Beach, CA.

Only 14 of the variables are actually used in research studies, seven numerical and seven categorical. To facilitate the evaluation of this work, data has been cleansed to remove missing data while keeping the maximum possible number of non-missing variables and individuals. This process is described in Table 2. Although in general only the Cleveland sub-dataset is used in research experiments due to its higher quality and number of individuals, in these experiments all datasets have been used with the purpose to assess the stability among all the sources.

The stability metrics have been evaluated on this dataset as follows. First they have been univariately measured, in both numerical and categorical variables, comparing the results with classical statistical univariate tests. Second, they have been measured for each combination of variables, containing pairs of numerical, categorical and mixing types. Finally, the stability metrics have been measured using all the variables.

For this evaluation, the discrete Jensen-Shannon distance (Equations 5 and 6) was used

	Original (14 variables)		**Cleansed (11 variables)**	
Source	**Individuals**	**Total missing values**	**Individuals**	**Total missing values**
Cleveland	303	6	303	0
Hungarian	294	782	261	0
Switzerland	123	284	45	0
VA	200	699	129	0
All	920	1771	738	0

Table 2: Data cleansing of the UCI Heart Disease dataset carried out in this work

as the reference PDF distance. In the case of numerical variables, their corresponding discrete PDFs were obtained from their KDE estimations using Matlab [30]. Gaussian kernels and automatic bandwidth selection [31] were used.

5.2.1 Univariate evaluation

For each variable, the GPD and SPO metrics were measured. Additionally, depending on whether the variable was numerical or categorical the classical ANOVA and χ^2 tests were performed reporting the corresponding p-values. Note that in the numerical case the ANOVA makes the assumption that variables are unimodal Gaussians, what may not be true. Results are shown in Table 3, which have been ordered by their GPD. Additionally, Figures 6, 7 and 8 show the probability distributions and 2D simplicial projections of the different variables.

	GPD (Ω)	**p-value**		**SPO ($\mathbb{O}$)**			
		ANOVA	χ^2	**Cleveland**	**Hungarian**	**Switzerland**	**V.A.**
trestbps	.1156	.3001	-	.0908	.0733	.1174	.0959
fbs	.1550	-	3e-10	.0219	.1364	.1048	.2431
exang	.2228	-	8e-13	.1768	.1871	.1609	.2031
sex	.2299	-	2e-10	.2201	.1549	.1562	.2195
cp	.2827	-	1e-16	.1895	.3016	.2563	.1759
age	.3054	6e-37	-	.0863	.4426	.1433	.3252
thalach	.3642	5e-37	-	.3897	.2019	.3497	.2480
restecg	.3709	-	2e-56	.4847	.2725	.1668	.2874
oldpeak	.4635	4e-10	-	.3377	.3912	.3924	.3925
num	.6302	2e-38	-	.4491	.6203	.5528	.4360
chol	.6737	2e-92	-	.4030	.3915	.9706	.4353

Table 3: Results of univariate evaluation on the UCI Heart Disease dataset. The variability and outlyigness measurements (columns) are shown for each variable (rows). Variables are sorted by the their GPD metric. The ANOVA or χ^2 p-value is shown according to whether the variable is numerical or categorical.

It can be observed that the GPD metric and the p-values of statistical tests are in general inversely proportional (Spearman correlation of $-.7182$, combining ANOVA and χ^2 p-values), i.e. the larger the GPD measurement the more significant the differences are found by the tests. This reinforces the consistence of the metric, which in addition shows its independence with respect to the type of variable. However, such correlation must be interpreted with caution. First, the behaviour of p-values do not need to be linear, and depends on the number of individuals or outliers (see Figure 5 for further details). As an example, the *trestbps* variable, shows a large p-value. As it can be observed (Figure 6(a)), its PDFs are quite similar except an outlier in the V.A. sample. Removing such outlying individual largely reduces the p-value to .1272, while the GPD and the V.A. SPO are only reduced to .1062 and .0739, respectively. On the other hand, statistical tests may not be accurate on multi-modal distributions, where the stability metrics are robust. Such problem can be observed in the *oldpeak* variable 8(a)), where ANOVA provides a p-value larger than its numerical predecessors.

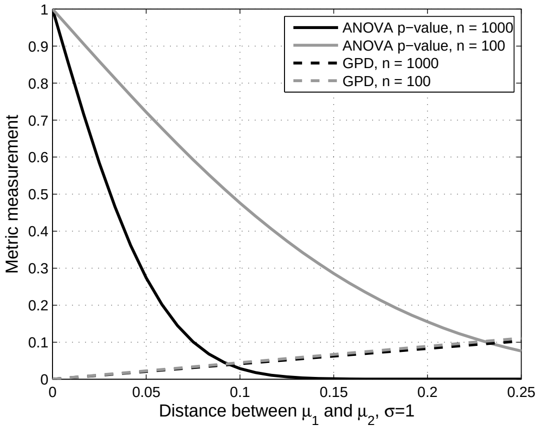


Figure 5: Comparison of the behaviour of the ANOVA p-value and the GPD (Ω) with different number of individuals. Two simulated Gaussian distributions with equal standard deviation were incrementally separated, where n random points were generated in each case. Probability density functions for GPD were estimated using KDE.

The results also show how outlying sources can be identified by the SPO metric. First, in the *age* variable, the respectively younger and older patients of Hungarian and V.A. datasets have their effect on their SPO metrics (Figure 7(b)). Regarding to the *chol* (serum cholesterol) variable, the Switzerland dataset showed an extreme outlyingness, probably caused by a wrong codification of the missing values: while in the Heart Disease dataset missing values are coded with -9, these seem to be coded with 0 (Figure 8(c)). In the *thalach* (maximum heart rate achieved) variable the projection shows the dissimilarity found among all sources (Figure 7(c)). Finally, the *num* variable corresponds to the heart disease diagnosis, and is the dependent variable for the data mining purposes of the dataset (note that studies with the Heart Disease dataset generally group positive values into a single positive class). However, it can be observed that there are large differences among the datasets. Specifically, the Hungarian dataset do not have patients with a value larger than 1, and Switzerland has very few healthy patients (0 value) in comparison with the others (Figure 8(b)).

5.2.2 Bivariate evaluation

Results of bivariate evaluation are shown in Table 4. As described in 5.1.2, a low number of individuals makes histograms or density estimations to be more noisy due to data sparsity, thus, the low number of individuals on the evaluated dataset makes the GPD metric to tend being slightly higher in this bivariate test. However, these measurements are comparable among them, which permits discovering interactions of pair of variables (concretely of their joint probability) with respect to the data source. It can be observed that the large univariate variability of *chol* is reflected in all of its joint GPDs. On the other hand, the combinations including the dependent variable,*num* in this case, should take special attention by researchers as variability may indicate possible conflicts when developing predictive models based on the multiple datasets.

	sex	**cp**	**trestbps**	**chol**	**fbs**	**restecg**	**thalach**	**exang**	**oldpeak**	**num**
age	.4123	.4515	.3516	.7562	.3416	.4992	.4917	.3999	.4006	.5469
sex	-	.3622	.2871	.7084	.2939	.4392	.4163	.2995	.5197	.6456
cp	-	-	.3687	.7160	.3550	.4939	.4703	.3344	.5568	.6714
trestbps	-	-	-	.6893	.2125	.4194	.3988	.2683	.2927	.4945
chol	-	-	-	-	.7005	.8357	.7367	.7065	.7080	.7797
fbs	-	-	-	-	-	.4138	.4198	.2947	.5042	.5928
restecg	-	-	-	-	-	-	.5287	.4420	.5950	.7063
thalach	-	-	-	-	-	-	-	.4022	.4580	.5789
exang	-	-	-	-	-	-	-	-	.4919	.6277
oldpeak	-	-	-	-	-	-	-	-	-	.5512

Table 4: Results of bivariate evaluation on the UCI Heart Disease dataset. Each cell shows the GPD (Ω) of the joint probability of the variables in the corresponding row and column.

5.2.3 Multivariate evaluation

The stability metrics were measured using all the available variables to assess the general stability of the complete dataset. To illustrate this example the PCA dimensionality reduction method with dummy coding of categorical variables was used. PCA was applied to the full dataset containing data from the four sources. The first three components were used for the analysis. Figure 9(a) shows dataset projection on these three first components, where the source of each individual is identified. It can be observed that there is a clear dissimilarity on the distributions of each source. The stability metrics were calculated on these distributions. Figure 9(b) shows a 2-dimensional simplicial projection of the 3-simplex obtained with the method, which yielded the stability metrics shown in Table 5. The observed dissimilarity among the sources is reflected on the metrics. The 2-dimensional sphere in Figure 9(b) represents the upper variability bound defined by the 1R-simplex where all the pairwise dissimilarities are maximum —in such situation all points would be located in the sphere. Thus, the obtained simplex and metrics reflect a large variability among all sources, without a clear cluster of data sources defining an approximate centroid of the problem. The most outlying source corresponds to the Switzerland sub-dataset. That may be due to the data quality problems present in the dataset, such as the apparently wrong codification of missing values, the low number of individuals after the cleansing procedure, as well as the difference in the target variable.

	GPD (Ω)	SPO ($\mathbb{O}$)			
		Cleveland	**Hungarian**	**Switzerland**	**VA**
4 sources	.5840	.4753	.4647	.5195	.4477

Table 5: Results of multivariate evaluation on the UCI Heart Disease dataset

6 Discussion

6.1 Significance

The common methods to assess the variability of multi-source biomedical data are generally suited to univariate measurements, and most take parametric or homoscedasticity assumptions on them. The evaluation results of the stability metrics developed in this work show that these metrics are a robust alternative to classical methods on multi-type, multi-modal and multivariate data, or a complementary tool when classical assumptions are met.

The GPD metric theoretically aims to increase as the global pairwise dissimilarity among the PDFs of data sources increases. That was validated by the evaluation results. Thus, the purpose to measure the degree of variability of multi-source data is accomplished. This is analogous to classical methods, but with the advantage of being suited to multi-type, multi-modal and multivariate data. Additionally, it has been shown that the GPD keeps stable as the sample size decreases in comparison with the p-values of classical statistical methods such as ANOVA Figure 5.

The SPO metric provides additional information about the outlyigness of each data source with respect to a latent central tendency of all the sources' distributions. To our knowledge such information is not provided by any classical test. On numerical data, ANOVA provides the sum-of-squares measurement as a measurement of the variability between groups. That is conceptually equivalent to the intermediate PDF dissimilarity matrix obtained during spatial stability calculus. The PDF dissimilarity matrix, however, is bounded and suited to the aforementioned features of data distributions.

Regarding to data quality, Weiskopf et al. [5] identified some methods to measure the *concordance* of datasets based on comparisons with gold standard equivalent repositories. The stability metrics permit measuring such degree of dataset concordance without requiring an additional gold standard dataset. Hence, the GPD metric provides the degree of concordance among datasets, while the SPO metric provides the degree of concordance of specific datasets with respect to a latent reference to all the datasets. Hence, the GPD and SPO can be defined as a composite measurement method of a spatial stability data quality dimension. The spatial stability can therefore be assessed under data quality assurance protocols.

One of the most practical use cases where the proposed methods can be used is the initial data understanding and data preparation stages of multi-source biobanks based research. It includes data mining or clinical trials. The GPD metric can be used to find global dissimilarities among data sources' PDFs. Large values could be caused by a low overall probabilistic concordance, or by outlying specific sources, due to possible centre or user biases. Such source outlyingness would be measured by the SPO metric. Researchers could decide to remove anomalous sources from their study or take the appropriate decisions to correct possible biases. As an example, in the development of predictive models outlying sources may reduce the global effectiveness and generalisation of models. Researchers may even consider detected variabilities as an outcome of their studies. In addition, the spatial stability plot may help to visually identify patterns among a large number of sources, with the possibility to use the intermediate PDF dissimilarity matrix as the input of subgroup discovery algorithms such as hierarchical clustering.

6.2 Limitations

Using the spatial stability metrics may require some attention under some situations, as well as in most actual data mining methods. Results showed that metrics are scalable to the number of variables. This is true according to the theoretical definition of metrics. However, in practice, the curse of dimensionality may affect to the metrics. Hence, as the number of variables increases, the probabilistic space becomes sparser. Specifically, the sparsity of a low number of data points —i.e., individuals— across the probabilistic space may cause the PDF estimations to be inaccurate —e.g., sparse, unsmoothed or 'peaky' PDFs—, leading to anomalous PDF distances. Such a variance of PDF distance estimators related to dimensionality has been discussed in other studies [32].

Nevertheless, as in most data mining tasks, the curse of dimensionality can be relaxed using proper dimensionality reduction methods or selecting a subset of appropriate study variables. In this work, PCA was used in the multivariate evaluation experiment. However, other non-linear methods or methods with a more intelligent treatment of categorical variables may be more suitable with multi-modal or categorical data. E.g., if distances among categories can be specified, the ISOMAP algorithm could be used to generate a dimensionality reduced manifold conserving distances between data points.

On the other hand, even when no dimensionality reduction is required, the PDF estimation method may also imply some variance on the PDF distances and, thus, to the stability metrics. The estimation of categorical histograms is straightforward. However, numerical data can be estimated using both histograms or other smoothing methods such as KDE, which may require tuning specific parameters such as the bin size (in the case of histograms) or kernel bandwith (in the case of KDE). As a consequence, an inadequate parametrization may lead to inaccurate PDFs. With the purpose to accurately estimate PDFs, parameters can be selected manually, where the optimum values are selected by a user, or automatically, using different methods to select them [31, 33]. In this work, the KDE bandwidth was selected using the latter approach, simulating a totally automatic spatial stability assessment. The automatic method provided reliable estimations. However, the use of other method or some manual adjustments on the kernel bandwidths may have provided slightly different results. Nevertheless, in the proposed method to obtain the stability metrics, the PDF estimation step is flexible to the use of different estimation methods suited to specific purposes or based on semantic knowledge about the problem.

Other aspect avoided in this work but which may be present on real multi-source biomedical data is the patient overlap. Weber [34] showed that the patient overlap among different sources may limit the effectiveness of tools oriented to multi-site datasets. Thus, if it is to happen, it should be considered before applying any method. However, if the number of individuals is sufficiently high in comparison with those overlapping patients, that problem may be of little significance.

6.3 Future work

Some of the classical methods, such as ANOVA or χ^2 tests, have associated p-values indicating the statistical significance on the difference between the univariate measurements. They allow taking decisions based on the rejection of a null hypothesis. The stability metrics do not currently provide such a p-value, hence, its interpretation aimed to decision making may require further understanding. The GPD can be considered a estimator equivalent to the notion of normalized standard deviation of PDFs. As a descriptive estimator, further work can be carried out to characterize its measurements on different contexts and problems. First, the GPD behaviour can be characterized according to different changes on different types of distributions, similarly to the previous work [2] discussed in Section 2.1. Second, the GPD outcomes can be associated to evaluation indicators of different target problems combining multi-source data. As an example, it may help understanding which GPD thresholds are sufficient to maintain acceptable error bounds in predictive modelling combining multi-centre data. On the other hand, it is also left for future work studying the possibility to provide confidence intervals on the stability metrics.

Nowadays many biomedical studies still count with low sample sizes, what may lead to the aforementioned limitations, specially in high dimensions. Hence, further work should be carried out with the purpose to characterize this effect to obtain possible calibrations or error bounds for the metrics. Additionally, such work may be combined with the study of the proper dimensionality reduction methods suited to the analysed data.

It may also be noted that as the Jensen-Shannon distance was used in this work as PDF distance for its symmetry, smoothness and bounds, that distance is at a small constant to the Hellinger distance [2, 35]. Hence, each of them may be used interchangeably for the proposed metrics. Further studies may identify specific features for their selection.

Other interesting capabilities of the method emerge as future work aimed to the data preparation procedures. The method can be used to assess the stability of other data quality features such as missing data. The GPD and SPO metrics represent additional features of the dataset which may improve the development of models or hypotheses on multi-source data. In an environment with a large number of sources, such a large set of hospitals in a country, or a large number of users in a hospital, the simplicial projection can be used to obtain a clustering of these sources, as well as to provide 2D or 3D visualizations of the source dissimilarities. Hence, further visual analytics methods for data source spatial stability will be studied to provide more informative visualizations (e.g., considering sample sizes or other source features) and interactive control panels. Finally, measuring the stability metrics through a set of temporal batches can provide a temporal monitoring of the inter-source variability as well as help to detect and monitor source biases.

Further discussions can be made deriving the application of the developed stability metrics to other purposes. Data source stability, as studied in this work, can be classified as a representation learning problem. Representation learning [36] aims to find latent prior knowledge, namely ‘priors’, about data to facilitate the data understanding and model development on data mining problems. Hence, the GPD or SPO metrics may be used to represent such a prior knowledge of data. For instance, in a multi-source dataset each source outlyingness can be included as an additional variable to compensate possible dissimilarities on sources when developing data models. Similarly, the metalearning field of study [37] aims to find metaknowledge about models or data to guide the search of the most appropriate model for a specific problem. Thus, the stability metrics could be used to characterize particular datasets, where their effectiveness as a metaknowledge feature to choose apropriate models could be studied.

7 Conclusions

When multi-source data samples are expected to represent the same, or a similar population, variabilities among the sources' PDFs may hinder any data exploitation or research processes with such data. This work constructs stability metrics for assessing such variabilities. As an objective, the metrics should be robust to multi-type, multi-modal and multi-dimensional data as well as bounded and comparable among domains. The here developed method based on simplicial projections from PDF distances have demonstrated capabilities to accomplish these hypothesis, providing metrics for measuring the global probabilistic deviation of data, the source probabilistic outlyingness of each data source, and a interpretable stability plot visualization of the inter-source variability. The metrics can be used as a complementary or alternative method to classical univariate statistical tests, with the advantages of being independent to the type of variable, dealing with multi-modal distributions, and providing additional visualizations. Additionally, the GPD metric, Ω, stands as an estimator equivalent to the notion of the normalized standard deviation of a set of PDFs, a concept that may be used in several different purposes.

In practice, the spatial stability metrics can be used as part of data quality assurance protocols or audit processes. The GPD and SPO metrics conform a spatial stability data quality dimension to assess the multi-source probabilistic concordance of data, and without the need of a gold standard reference dataset. Hence, the stability metrics may help assuring the quality of —increasingly larger— biobanks-based research studies involved with multi-center, multi-machine or multi-user data.

Acknowledgements

This work was supported by own IBIME funds under the UPV project *Servicio de evaluación y rating de la calidad de repositorios de datos biomédicos* [UPV-2014-872], and the EU FP7 Project *Help4Mood – A Computational Distributed System to Support the Treatment of Patients with Major Depression* [ICT-248765]. The authors thank Prof. dr. Patrick J.F. Groenen, from the Erasmus University, Rotterdam, for his feedback on MDS.

Conflict of Interest Statement

The Authors declare that there is no conflict of interest.

References

[1] McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. PLoS ONE. 2013 03;8(3):e55811.

[2] Sáez C, Robles M, García-Gómez JM. Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE; 2013. p. 3226–3229.

[3] Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst. 1996;12(4):5–33.

[4] Cruz-Correia RJ, Pereira Rodrigues P, Freitas A, Canario Almeida F, Chen R, Costa-Pereira A. Data Quality and Integration Issues in Electronic Health Records. In: Information Discovery On Electronic Health Records. V. Hristidis (ed.); 2010. p. 55–96.

[5] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan;20(1):144–151.

[6] Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM. Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform. 2012;180:721–725.

[7] A Asuncion, D J Newman. UCI Machine Learning Repository;. University of California, Irvine, School of Information and Computer Sciences http://archive.ics.uci.edu/ml/ (Accessed 27th March 2014).

[8] Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology. 1989;64(5):304 – 310.

[9] Markus HS, Ackerstaff R, Babikian V, Bladin C, Droste D, Grosset D, et al. Intercenter Agreement in Reading Doppler Embolic Signals: A Multicenter International Study. Stroke. 1997;28(7):1307–1310.

[10] Verwey NA, van der Flier WM, Blennow K, Clark C, Sokolow S, De Deyn PP, et al. A worldwide multicentre comparison of assays for cerebrospinal fluid biomarkers in Alzheimer's disease. Annals of Clinical Biochemistry. 2009;46(3):235–240.

[11] Mattsson N, Zetterberg H, et al. Lessons from multicenter studies on CSF biomarkers for Alzheimer's disease. International journal of Alzheimer's disease. 2010;.

[12] Dargaud Y, Wolberg AS, Luddington R, Regnault V, Spronk H, Baglin T, et al. Evaluation of a standardized protocol for thrombin generation measurement using the calibrated automated thrombogram: An international multicentre study. Thrombosis Research. 2012;130(6):929 – 934.

[13] Pagani E, Hirsch JG, Pouwels PJW, Horsfield MA, Perego E, Gass A, et al. Intercenter differences in diffusion tensor MRI acquisition. Journal of Magnetic Resonance Imaging. 2010;31(6):1458–1468.

[14] Jarman B, Gault S, Alves B, Hider A, Dolan S, Cook A, et al. Explaining differences in English hospital death rates using routinely collected data. BMJ. 1999 6;318(7197):1515–1520.

[15] Rubner Y, Tomasi C, Guibas L. The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision. 2000 Nov;40(2):99–121.

[16] Endres DM, Schindelin JE. A new metric for probability distributions. IEEE Transactions on Information Theory. 2003;49(7):1858–1860.

[17] Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory. 1991;37:145–151.

[18] Parks HR, Wills DC. An Elementary Calculation of the Dihedral Angle of the Regular n-Simplex. The American Mathematical Monthly. 2002 Oct;109(8):756–758.

[19] Parzen E. On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics. 1962 09;33(3):1065–1076.

[20] Bowman AW, Azzalini A. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series). Oxford University Press, USA; 1997.

[21] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000 Dec;290(5500):2319–2323.

[22] Hershey JR, Olsen PA. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. vol. 4; 2007. p. 317–320.

[23] Amari SI, Nagaoka H. Methods of Information Geometry (Translations of Mathematical Monographs). American Mathematical Society; 2007.

[24] Torgerson W. Multidimensional scaling: I. Theory and method. Psychometrika. 1952;17(4):401–419.

[25] Borg I, Groenen PJF. Modern Multidimensional Scaling: Theory and Applications. Springer; 2010.

[26] Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964;29(1):1–27.

[27] De Leeuw J. Fitting distances by least squares. Tech Rep No 130, Interdivisional Program in Statistics, UCLA. 1993;.

[28] Whitney H. Differentiable manifolds. Annals of Math. 1940;41:645–680.

[29] Wong PC, Foote H, Leung R, Adams D, Thomas J. Data signatures and visualization of scientific data sets. Computer Graphics and Applications, IEEE. 2000;20(2):12–15.

[30] Ihler, A and Mandel, M . Kernel Density Estimation Toolbox for MATLAB ;. http://www.ics.uci.edu/~ihler/code/kde.html (Accessed 6th May 2014).

[31] Silverman BW. Density Estimation for Statistics and Data Analysis. Springer US; 1986.

[32] Carvalho ARF, Tavares JMRS, Principe JC. A Novel Nonparametric Distance Estimator for Densities with Error Bounds. Entropy. 2013;15(5):1609–1623.

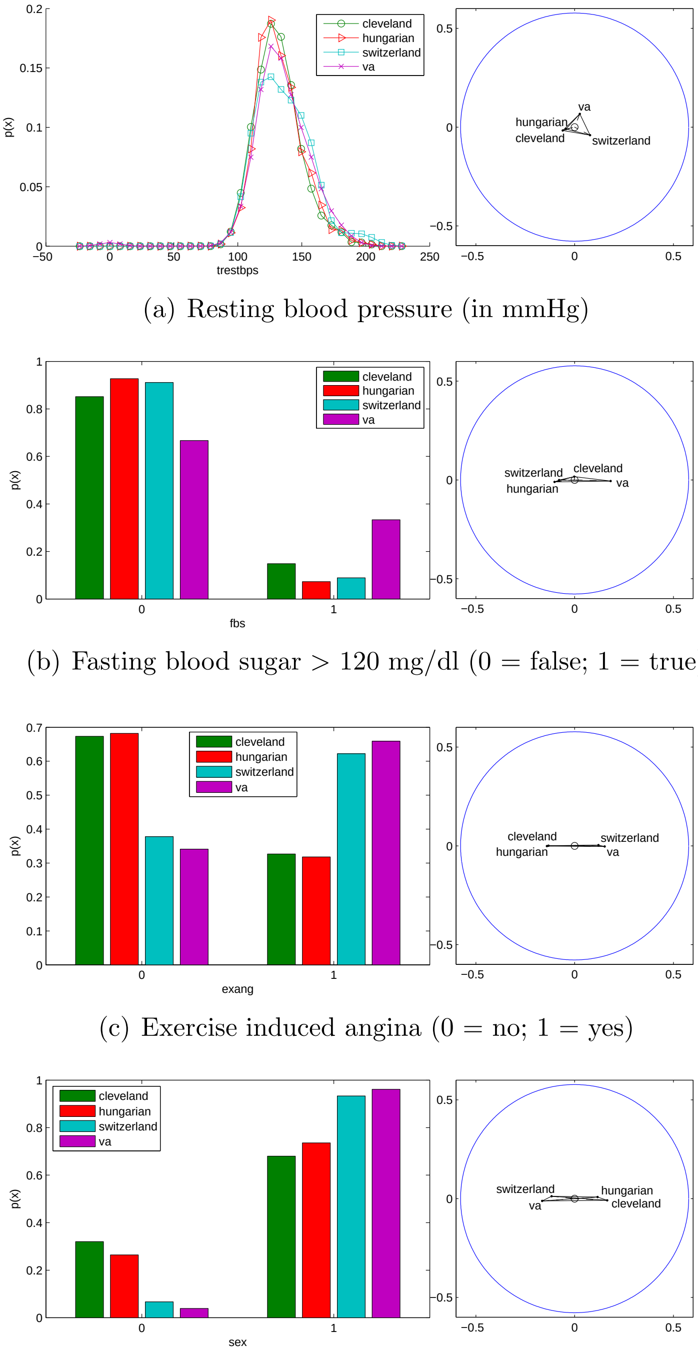
[33] Shimazaki H, Shinomoto S. A Method for Selecting the Bin Size of a Time Histogram. Neural Comput. 2007 Jun;19(6):1503–1527.

[34] Weber GM. Federated queries of clinical data repositories: the sum of the parts does not equal the whole. J Am Med Inform Assoc. 2013 Jun;20(e1):e155–161.

[35] Jayram TS. Hellinger Strikes Back: A Note on the Multi-party Information Complexity of AND. In: Proceedings of the 12th International Workshop and 13th International Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. APPROX '09 / RANDOM '09. Berlin, Heidelberg: Springer-Verlag; 2009. p. 562–573.

[36] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2013;35(8):1798–1828.

[37] Brazdil PB, editor. Metalearning: applications to data mining. Cognitive technologies. Berlin: Springer; 2009.



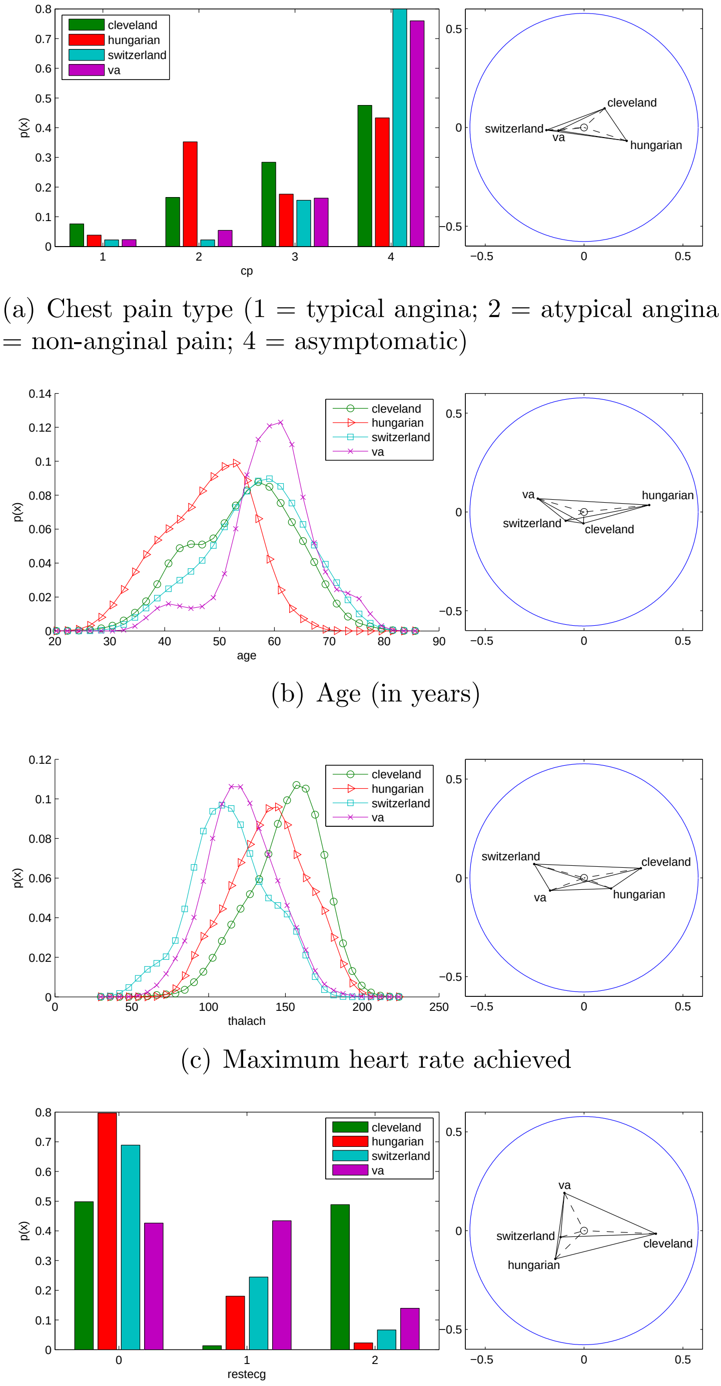
(a) Resting blood pressure (in mmHg)

(b) Fasting blood sugar > 120 mg/dl (0 = false; 1 = true)

(c) Exercise induced angina (0 = no; 1 = yes)

(d) Sex (0 = female; 1 = male)

Figure 6: Univariate probability distributions and 2-simplex stability plots for variables *trestbps*, *fbs*, *exang* and *sex*. The 2-dimensional sphere represents the upper variability bound where all the pairwise dissimilarities would be maximum.



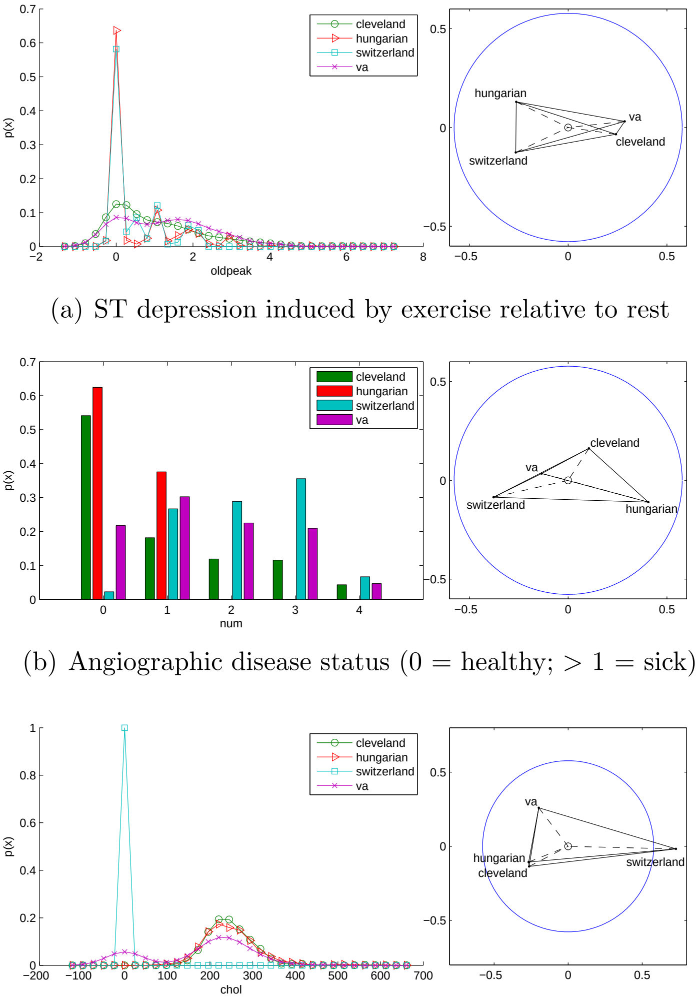
(a) Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)

(b) Age (in years)

(c) Maximum heart rate achieved

(d) Resting electrocardiographic results (0 = normal; 1 = ST-T wave abnormality; 2 = left ventricular hypertrophy)

Figure 7: Univariate probability distributions and 2-simplex stability plots for variables *cp*, *age*, *thalach* and *restecg*. The 2-dimensional sphere represents the upper variability bound where all the pairwise dissimilarities would be maximum.

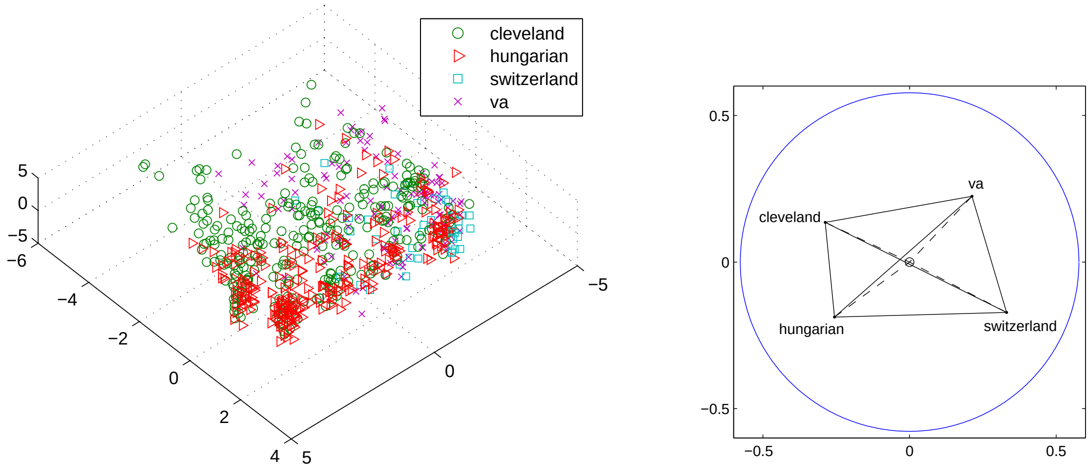


(a) ST depression induced by exercise relative to rest

(b) Angiographic disease status (0 = healthy; > 1 = sick)

(c) Serum cholesterol (in mg/dl)

Figure 8: Univariate probability distributions and 2-simplex stability plots for variables *oldpeak*, *num* and *chol*. The 2-dimensional sphere represents the upper variability bound where all the pairwise dissimilarities would be maximum.



(a) The UCI Heart Disease dataset on its three first PCA components. Data sources are identified.

(b) 2-simplex stability plot of stability

Figure 9: Visualizations of multivariate stability on the UCI Heart Disease dataset.