Sáez Silvestre, C.; Pereira Rodrigues, P.; Gama, J.; Robles Viejo, M.; García Gómez, JM. (2014). Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Mining and Knowledge Discovery. 28:1-1. doi:10.1007/s10618-014-0378-6.

Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality

Carlos Sáez[1,2], Pedro Pereira Rodrigues[2,3], João Gama[3], Montserrat Robles[1], and Juan Miguel García-Gómez[1]

[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] Laboratório de Inteligência Artificial e Apoio à Decisão (LIAAD)-INESC, Universidade do Porto, Porto, Portugal

Abstract

Knowledge discovery on biomedical data can be based on on-line, data-stream analyses, or using retrospective, timestamped, off-line datasets. In both cases, changes in the processes that generate data or in their quality features through time may hinder either the knowledge discovery process or the generalization of past knowledge. These problems can be seen as a lack of data *temporal stability*. This work establishes the temporal stability as a data quality dimension and proposes new methods for its assessment based on a probabilistic framework. Concretely, methods are proposed for 1) monitoring changes, and 2) characterizing changes, trends and detecting temporal subgroups. First, a probabilistic change detection algorithm is proposed based on the Statistical Process Control of the posterior Beta distribution of the Jensen-Shannon distance, with a memoryless forgetting mechanism. This algorithm (PDF-SPC) classifies the degree of current change in three states: In-Control, Warning, and Out-of-Control. Second, a novel method is proposed to visualize and characterize the temporal changes of data based on the projection of a non-parametric information-geometric statistical manifold of time windows. This projection facilitates the exploration of temporal trends using the proposed IGT-plot and, by means of unsupervised learning methods, discovering conceptually-related temporal subgroups. Methods are evaluated using real and simulated data based on the National Hospital Discharge Survey (NHDS) dataset.

Keywords: Data quality, Change detection, Information theory, Information geometry, Visual analytics, Biomedical data

1 Introduction

Knowledge discovery on biomedical data is generally performed over healthcare repositories or research biobanks. Either when the research repositories are generated from routine clinical data or when they are specifically designed for a research purpose, it is well accepted that the efficiency of the research processes and the reliability on their information and results are improved if the repository has been assessed for data quality (Cruz-Correia et al, 2010; Weiskopf and Weng, 2013).

The data quality research has gained attention since the work by Wang and Strong (1996). Following their approach, many studies have been developed to define what characteristics of data are related to its quality, generally known as data quality dimensions. Additionally, the increasing establishment of electronic health records (EHR) and the increase of available data is wide-spreading the necessity of biomedical data quality assessment procedures for maintaining high-quality, curated biomedical information repositories (Weiskopf and Weng, 2013).

Time is a factor that has been studied in relation to the biomedical data quality in some works. Recently, Weiskopf and Weng (2013) performed a systematic review on methods and dimensions of biomedical data quality assessment. From a resultant pool of 95 articles, only four were related to the *currency* of data. According to these studies, currency refers to the degree of how up-to-date the measurements of a patient are, and it is measured based on temporal thresholds. However, Cruz-Correia et al (2010) and Sáez et al (2012) introduced that another aspect of data quality is related to the fact that when data is collected for long periods of time, the processes that generate such data do not need to be stationary. This may be due to several reasons, such as changes in clinical protocols, environmental or seasonal effects, changes in the clinical staff, or changes in software or clinical devices. Thus, the non-stationary biological and social behaviour, as the source of biomedical data, may lead to different types of changes in data probability distribution functions (PDFs), namely gradual, abrupt or recurrent. These changes may also lead to partitions of data into subgroups of conceptually and probabilistically-related time periods, namely *temporal subgroups*. Therefore, if it is assumed that the data generating processes are stable through time, undesired and unexpected data changes may lead data to fail meeting users —i.e., data analysts— expectations, thus being considered as a lack of data quality.

This work proposes new methods for the assessment of temporal changes in biomedical data PDFs, which can be used as a framework under a *temporal stability* data quality dimension. This is related to assessing the changes causing non-stationarity of data time series (Brockwell and Davis, 2009). Hence, methods are proposed to 1) monitor changes, and 2) characterize changes, trends and detect temporal subgroups. In addition, due to the heterogeneous characteristics of biomedical data (Sáez et al, 2013), methods must be robust to different variable types, as well as to multivariate, multi-modal data. Furthermore, in order to improve the scalability of the methods, their outcomes should be provided as comparable among different domains, hence requiring bounded metrics. As a consequence, a probabilistic framework is established to support the proposed methods, comprising 1) a non-parametric synopsis of PDFs, using an incremental, memoryless non-forgetting approach (Rodrigues et al, 2010), and 2) a PDF distance measurement based

on information-theoretic probabilistic distances (Csiszár, 1967; Lin, 1991), concretely in the Jensen-Shannon distance (Endres and Schindelin, 2003).

The first proposed method is a probabilistic change detection algorithm to monitor changes in non-parametric PDFs through time. It is based on the concepts of the Statistical Process Control (SPC) by Gama et al (2004), originally designed for drift detection in the performance of machine learning models. The new algorithm (PDF-SPC) is based on the monitoring of the incrementally estimated posterior Beta distribution of the Jensen-Shannon PDF distance, classifying the degree of current change in three states: In-Control, Warning, and Out-of-Control.

The second proposed method is a novel approach to visualize and characterize the temporal changes of data based on the projection of a latent, non-parametric information-geometric statistical manifold (Amari and Nagaoka, 2007) of time windows. Concretely, a dissimilarity matrix is obtained from the PDF distances among the different time windows, where multi-dimensional scaling is used afterwards to project the temporal statistical manifold (or a dimensionally reduced version). Hence, being the PDFs of time windows projected in a geometric space, this permits visualizing and characterizing the temporal changes that occur in data, as well as to apply unsupervised learning methods, such as clustering, to obtain conceptually-related subgroups of temporal windows.

The interpretation of the results provided by the proposed methods is facilitated by visual methods, namely PDF-SPC control charts and information-geometric temporal plots (IGT-plots) of the statistical manifolds. Also, dendrograms and dissimilarity heatmaps can be used as complementary visualizations. Additionally, as a by-product of the probabilistic framework, the continuous estimation of PDFs leads to *probability mass temporal maps* which, similarly to spectrograms, help understanding the temporal changes of probability distributions.

The rest of the paper is organized as follows. Section 2 describes the required background. Section 3 describes the probabilistic framework and the two proposed methods. Section 4 describes the National Hospital Discharge Survey (NHDS) dataset used in the evaluation. Section 5 describes the evaluation and its results. Section 6 discusses the study, and compares it with state-of-the-art related work. Finally, Section 7 concludes the paper.

2 Background

Biomedical data are generally gathered in two ways for its analysis: *on-line* and *off-line*. The classical method for accessing research data is as an off-line dataset, e.g., a comma-separated values file or a small relational data base. However, the continuous increase on the amounts of available clinical data is changing the tendency to on-line methods, where data is analysed through the continuous observation of batches, generally aiming to optimise processing and storage resources (Gama and Gaber, 2007; Rodrigues and Correia, 2013). On the other hand, when the purpose is to monitor biomedical indicators in real time, the on-line analysis is straightforward. This work aims to apply to both scenarios, thus, providing users feedback about changes on their off-line dataset or during on-line processes.

This section describes some previous theoretical background which is required for the new methods proposed in this work. Concretely, this background is divided in two main topics: the probabilistic framework to compare biomedical data distributions and the change detection methods.

2.1 Probabilistic distances on biomedical data distributions

Biomedical data show heterogeneous conditions. They are generally based on multi-modal distributions —i.e., various inherent generative functions, such as a mixture of affected and unaffected patients. Studies may be uni- or multivariate, and may be composed of different types of variables —i.e., continuous, discrete ordinal and non ordinal, or mixed. Under these conditions, comparing different data samples or batches through time based on classical statistics may not be enough representative, or even not valid. Additionally, in order to measure the magnitude of changes it is interesting to provide a metric for such comparisons which, ideally, should be bounded to facilitate its comparability on different domains.

Sáez et al (2013) studied the behaviour of different PDF dissimilarity metrics with respect to these conditions. The results of such study are summarized in Table 1. 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 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)(Rubner et al, 2000) 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 the optimal minimum cost of transforming one histogram into another. Then, information-theoretic distances permit constructing over the theory of a probabilistic framework.

Condition	**T**	**KW**	**KS**	**JF**	**JSD**	**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: Summary table of the study by Sáez et al (2013). It shows the capacity of PDF distances (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), JSD: Jensen-Shannon$^{1/2}$, EMD: Earth Mover's Distance. The meaning of '-' is that the corresponding distance is not designed for the corresponding feature.

Focusing on the information-theoretic distances, the Jeffrey distance is a symmetrized, metric version of the Kullback-Leibler divergence. However, it is not bounded and, as showed by Sáez et al (2013), when the probability mass in any region of the support in any of the compared PDFs tends to zero, the metric tends to infinite. In contrast, the

Jensen-Shannon distance (JSD), square root of the Jensen-Shannon divergence (Endres and Schindelin, 2003), is a metric bounded between zero and one, and it was smoothly convergent to one on that situation. As a consequence, in this work the JSD was selected as the distance between PDFs.

2.2 Change detection

Change detection methods have been widely studied in data streams, specially when data are generated as a continuous flow and limited processing or storage resources are available (Gama, 2010). Change detection aim at identifying changes on sufficient statistics of the sample measured through time (Basseville and Nikiforov, 1993; Gama and Gaber, 2007). The selection of the change detection method and the corresponding sufficient statistic depend on the purpose, and generally follow two approaches: 1) monitoring data distributions, such as the evolution of the average; or 2) monitoring the evolution of performance indicators, such as the fitness of data mining models or patterns (Klinkenberg and Renz, 1998).

Changes can be classified according to their causes and to their behaviour, e.g., their rate of change. Regarding to the causes, changes can occur due to modifications in the context of data acquisition, e.g., changes in clinical protocols. On the other hand, related to their behaviour, changes may be characterized as 1) gradual, 2) abrupt and 3) recurrent. In the literature, gradual changes are also known as *concept drifts*, while abrupt as *concept shift*. Abrupt changes do not necessarily imply changes with a large magnitude. In fact, the early warning of small changes may be of crucial importance to prevent larger problems caused by the accumulation of such small changes (Basseville and Nikiforov, 1993). The proper change detection methods will depend on these requirements.

In this study, the focus is to detect and characterize changes in the PDF of data. This is usually based on monitoring temporal windows of the current PDF with respect to a reference window. This involves three related aspects: 1) the type of window scheme, 2) the synopsis of the windowed data into a sufficient statistic, and 3) the change detection method on the sufficient statistic.

Time windows schemes define the characteristics of the temporal period which data is synopsed —i.e., aggregated— to be monitored. The simplest approach is to use sliding windows of fixed size (Gama and Gaber, 2007; Mitchell et al, 1994). Thus, for a window size of w observations, when the individual i is observed, the $i - w$ is forgotten. This approach is useful on sensor data, which are expected to arrive in a continuous stream. However, biomedical data do not necessarily have a constant flow, e.g., the number of patient discharges presents large variations during the day, among the days of the week, or have a seasonal effect. This, in addition to the social organization of time, may lead to an inaccurate statistical sampling. Hence, a solution comes by using sliding windows within temporal semantic landmarks (Gehrke et al, 2001), i.e., aggregating daily, weekly, or monthly data, independently of the number of individuals within each semantic block. These approaches use a catastrophic-forget, i.e., the outside-window information is ignored. However, as concepts may evolve smoothly, old data may still be important (Gama et al, 2004). In tilted windows, current information is an aggregation at increasing levels of granularity from past to current data (Han et al, 2012). Thus, old data are still

used but latter examples are given more importance. Other approach to synopse data without forgetting is using weighted sliding windows. Hence, each observation is weighted according to its age, getting older data less weight. Due to the fact that the amount of memory is limited, specially in ubiquitous streams scenarios, weighted sliding windows weight data individuals within a window. Thus, there is still a minor outside-window forgetting. In order to overcome this issue, Rodrigues et al (2010) proposed the incremental memoryless fading windows. It uses all previous data in an incremental manner, i.e., only the last observation is maintained in memory, approximating weighted windows within specific error bounds.

Synopsis methods aim to aggregate or summarize data within a window as the basis for the sufficient statistic to be monitored for changes. Simplest methods may just calculate the window central tendency —e.g., a weighted average according to weighted windows schemes— and dispersion. In scenarios where the Gaussian behaviour is not the default, other methods such as histograms or wavelets (Chakrabarti et al, 2001) may result more suitable. Thus, frequency histograms or wavelet coefficients are calculated on the window data as a compact aggregation of its information. The synopsed sufficient statistic of a current window i can be calculated, as previously mentioned, according to weighted past information. Hence, memoryless fading windows provide α-fading sufficient statistics considering all previous data points. The general form of an α-fading statistic $\Upsilon_\alpha(i)$ of the variable v is

$$\Upsilon_\alpha(i) = \begin{cases} v_1, & \text{i} = 1 \\ v_i + \alpha \cdot \Upsilon_\alpha(i-1), & \text{i} > 1 \end{cases} \tag{1}$$

with $0 < \alpha < 1$. Hence, $\Upsilon_\alpha(i)$ is the α-fading statistic obtained from the synopsis of data in the landmarked window i.

Detection methods have been proposed depending on the type and purpose of the analysed data (Sebastião and Gama, 2009). The Page-Hinkley Test (Mouss et al, 2004) is one of the most referred when the monitored data is assumed to show a Gaussian behaviour —e.g., in industrial processes. Data streams do not necessarily need to follow a Gaussian distribution. To deal with this, Kifer et al (2004) proposed a non-parametric change detection method based on a relaxation of the total variation distance between PDFs. On the other hand, with foundations on the Statistical Quality Control by Shewhart and Deming (1939), Gama et al (2004) proposed a Statistical Process Control method to detect changes in the performance indicators of machine learning models — i.e., the classification error-rate. Their SPC defines three possible states for the system: In-Control, Warning and Out-of-Control. The state is selected according to the confidence interval of the current error-rate to be generated from the original distribution. Thus, an Out-of-Control state is associated to a concept drift, leading to the re-learn of a new classification model with the observations since the last Warning state —as a meaningful reference of the beginning of the new concept. .

3 Proposed methods

This section describes the proposed methods for the assessment of the temporal stability DQ dimension. The proposed methods are based on a common probabilistic framework defined by the measurement of the distance between the PDF of different temporal windows. This framework is described first in this section. Then, the new methods for change monitoring and for the characterization and subgroup discovery are described.

3.1 Probabilistic framework

The framework defines the methods to 1) estimate the PDF of the data within a window, and 2) measure the PDF distance between two windows.

In terms of change detection, the method to estimate the window PDF can be defined according to a time window scheme and synopsis method. A prior consideration is that the social organization of time is reflected in temporal biomedical data. Thus, depending on the hour, weekday, week, month or year there will always be an implicit biased behaviour. As a consequence, the use of such a temporal landmarked windows (with a granularity according to the characteristics of the study) is recommended for a proper sampling. Hence, sufficient statistics will aggregate the data within such windows.

On the other hand, in Section 2.2 it was defined that the flow at which biomedical data is generated is not generally constant —i.e., the number of individuals per time period. This may depend on the aforementioned social organization, but also on other contextual factors. Therefore, the data samples in different landmarked windows may not be enough representative, and may lead to inaccurate sufficient statistics. In order to overcome that issue, the landmarked windows are combined with a memoryless fading windows scheme. The initial landmarked window and the fading window are used in different tasks. While the former contains the data points which are synopsed to obtain the sufficient statistic, the later contains the set of sufficient statistics which are gradually weighted. In addition to the computational advantages of memoryless fading windows, as an approximation to weighted windows they contribute to the non-forgetting of past data, which is important for the tracking of gradual changes.

A requirement for the temporal stability methods is that they must be robust to the heterogeneous conditions of biomedical data. Hence, synopsis methods should capture such information for further analyses. With such a purpose, histograms stand as a proper method as they can be obtained for continuous, discrete, and even for mixed types problems, as well to multivariate data.

On discrete variables, histograms may exactly correspond to their PDF, where each bin contains the probability mass associated to a value on the distribution support. However, on continuous distributions histograms must be defined according to a set of non-overlapping intervals, leading to a discrete number of bins approximating the original continuous PDF. Different techniques exist to obtain the proper number of bins on continuous data (Guha et al, 2004; Shimazaki and Shinomoto, 2010). Additionally, when the problem is purely continuous, kernel density estimation (KDE) methods (Bowman and Azzalini, 1997; Parzen, 1962) can be used to obtain a generative and smoothed PDF.

As a consequence, each window PDF, further on P_i, will be approximated as an α-fading averaged histogram where the probability mass of each bin is

$$H_{b,\alpha}(i) = \frac{S_{b,\alpha}(i)}{N_{b,\alpha}(i)}, \tag{2}$$

where, following Equation 1, $S_{b,\alpha}(i)$ is the α-fading sum of the raw probability mass of bin b at window i, and $N_{b,\alpha}(i)$ the corresponding α-fading increment (i.e., the α-fading account of averaged bins), with $0 < \alpha < 1$.

The memoryless approximation of the α-fading averaged histogram is not error free in comparison to a weighted approximation. It is proved (Rodrigues et al, 2010) that the error can be bound within a confidence interval of $\pm 2\epsilon R$ setting $\alpha = \epsilon^{\frac{1}{w}}$, where $R = 1$ is the variable range —as a probability mass—, and w corresponds to the window size to approximate.

On the other hand, the framework establishes a method for the measurement of the distance between the PDFs of two windows. Such method should be 1) robust to multivariate, multi-type and multi-modal data, 2) bounded and 3) smoothly convergent with near-0 probability bins. As discussed in Section 2.1, a method that fulfils these properties is the Jensen-Shannon distance. Hence, the distance between the PDFs of two windows, P_i and P_j is

$$d(P_i, P_j) = JSD(P_i||P_j) = JS(P_i||P_j)^{1/2} = \left(\frac{1}{2}KL(P_i||M) + \frac{1}{2}KL(P_j||M)\right)^{1/2} \tag{3}$$

where $M = \frac{1}{2}(P_i + P_j)$, and $KL(P||Q)$ is the Kullback-Leibler divergence between distributions P and Q. Considering the histogram approximation of the PDFs, the discrete Kullback-Leibler divergence is calculated as

$$KL(P||Q) = \sum_b \log_2\left(\frac{P_b}{Q_b}\right) P_b \tag{4}$$

where P_b and Q_b are the approximated probability mass at bin b. When using the base 2 logarithm to calculate the Kullback-Leibler divergence, the Jensen-Shannon distance is bounded between zero and one.

3.2 Change monitoring

With the purpose of monitoring changes as part of the temporal stability data quality assessment, a new change detection algorithm is proposed. The degree of change between the PDFs of two time windows is given by their Jensen-Shannon distance. The JSD is $[0, 1]$-bounded and always positive. Thus, in a stable system and considering some Gaussian noise, monitoring the JSD between the PDFs of the current window and a reference past window will provide a stable signal close to zero. Then, the objective would be monitoring a sufficient statistic associated to the data stability, i.e, a sufficient statistic of the distribution of the Jensen-Shannon distances. The proposed change detection and monitoring method is based on the concepts of SPC by Gama et al (2004) —originally aimed to monitoring the error rate of predictive models— to monitor the data stability based on the Beta distribution of the JSD.

Suppose a sequence of PDF estimations $\{P_i\}$. Using the first element as a reference, $P_{ref} = P_1$, the JSD of further elements $P_2, ..., P_i$ with respect to the former provides a sequence of distances $\{d_i\}$. In a stable system, d will approximately be normally distributed around a central tendency distance associated to a latent noise. More strictly, as the JSD is $[0, 1]$-bounded, d is a Beta random variable. Hence, in a stable system, after a transitory state, the mean value μ of the $Beta(\alpha, \beta)$ distribution B given by $\{d_i\}$ will remain stable. Additionally, an upper confidence interval u^z for B is given by the inverse cumulative distribution function $iCDF(.5 + z/2)$, with $0 < z < 1$ —e.g., for an upper confidence interval at 95% then $z = .95$.

The proposed PDF-SPC method manages three registers during the monitoring, $u_{min}^{z_1}$, $u_{min}^{z_2}$ and $u_{min}^{z_3}$, with $z_1 < z_2 < z_3$. For each new distance d_i, which updates the Beta distribution B, if the new $u_i^{z_1}$ is lower than $u_{min}^{z_1}$, the three registers are updated based on B_i. Hence, the values of z_1, z_2 and z_3 depend on the desired confidence levels —e.g, based on the three-sigma rule the upper confidence intervals can be set to u_{min}^{68}, u_{min}^{95} and u_{min}^{997}.

Given a new distance d_i, three possible states are defined for the system:

- In-Control: while $u_i^{z_1} < u_{min}^{z_2}$. The monitored PDF is temporary stable.
- Warning: while $u_i^{z_1} \geq u_{min}^{z_2} \wedge u_i^{z_1} < u_{min}^{z_3}$. The monitored PDF is changing but without reaching an action level. Its causes may be noise or a gradual change. Hence, an effective change should be confirmed based on further data.
- Out-of-Control: whenever $u_i^{z_1} \geq u_{min}^{z_3}$. The current PDF has reached a significantly higher distance from the past reference. The current B_i is different from the reference with a probability of z_3.

Reaching the Out-of-Control state means that a new concept is established. As a consequence, in order to continue the change monitoring process, the PDF-SPC algorithm (Algorithm 1) will replace the reference PDF with the current concept. Hence, if the Out-of-Control state is reached after P_j is observed, then $P_{ref} = P_j$.

As well as the estimation of PDFs is based on a α-fading incremental approach, with the purpose to avoid storing in memory all the observations of d_i, the distribution B is updated using an incremental approach. Hence, the estimation of the parameters of B, $\hat{\alpha}$ (Equation 5) and $\hat{\beta}$ (Equation 6) is based on the Maximum Likelihood Estimation using a recursive estimation of the sample geometric mean, $\hat{G}$ (Equation 7).

$$\hat{\alpha} = \tfrac{1}{2} + \frac{\hat{G}(d_i)}{2(1 - \hat{G}(d_i) - \hat{G}(1 - d_i))} \tag{5}$$

$$\hat{\beta} = \tfrac{1}{2} + \frac{\hat{G}(1 - d_i)}{2(1 - \hat{G}(d_i) - \hat{G}(1 - d_i))} \tag{6}$$

$$\hat{G}(x_i) = \left(\left(\hat{G}(x_{i-1}) \right)^{i-1} x_i \right)^{1/i} \tag{7}$$

input: P_{ref}, current reference PDF
Sequence of PDFs: $\{P_i\}$
begin
    Let P_i be the current PDF
    Let $d_i = JSD(P_i||P_{ref})$
    Let B be a $Beta(\alpha, \beta)$ distribution
    Re-estimate B with d_i
    if $u_i^{z_1} < u_{min}^{z_1}$ **then**
        $u_{min}^{z_1} = u_i^{z_1}$
        $u_{min}^{z_2} = u_i^{z_2}$
        $u_{min}^{z_3} = u_i^{z_3}$
    end
    if $u_{min}^{z_1} < u_{min}^{z_2}$ **then**
        `/* In-Control */`
        $Warning? \leftarrow False$
    else
        if $u_{min}^{z_1} < u_{min}^{z_3}$ **then**
            `/* Warning Zone */`
            if $NOTWarning?$ **then**
                $Warning? \leftarrow True$
            else
                nothing
            end
        else
            `/* Out-of-Control */`
            $P_{ref} = P_i$
            $Warning? \leftarrow False$
            Re-start B, $u_{min}^{z_1}$, $u_{min}^{z_2}$, $u_{min}^{z_3}$
        end
    end
end

Algorithm 1: The PDF-SPC change monitoring algorithm

The PDF-SPC permits identifying timestamps related to concept changes, i.e., whenever Warning and Out-of-Control states are reached. As a possible initial indicator of further larger changes (Basseville and Nikiforov, 1993), that information is specially useful to rapidly react to, or even to predict, changes. On the other hand, Widmer and Kubat (1996) suggested that two concepts may coexist before a change is achieved. The Warning state is fired when there is a suspect for a change, which may be confirmed once there is enough evidence by the Out-of-Control state. Hence, the temporal distance between a Warning an Out-of-Control states may be an indicator of such period of coexistence of concepts and, thus, of the rate of change. However, other descriptive information to characterize the behaviour of changes may be missed, e.g., whether concepts can be grouped into meaningful, possibly recurrent, groups. A promising novel method to deal with this problems is described next.

3.3 Characterization and temporal subgroup discovery

With the purpose to characterize the behaviour of changes and facilitate the discovery of temporal subgroups, a novel method is proposed. As the PDF-SPC monitors the degree of changes, this new method aims to describe them, facilitating their characterization, e.g., into gradual, abrupt or recurrent, and analysing the evolution of data inherent concepts.

According to the probabilistic framework, each time window can be seen as an individual characterized by its PDF estimation. The Information Geometry field states (Amari and Nagaoka, 2007) that probability distributions lie on a Riemannian manifold whose inner product is defined by the Fisher Information Metric of a specific family of probability distributions. The geodesic distances between the points associated to PDFs are approximated by their PDF divergences, such as the Jensen-Shannon. Hence, the JSDs among each pair of PDFs can be used to approximate a non-parametric —i.e., family-independent— statistical manifold where the temporal PDF estimations lie and, as a consequence, allow the discovery of related trends and subgroups. In addition, due to the JSD bounds, the maximum possible distance among any pair of PDF points is one. That means that the approximated statistical manifold is bounded by a hyper-ball of diameter one. Hence, the studied PDFs will lie on space comparable among different problems, as it will be known that: 1) equal PDFs will co-locate and 2) completely separable PDFs will be located at the hyper-ball surface —i.e., at a distance of one.

Suppose a sequence of PDF estimations $\{P_i\}$, with $1 < i < n$. The $\binom{n}{2}$ pairwise distances $d(P_i, P_j)$ define a n-by-n symmetric dissimilarity matrix $D = (d_{11}, ..., d_{nn}), d_{ij} : d(P_i, P_j)$. Hence, D can be used as the input of a compatible[1] clustering method, such as a complete linkage hierarchical clustering, which will provide a set of groups G_k, each related to a data inherent temporal concept.

The approximated statistical manifold provides information about the layout of PDFs in such a latent space, e.g., to discover conceptual subgroups. However, much more information can be taken considering that there is an implicit temporal order among such PDF points. While the distances among subgroups indicate the concept dissimilarity, the layout of the temporal order among their points provides information about how concepts

[1]Note that Jensen-Shannon distances are not euclidean, hence, compatible clustering methods or euclidean transformations should be used.

evolved through time. Hence, a temporal continuity through the points of a subgroup, e.g., along the vector defining its largest variance, is an indicator of a gradual change. On the other hand, a temporal alternation among different subgroups every certain time period may be an indicator of recurrent abrupt changes among probabilistically distinguished concepts. Similarly, a temporal fluctuation through a direction within a subgroup, e.g., along one of its variance vectors, may be an indicator of a recurrent gradual change among closer, probabilistically-contiguous concepts.

Hence, in order to permit such analysis it is needed to translate the dissimilarity matrix D into a set of points in a geometric space. Considering that the distances in D are not euclidean, the use of a multidimensional scaling (MDS) (Borg and Groenen, 2010; Torgerson, 1952) method is suitable to obtain an embedding of the PDFs into a euclidean space.

Given a dissimilarity matrix D, the objective of MDS is to obtain the set $P = (\vec{p}_{11}, ..., \vec{p}_{nc})$ of points in a $\mathbb{R}^c$ euclidean space such that $c = n - 1$. This is done by finding the best approximation of $||\vec{p}_i - \vec{p}_j|| \approx d_{ij}$, where $|| \cdot ||$ is the euclidean norm between points $\vec{p}_i$ and $\vec{p}_j$. This approximation can be solved by the minimization of the loss function:

$$\min_{P_1,...,P_S} \sum_{i<j} (||\vec{p}_i - \vec{p}_j|| - d_{ij})^2 \tag{8}$$

A dimensionally reduced MDS projection into 2 or 3 dimensions can be obtained to facilitate the processing and visualization of such information. Hence, simply based on the calculus of an inter-window PDF dissimilarity matrix and a MDS projection, an information-geometric temporal plot (IGT-plot) stands as a powerful visual analytics tool to explore, characterize and understand changes from a probabilistic perspective. Illustrative examples of such visualization are shown in the next section.

4 Data

This section describes the data used to evaluate the proposed methods and proposes a visualization method for monitoring PDFs.

The data used in the evaluation is the publicly available NHDS dataset (NHDS, 2014). Using only adult patients (age $\geqslant$ 18), the dataset contains 2,509,113 hospital discharge records of approximately 1% of the US hospitals from 2000 to 2009. The minimum date granularity is the discharge month. Hence, the following experiments are based on a monthly basis aggregate landmarked windows, with a total of 120 months (the time windows will be referred further on as their month index). The NHDS dataset contains several demographic, diagnosis and discharge status information. However, for the purpose of this evaluation the age and sex variables are sufficiently representative, as it is shown next.

With the purpose to illustrate the examples a *probability mass temporal map* visualization is proposed, which results as a novel visual method for the monitoring of biomedical variables. It is based on the idea of *dense pixel* visualizations (Keim, 2000), where the range of possible values are associated to a coloured pixel according to a user-specified colormap. That method has already been used to visualize sensor monitorings (Rodrigues

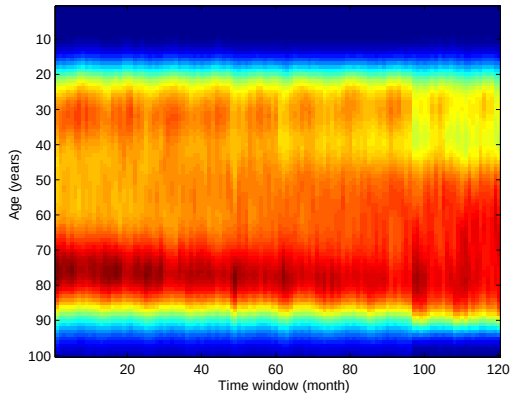
and Gama, 2010). In this case, the method is adapted to visualize the evolution of PDF estimations, where the domain axis identifies the temporal window and the range corresponds to the probability bins. Hence, each row of the map can be seen as a signal of the probability mass evolution for a given support value. In principle, the method is suitable to variables where there is an order in the variable support, i.e. as numerical data or discrete ordered. However, it may also be useful to visualize the joint probability of ordered and non ordered variables, using the latter to divide the range axis of the map on repeated supports of the former.

Figures 1(a) and 1(b) show probability mass temporal maps of the age variable (given in years). As a univariate numerical variable, PDFs at each window are estimated based on KDE to obtain an smoother histogram. In Figure 1(a), an outside-window forgetting window scheme is used. In Figure 1(b), the memoryless fading window scheme is used (an error of $\epsilon = 0.05$ was used with a smoothing window of 12 months). It can be seen that the non-forgetting approach of fading windows leads to a smoother temporal estimation, which may avoid undesirable noise caused by non-representative windows. Note that the Gaussian-kernel estimation of KDE causes that some probability mass from the lower tails of continuous Gaussian kernels is given to the bins under 18 years.

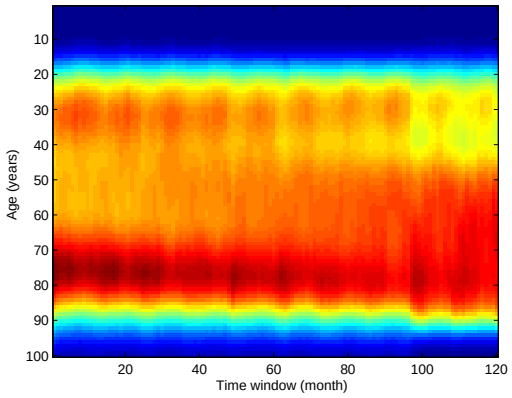
It can be observed that the age variable shows a multimodal behaviour which contains several temporal artefacts of special interest for evaluating change monitoring methods. First, there is an abrupt change in month 97. That change is documented (NHDS, 2010) as a change in the codification of the variable, where ages 91 and over were recoded to 90. Second, a gradual shift is observed as mid-age patients (age $\approx$ 55) get more probability mass through time. This has associated a decrease in the mass of younger and older patients. This change may be related to a long-term contextual change —e.g., socio-economic change or due to changes in the sampled hospitals/population — possibly associated to the increase in life expectancy in the US (Arias, 2014; National Research Council, 2011). Finally, a recurrent change is observed in young (age $\approx$ 30) patients with a periodicity of 12 months. That change is associated to the increase of births in the summer period, possibly due to the live births seasonality documented for the US (Cesario, 2002) and other countries (Wellings et al, 1999). Such effect can be observed in the marginal map for female patients (see Supplementary Material).

On the other hand, Figure 1(c) shows the temporal evolution of the probability masses of the sex variable. There is a minor gradual change in the probabilities of male and female patients, due to the increase in the males/females ratio in the US during the period of study (Howden and Meyer, 2011).

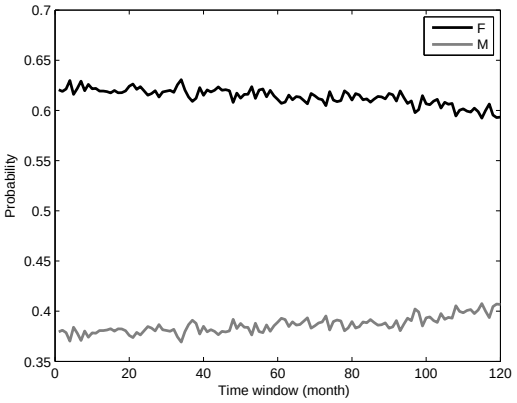
Finally, Figure 1(d) shows the evolution of the joint probability of age and sex. In this case, while including a discrete non-ordered variable —namely categorical—, it is not possible to directly apply KDE. Hence the raw synopsed histogram is used instead. In addition, to facilitate the visualization, the 2-dimensional histogram was vectorized partitioning by the sex variable, hence the upper half of the map identifies the evolution of age in females, and the lower in males. As a raw, non-KDE smoothed histogram, it can be observed that, first, ages below 18 do not get mass (as only adult patients were included), and second, the aforementioned change in month 97 makes the age of 90 get the highest mass (as it includes any older patients). That large difference in probability masses causes that lower values get close colors in the map. In this case, the visualization



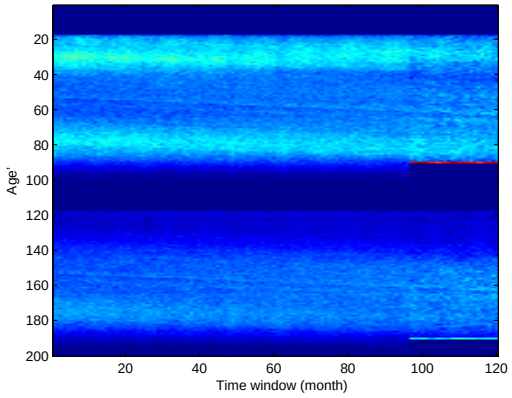
(a) Probability mass temporal map of variable age with non-weighted outside-window forget



(b) Probability mass temporal map of variable age with memoryless fading windows



(c) Temporal evolution of the probability of variable sex (F: female, M: male)



(d) Temporal evolution of the joint probability of variables age and sex (the 2-dimensional joint distribution was vectorized on Age' partitioning by sex as: 0-100 for female age, 101-200 for male age)

Figure 1: Visualizations for the monitoring of the NHDS variables

can be improved applying a logarithm function with a tuning parameter to the array of PDFs to visualize, $\log(\vec{P} + z)$, which assigns larger values of the colormap to the intermediate masses.

Hence, the combination of the age and sex variables in the evaluation study accomplishes the three heterogeneous characteristics of biomedical data to which methods must be robust: age is clearly multimodal, each is of different type, and changes can be studied on their joint —i.e., multivariate— distribution.

5 Evaluation

In this section the proposed methods are evaluated with the real changes present in the NHDS data described in previous section, as well as with simulated changes applied on

it.

5.1 Change monitoring

The PDF-SPC algorithm was applied first to a continuous univariate problem based on the age variable with the purpose to evaluate its behaviour with respect to the present changes. Second, it was evaluated on the sex variable, categorical, where a small gradual drift occurs. Then, it was evaluated on the multivariate and mixed-types problem based on the joint probability of age and sex. Finally, a simulated abrupt shift was introduced in the latter problem as a change in the joint probability of age and sex but not in their respective univariate estimates, with the purpose to evaluate the behaviour of the SPC algorithm on that multivariate change. The confidence levels were set to $z_1 = .68$, $z_2 = .95$ and $z_3 = .997$.

Figure 2 shows the results of these four evaluations. The age variable monitoring, Figure 2(a), shows that the three types of changes are detected. First, after the transitory state there is a continuous increase in the PDF distance with respect to the reference window, associated to the gradual movement of mass to the mid-age range. This leads to a warning state in month 46. Second, the abrupt change in month 97 was clearly detected. Third, the recurrent change on age ≈ 30 is captured as a periodic change in the probabilistic distance to the reference, however, the selected confidence levels avoid firing any change from them.

The sex variable monitoring, Figure 2(b), shows the gradual switch as an increase in the monitored distances. However, as expected the magnitude of the change is much lower —note that the Jensen-Shannon distance is $[0, 1]$-bounded, hence magnitudes are comparable. The recurrent change which was easily observed in the age variable is detected in this case as well. Given the 12-month periodicity, the phase displacement with respect to the age monitoring may just be due to the selected reference window.

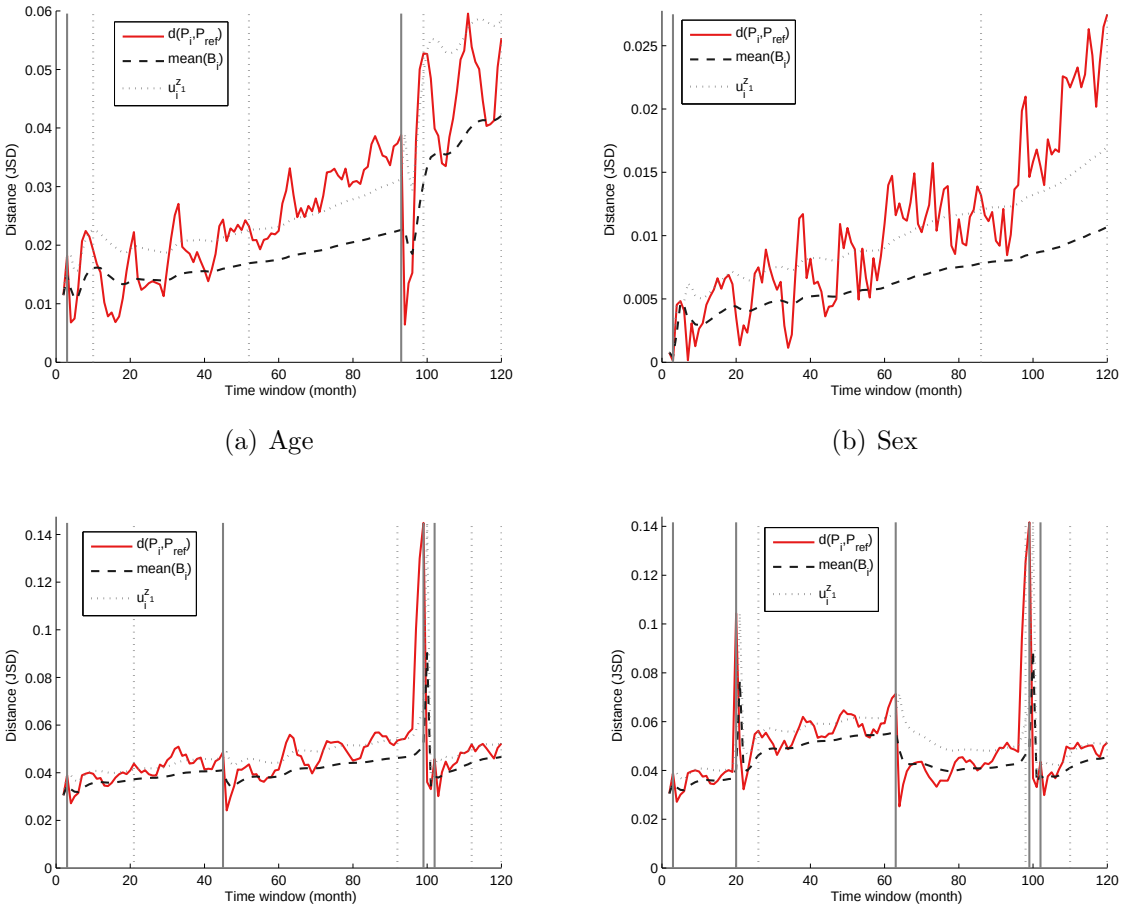
The monitoring of the joint probability of age and sex, Figure 2(c), also captures a gradual change as the mean distance also increases. However, maybe due to the sum of changes in both variables causes the change to be detected before. Hence, a change is fired after month 43. Additionally, the codification change in age is also detected, although a couple of iterations later.

In the last experiment, a multivariate change was introduced in month 20, maintaining the new concept until the end. Thus, the sex of n patients was switched, where n corresponds to the minimum amount of patients from any of the two sexes at each time window —males in all cases. Whilst the change is not detected univariately, the multivariate monitoring clearly detects the change in month 20.

5.2 Characterization and temporal subgroup discovery

The proposed methods for change characterization and temporal subgroup discovery were applied to two of the previous scenarios: in the age variable monitoring and in the monitoring of joint age and sex variables with simulated change. It is expected that characterizations and subgroups are related to the concept changes detected by the SPC method.

Figure 3(a) shows the 2-dimensional IGT-plot associated to the statistical manifold where the temporal PDFs estimated from variable age lie. Each PDF is represented as



(a) Age

(b) Sex

(c) Age+Sex

(d) Age+Sex with forced joint change

Figure 2: Control charts of the PDF-SPC evaluations. Vertical dotted lines indicate the entering into a Warning state. Vertical continuous lines indicate a change detection as an Out-of-Control state.

the index corresponding to its temporal window —i.e., the month—, allowing temporal changes to be characterized. It can be observed that there are two well differentiated groups which, looking at their indices, correspond to the concepts before and after the codification change in month 97. Looking at the first subgroup, there is a linear temporal continuity through its larger variance (Arrow A). That continuity is a clear indicator of the gradual change moving probability mass to the mid-age patients. For the visual representation, a colormap has been used to assign cooler and warmer colors to winter and summer months, respectively. Hence, it can be observed that the second variance component of the first subgroup (Arrow B) is associated to the mentioned 12-month periodic change. In addition, the same change direction is shown in the second subgroup.

The apparent subgroups were confirmed with a complete linkage hierarchical clustering based on the PDFs dissimilarity matrix. Figure 3(b) shows a heatmap of the symmetric PDFs dissimilarity matrix, where the color temperature represent a larger probabilistic distance between the PDFs P_x and P_y. The differences among the two main groups can

be observed, as well as a recurrent distance increase with a 12-month periodicity. Figure 3(c) shows the dendrogram obtained from the clustering method, which confirms such temporal groups.

On the other hand, Figure 4(a) shows the IGT-plot of the second scenario. In this case, it can be observed that there are three well differentiated groups. Looking at the indices, the first and second subgroups are separated by the forced multivariate change in month 20, while the second and third by the univariate change in age. In addition, there are transitory PDFs between the groups, which may be due to the smoothing produced by the fading windows approach. Figures 4(b) and 4(c), as in the previous example, confirm the discovered temporal subgroups. In this scenario, the 12-month recurrent change is hindered by the magnitude of the other changes, however, it can slightly be observed in the variance directions of the subgroups (note the separability among the seasonal colors) as well as in the dissimilarity matrix heatmap. However, the change detected in month 62 does not establish a subgroup change. Nevertheless, it represents the resultant change from an accumulated gradual change on both variables whose magnitude, as it can be observed in Figure 2(d), is lower. The equivalent results for the age+sex scenario without the forced change are shown in the Supplementary Material.

6 Discussion

This section highlights the significant points of this work, discusses it with related work and the limitations of the proposed methods and, finally, suggests future lines of work.

6.1 Significance

First, the PDF-SPC algorithm has shown to accurately detect the changes present in the evaluated data according to their evidences (Section 4). The resultant monitoring charts provide information about the magnitude and type of changes, showing the current probabilistic distance with respect to the reference concept. Based on the Jensen-Shannon distance, the magnitude of changes is $[0-1]$-bounded and hence comparable among different problems, e.g., the magnitude of changes in sex (Figure 2(b)) is probabilistically an average of half the magnitude in age (Figure 2(a)). Additionally, based on an incremental approach, the method is suitable to on-line analyses with a reduced storage and computational cost.

Second, the methods for change characterization and temporal subgroup discovery based on information geometry have shown to detect temporal subgroups present on the evaluated data, as well as to help characterizing the type of changes based on the temporal tendencies of the data points associated to the PDFs of time windows. To the knowledge of the authors, this is the first study of non-parametric change detection and characterization based on information-geometric statistical manifolds, with the potential to be an important step forward. To date, most change detection methods provide information about the magnitude of changes, their classification according to the rate of change, which regions in the variables of study show a major contribution to changes, or even aim to their prediction. However, the temporal projection of a non-parametric information-geometric statistical manifold constructed from consecutive time-windows permits describing and

analysing the behaviour of changes, as the evolution of a probabilistic concept through such manifold. In this study, the method has been used on one hand to construct the IGT-plots, as a novel visualization tool for the exploration of temporal changes in data PDF. In other hand, the obtained PDF points have been used for unsupervised learning purposes with the purpose to find temporal subgroups. However, these are only the first steps of many further research possibilities which still remain opened based on this approach.

Analysing the two proposed methods together, the evaluation results have demonstrated the consistency between the PDF-SPC change monitoring algorithm and the information-geometric based methods for the characterization and subgroup discovery, since the change levels and detections in the PDF-SPC monitoring are associated to the obtained temporal characterization and subgroups. In addition, both methods result suitable to the heterogeneous biomedical data conditions posed as requirements. The use of the probabilistic distances approach permits measuring changes in multi-modal distributions, as previously demonstrated by Sáez et al (2013). This, in combination with synopsing data into histograms, allows the analysis of uni and multivariate continuous, discrete ordinal and non-ordinal, as well as mixed distributions. In addition, methods have shown to be robust to detect changes on multivariate variable interactions.

As a consequence, the proposed methods have shown to be useful tools for data quality assessment focusing in the temporal stability dimension. This work has focused to the change monitoring and characterization on data distributions. The same concepts and methods can be applied to monitor other data quality features, such as monitoring the degree of missing, inconsistent, or incorrect data. These could be used to audit the quality of multi-centric or multi-user data gathering for research repositories, clinical trials, or claims data. Concretely, the latter are known to be far from perfect (Solberg et al, 2006), where these processes may be of special interest. Hence, the proposed methods can be used as exploratory data quality assessment solutions. Furthermore, as based on probabilistic metrics, they might also be used with quantitative decision making purposes. However further research is required to define these criteria.

6.2 Comparison with related work

Basic statistical methods, similarly to Shewhart control charts, have been used in the medical monitoring. E.g., laboratory systems have well established temporal quality controls based on the Levey-Jennings charts and Westgard rules (Westgard and Barry, 2010). Thus, a batch is considered Out-of-Control using basic statistics based on reference chemical reactives. On the other hand, other studies have used more complex change detection methods. Rodrigues et al (2011) proposed a method to improve the monitoring of cardiotocography signals using the memoryless fading window approach. Sebastião et al (2013) applied a Page-Hinkley change detection test combined with a time-weighted mechanism for the monitoring of depht anaesthesya signals. Similarly to the PDF-SPC, these studies focus to the monitoring of data itself, based on quality control references or in physiological signals.

On the other hand, Stiglic and Kokol (2011) proposed a method to facilitate the interpretation of changes in the performance of clinical diagnosis classification models

by means of a bivariate analysis of class labels. Using the NHDS dataset, they found a change in the performance of models to predict chronic kidney disease by the end of year 2005. Their visual method provided the insights to confirm that the change was due to change in the ICD9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) title and description of concepts D403 and D404, related to the investigated disease. Additionally, a decrease in the performance measures was found by the start of year 2008. Interestingly, that change is correlated in time with the change in the codification of age found in this work.

In the generic change detection domain this is not the first study using non-parametric PDF distance measures for detecting changes. In their useful approach, Kifer et al (2004) proposed a relaxed total variation distance among PDFs for change detection. They discarded using information-theoretic distances claiming discrete distributions were needed. However, based on the Kullback-Leibler divergence they can be used on purely continuous and, as demonstrated in this work, even in mixed-types multi-variate distributions. Dasu et al (2009) and Sebastião et al (2010) did use the Kullback-Leibler divergence in their respective studies. However, the Kullback-Leibler neither satisfies the properties of a metric nor is bounded, as required in this study.

The higher dimensionality is a challenge for change detection methods. Several solutions have been proposed in the general change detection domain. Aggarwal (2003) proposed a method based in a physical model to measure the velocity of changes in probability masses of continuous data using KDE. To deal with higher dimensionality and facilitate the understanding on changes, he proposed picking sub-projections in which the greatest amount of change has occurred. Hrovat et al (2014) applied a strategy to detect relevant subgroups on which to make the analysis, with the purpose to detect temporal trends on biomedical data. Papadimitriou et al (2005) presented a method capable to find the key trends in a numerical multivariate time series. The method internally used principal component analysis (PCA), which may not result effective with multi-modal distributions nor reducing dimensionality including categorical data, both aspects generally present in biomedical data. The previously mentioned Dasu et al (2009) approach measures changes between two PDFs embedded into a reduced structure using an extension of kd-trees. Thus, a distance metric needs to be defined between data points, what may be complicated when non-ordinal data is present. It can be deduced, hence, that dealing with non-ordinal discrete data may represent another challenge. This work deals with these problems using the synopsis on non-parametric discrete histograms.

Other approach for change detection suited to multi-modal distributions consists on monitoring cluster evolutions. Spiliopoulou et al (2006) presented the MONIC framework based on that idea, where different types of cluster changes are characterized. It is important to distinguish between such approaches and the temporal clustering presented in this work. Whilst the former clusters data within time windows, in this work what is clustered are the time windows.

6.3 Limitations

In high dimensions, the histogram synopsis method involves by default a larger probabilistic space. Hence, data points may become sparse, leading to ineffective distribu-

tion comparisons, such as larger PDF distances. Due to the heterogeneous conditions of biomedical data, the application of non-linear dimensionality reduction methods, such as ISOMAP, or solutions as exposed in related work may alleviate such problem. Evaluations on high-dimensional simulated changes may help studying these.

On the other hand, on continuous data, KDE provides smoother PDF estimations than raw histogram estimations. Although the fading window approach already smooths the obtained PDFs using past data, KDE improves the smoothing within the current window. However, sometimes using data models instead of raw data may hide other relevant information present at lower levels of probabilistic magnitude. An interesting example arises from the evaluated data. The histogram estimation of the age variable is shown in Figure 1(d) splitted by sex. It can be observed that in both sexes there is a straight temporal gap beginning approximately at the age of 47, and continuing in a yearly basis. Considering that the NHDS represents the population of the United States, and that the first sample corresponds to year 2000, it is concluded that these population gap correspond to patients born around 1943, where the social effects of World War II reduced the birth rates. On the other hand, the use of methods to select the proper number of bins in histogram may be useful to overcome some of these issues, however, the proper number of bins may also vary through time, what may require further study to make compatible the on-line probability distance measurements as the support is changed.

6.4 Future work

It has been observed in the examples that a recurrent change is related to a 12-month periodic seasonal effect. During the experiments developed in this work, it was observed that applying a simple non-weighted sliding window scheme with a window size of 12 months completely removed such effect on the resultant monthly PDF estimations (see comparison in Supplementary Material). That effect may result useful for the detection of gradual changes, since higher frequent, recurrent changes which may hinder the former are removed. However, the long-term smoothing may cause abrupt changes not to be accurately detected. Hence, two interesting future work topics arise. First, automatically detecting the period of recurrent changes, e.g., based on signal processing methods. Second, using ensemble change detection models combining different window schemes focused to specific types of changes.

The study of proper dimensionality reduction methods for effectively measuring PDF distances on high dimensions is also an important future work. This can be complemented with methods to select appropriate variable or sample subgroups on which to make the analysis. In addition, the maintenance and compatibility through time of these methods to reduce the problem complexity can be studied.

Regarding to the temporal characterization and subgroup discovery methods, it is open as further work their improvement based on incremental approaches. Hence, incremental clustering (Rodrigues et al, 2008) and MDS (Brandes and Pich, 2007) methods could be used with such a purpose. On the other hand, the use of complementary methods to optimise the efficiency of the projections, such as Self-Organizing maps (Kohonen, 1982), may be studied.

Another interesting future work is to apply functional data analysis methods (Ramsay and Silverman, 2005) to model the probabilistic temporal evolution of data on the

information-geometric statistical manifold. They may provide smoothed tendency curves on which to characterize and measure changes.

Finally, the combination of the temporal stability methods presented in this work with metrics for the probabilistic *spatial stability* among multiple sources of biomedical data (Sáez et al, 2014), will lead to a future study aiming to a probabilistic spatio-temporal data quality assessment.

7 Conclusion

The probabilistic methods presented in this work have demonstrated their feasibility for the change detection, characterization and subgroup discovery of temporal biomedical data. The changes present in the evaluated and simulated NHDS datasets have been successfully detected, in addition, with a probabilistic interpretation, as provided by the proposed PDF-SPC and information-geometric projection methods. Further studies will be made to confirm the generalisation of the methods.

As part of a data quality assessment, the proposed methods can facilitate the data understanding and lead to better decisions when developing knowledge discovery studies, either on-line or off-line, based on these data. Used as an exploratory framework, they permit visualizing the temporal stability of large healthcare databases in an interpretable and rapidly manner. In addition, methods are built to be comparable among different domains, hence, they may be used as part of a biomedical data quality auditory process. This is an important subject, as poor levels of data quality may have direct consequences on patient care (Aspden et al, 2004) as well as in the biomedical research processes (Sáez et al, 2014; Weiskopf and Weng, 2013). This work has demonstrated that data stream and change detection methods can be successfully applied in the biomedical data context, thus, further studies can still be made to analyse the impact that a temporal stability assessment can provide to real, in-production healthcare repositories.

Finally, this work has contributed to the generic change detection field of study in two aspects. First, the extension of the widely accepted SPC method to the monitoring of changes in non-parametric PDFs based on information-theoretic distances. And second, the novel change characterization method based on information geometry. It is important to emphasize the contribution to the state-of-the-art of this method. In this work, it has demonstrated possibilities which have not received proper attention in the literature yet, such as discovering temporal subgroups or characterizing the direction and length of changes through the series of time-windows in the statistical manifold. However, a lot of new possibilities are opened, standing as the first step of a promising line of research in change detection.

Acknowledgements

The work by C Sáez has been supported by an Erasmus Lifelong Learning Programme 2013 grant. This work has been supported by own IBIME funds. The authors thank Dr. Gregor Stiglic, from the Univeristy of Maribor, Slovenia, for his support on the NHDS data.

References

Aggarwal C (2003) A framework for diagnosing changes in evolving data streams. In: ACM SIGMOD Conference, pp 575–586

Amari SI, Nagaoka H (2007) Methods of Information Geometry. American Mathematical Soc.

Arias E (2014) United states life tables, 2009. National Vital Statistics Reports 62(7)

Aspden P, Corrigan JM, Wolcott J, Erickson SM (2004) Patient Safety: Achieving a New Standard for Care. Committee on Data Standards for Patient Safety, The National Academies Press, Washington, D.C.

Basseville M, Nikiforov IV (1993) Detection of Abrupt Changes: Theory and Application. Prentice-Hall, Inc., Upper Saddle River, NJ, USA

Borg I, Groenen PJF (2010) Modern Multidimensional Scaling: Theory and Applications. Springer

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

Brandes U, Pich C (2007) Eigensolver methods for progressive multidimensional scaling of large data. In: Kaufmann M, Wagner D (eds) Graph Drawing, Springer Berlin Heidelberg, Lecture Notes in Computer Science, vol 4372, pp 42–53

Brockwell P, Davis R (2009) Time Series: Theory and Methods. Springer Series in Statistics, Springer

Cesario SK (2002) The "Christmas Effect" and other biometeorologic influences on childbearing and the health of women. J Obstet Gynecol Neonatal Nurs 31(5):526–535

Chakrabarti K, Garofalakis M, Rastogi R, Shim K (2001) Approximate query processing using wavelets. The VLDB Journal 10(2-3):199–223

Cruz-Correia RJ, Pereira Rodrigues P, Freitas A, Canario Almeida F, Chen R, Costa-Pereira A (2010) Data quality and integration issues in electronic health records. Information Discovery On Electronic Health Records pp 55–96

Csiszár I (1967) Information-type measures of difference of probability distributions and indirect observations. Studia Sci Math Hungar 2:299–318

Dasu T, Krishnan S, Lin D, Venkatasubramanian S, Yi K (2009) Change (detection) you can believe in: Finding distributional shifts in data streams. In: Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII, Springer-Verlag, Berlin, Heidelberg, IDA '09, pp 21–34

Endres D, Schindelin J (2003) A new metric for probability distributions. Information Theory, IEEE Transactions on 49(7):1858–1860

Gama J, Gaber MM (2007) Learning from Data Streams: Processing Techniques in Sensor Networks. Springer

Gama J, Medas P, Castillo G, Rodrigues P (2004) Learning with drift detection. In: Bazzan A, Labidi S (eds) Advances in Artificial Intelligence – SBIA 2004, Lecture Notes in Computer Science, vol 3171, Springer Berlin Heidelberg, pp 286–295

Gama Ja (2010) Knowledge Discovery from Data Streams, 1st edn. Chapman & Hall/CRC

Gehrke J, Korn F, Srivastava D (2001) On computing correlated aggregates over continual data streams. SIGMOD Rec 30(2):13–24

Guha S, Shim K, Woo J (2004) Rehist: Relative error histogram construction algorithms. In: In VLDB, pp 300–311

Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques. Elsevier : Morgan Kaufmann

Howden LM, Meyer JA (2011) Age and Sex Composition: 2010. 2010 Census Briefs US Department of Commerce, Economics and Statistics Administration, US Census Bureau

Hrovat G, Stiglic G, Kokol P, Ojstersek M (2014) Contrasting temporal trend discovery for large healthcare databases. Computer Methods and Programs in Biomedicine 113(1):251–257

Keim DA (2000) Designing pixel-oriented visualization techniques: Theory and applications. IEEE Transactions on Visualization and Computer Graphics 6(1):59–78

Kifer D, Ben-David S, Gehrke J (2004) Detecting change in data streams. In: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30, VLDB Endowment, VLDB '04, pp 180–191

Klinkenberg R, Renz I (1998) Adaptive information filtering: Learning in the presence of concept drifts. In: Workshop Notes of the ICML/AAAI-98 Workshop Learning for Text Categorization, AAAI Press, pp 33–40

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1):59–69

Lin J (1991) Divergence measures based on the shannon entropy. IEEE Transactions on Information theory 37:145–151

Mitchell TM, Caruana R, Freitag D, McDermott J, Zabowski D (1994) Experience with a learning personal assistant. Commun ACM 37(7):80–91

Mouss H, Mouss D, Mouss N, Sefouhi L (2004) Test of page-hinckley, an approach for fault detection in an agro-alimentary production system. In: Control Conference, 2004. 5th Asian, vol 2, pp 815–818 Vol.2

National Research Council (2011) Explaining Different Levels of Longevity in High-Income Countries. The National Academies Press, Washington, D.C.

NHDS (2010) United states department of health and human services. centers for disease control and prevention. national center for health statistics. national hospital discharge survey 2008 codebook

NHDS (2014) National Center for Health Statistics, National Hospital Discharge Survey (NHDS) data, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, available at: http://www.cdc.gov/nchs/nhds.htm

Papadimitriou S, Sun J, Faloutsos C (2005) Streaming pattern discovery in multiple time-series. In: Proceedings of the 31st International Conference on Very Large Data Bases, VLDB Endowment, VLDB '05, pp 697–708

Parzen E (1962) On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3):1065–1076

Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, New York

Rodrigues P, Correia R (2013) Streaming virtual patient records. In: G. Krempl, I. Zliobaite, Y. Wang, G. Forman (Eds.), Real-World Challenges for Data Stream Mining, Otto-von-Guericke, University Magdeburg, pp 34–37

Rodrigues P, Gama J, Pedroso J (2008) Hierarchical clustering of time-series data streams. Knowledge and Data Engineering, IEEE Transactions on 20(5):615–627

Rodrigues PP, Gama Ja (2010) A simple dense pixel visualization for mobile sensor data mining. In: Proceedings of the Second International Conference on Knowledge Discovery from Sensor Data, Springer-Verlag, Berlin, Heidelberg, Sensor-KDD'08, pp 175–189

Rodrigues PP, Gama J, Sebastião R (2010) Memoryless fading windows in ubiquitous settings. In: In Proceedings of Ubiquitous Data Mining (UDM) Workshop in conjunction with the 19th European Conference on Artificial Intelligence - ECAI 2010, pp 27–32

Rodrigues PP, Sebastião R, Santos CC (2011) Improving cardiotocography monitoring: a memory-less stream learning approach. In: Proceedings of the Learning from Medical Data Streams Workshop. Bled, Slovenia

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

Sebastião R, Gama J (2009) A study on change detection methods. In: 4th Portuguese Conf. on Artificial Intelligence

Sebastião R, Gama J, Rodrigues P, Bernardes J (2010) Monitoring incremental histogram distribution for change detection in data streams. In: Gaber M, Vatsavai R, Omitaomu O, Gama J, Chawla N, Ganguly A (eds) Knowledge Discovery from Sensor Data, Lecture Notes in Computer Science, vol 5840, Springer Berlin Heidelberg, pp 25–42

Sebastião R, Silva M, Rabiço R, Gama J, Mendonça T (2013) Real-time algorithm for changes detection in depth of anesthesia signals. Evolving Systems 4(1):3–12

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

Sáez C, Robles M, García-Gómez JM (2013) 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, pp 3226–3229

Sáez C, Robles M, García-Gómez JM (2014) Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Statistical Methods in Medical Research (forthcoming)

Shewhart WA, Deming WE (1939) Statistical method from the viewpoint of quality control. Washington, D.C. : Graduate School of the Department of Agriculture

Shimazaki H, Shinomoto S (2010) Kernel bandwidth optimization in spike rate estimation. J Comput Neurosci 29(1-2):171–182

Solberg LI, Engebretson KI, Sperl-Hillen JM, Hroscikoski MC, O'Connor PJ (2006) Are claims data accurate enough to identify patients for performance measures or quality improvement? the case of diabetes, heart disease, and depression. American Journal of Medical Quality 21(4):238–245

Spiliopoulou M, Ntoutsi I, Theodoridis Y, Schult R (2006) Monic: Modeling and monitoring cluster transitions. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '06, pp 706–711

Stiglic G, Kokol P (2011) Interpretability of sudden concept drift in medical informatics domain. 2010 IEEE International Conference on Data Mining Workshops 0:609–613

Torgerson W (1952) Multidimensional scaling: I. theory and method. Psychometrika 17(4):401–419

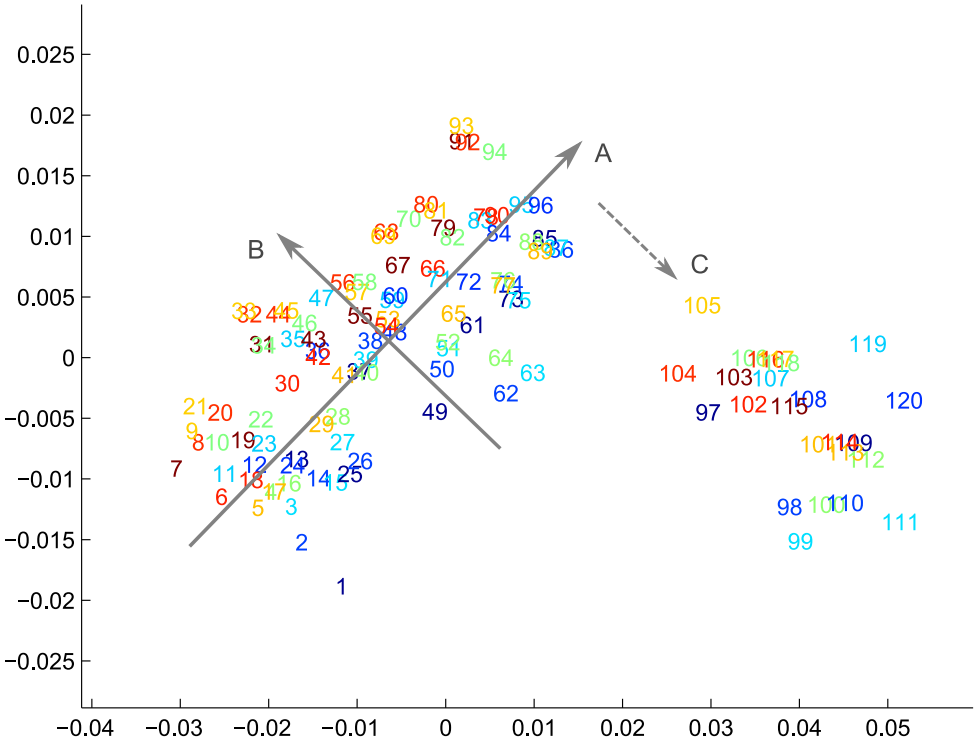
Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst 12(4):5–33

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

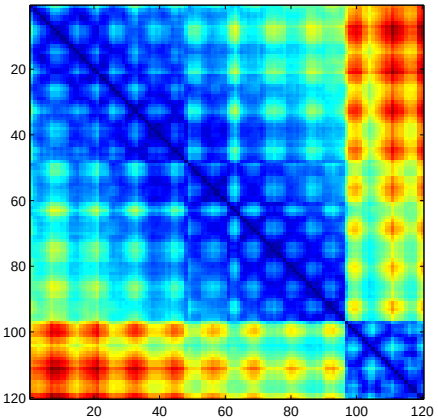
Wellings K, Macdowall W, Catchpole M, Goodrich J (1999) Seasonal variations in sexual activity and their implications for sexual health promotion. J R Soc Med 92(2):60–64

Westgard JO, Barry PL (2010) Basic QC practices: training in statistical quality control for medical laboratories. Westgard QC, Madison, WI

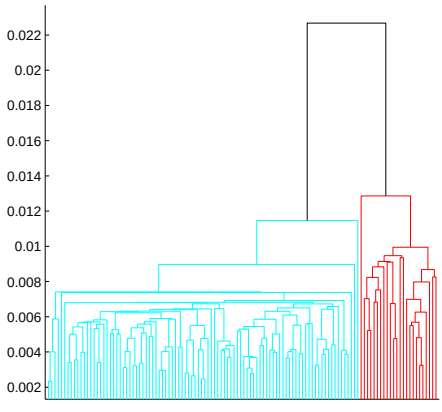
Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Machine Learning 23(1):69–101



(a) 2D information-geometric temporal plot (IGT-plot) of the statistical manifold of variable age, obtained with MDS from the probabilistic dissimilarity matrix. Points are represented by the index of the time window (months). Cooler and warmer colors are assigned to winter and summer months, respectively. Arrow A shows a temporal trend representing the gradual change. Arrow B represents the 12-month recurrent change. Finally, Arrow C represents the abrupt change separating the two temporal subgroups.

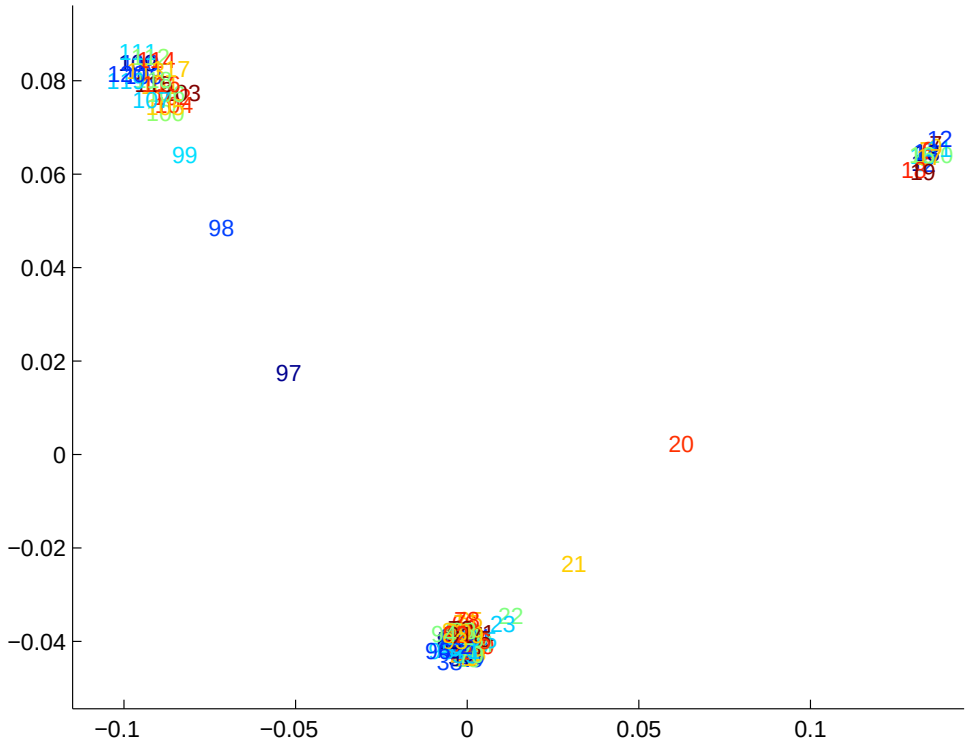


(b) Dissimilarity matrix heatmap

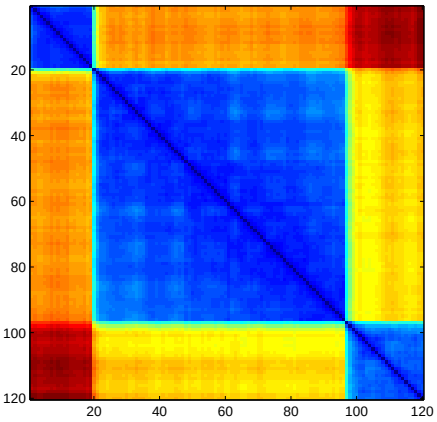


(c) Dendrogram for temporal subgroups

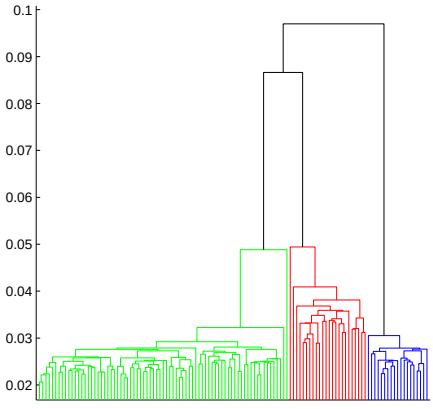
Figure 3: Change characterization and temporal subgroup discovery on the age variable. Change characterizations are shown by temporal trends on the projection. Two subgroups are clearly observed in the approximated statistical manifold (a) and dissimilarity matrix (b), which were confirmed with a complete linkage hierarchical clustering (c).



(a) 2D information-geometric temporal plot (IGT-plot) of the statistical manifold of variables age+sex with forced change at month 20 (obtained with MDS from the probabilistic dissimilarity matrix). Points are represented by the index of the time window (months). Cooler and warmer colors are assigned to winter and summer months, respectively.



(b) Dissimilarity matrix heatmap



(c) Dendrogram for temporal subgroups

Figure 4: Change characterization and temporal subgroup discovery on the joint age and sex variables with forced change at month 20. Change characterizations are shown by temporal trends on the projection. Three subgroups are clearly observed in the approximated statistical manifold (a) and dissimilarity matrix (b), which were confirmed with a complete linkage hierarchical clustering (c).