# Relative distances between homology groups to assess persistent defects in time series

Juan G. Diaz Ochoa
PerMediQ GmbH
Pelargusstr. 2, 70180 Stuttgart
March 31, 2019

Last: October 29, 2019

## Abstract

It is common to consider a data-intensive strategy to be an appropriate way to develop systemic analyses in biology and physiology. Therefore, options for data collection, sampling, standardization, visualization, and interpretation determine how causes are identified in time series to build mathematical models. However, there are often biases in the collected data that can affect the validity of the model: while collecting enough large datasets seems to be a good strategy for reducing the bias of the collected data, persistent and dynamical anomalies in the data structure can affect the overall validity of the model. In this work we present a methodology based on the definition of homological groups to evaluate persistent anomalies in the structure of the sampled time series. In this evaluation relevant patterns in the combination of different time series are clustered and grouped to customize the identification of causal relationships between parameters. We test this methodology on data collected from patients using mobile sensors to test the response to physical exercise in real-world conditions and outside the lab. With this methodology we plan to obtain a patient stratification of the time series to customize models in medicine.

**Key Words:** Time Series; Persistent Homology; Machine Learning; Medicine and Biology

## Introduction

High dimensional time series is are ubiquitous in several fields, from economics to biology and medicine. They are relevant because its combination helps to find relationships to infer causal relations between observations, for instance in the identification of the driving and driven systems from experimental time series (Paluš and Vejmelka, 2007). This allows the identification of the coupling between different parameters to derive and train descriptive and predictive models. Previous investigations focused on the definition of causality tests using time series (Paluš and Vejmelka, 2007), for instance using transfer entropy (Mao and Shang, 2017). But the understanding of causalities and the successful implementation of models requires a well-founded analysis of the influence of the sampled data[1], and in some cases *the inference of causality can be complicated by a bias in estimation from a limited amount of possibly noisy data* (Paluš and Vejmelka, 2007).

[1] https://arxiv.org/abs/1703.04730v2

This inference of causality is based on the notion of cooperative behavior of coupled complex systems, where synchronization and related phenomena has been observed in physical as well as biological systems (Pikovsky, 2008). Indeed, recently it has become very useful to deal biological systems from the point of view of theoretical informatics. For example, a bird flying in a forest *compute* their trajectory $\Gamma$ according to the tree distribution in the environment. Honeybees are also computing and creating a model of their environment (Roper et al., 2017). But also, a cancer cell adapting its response also represented by $\Gamma$ according to changes in its microenvironment, i.e. depending on the acidity of the tissue or even to toxic chemical substances for the cell (Rossetti et al., 2012). Such representations are useful considering that an environment is non-static: a storm can change the distribution of trees in a forest, or the change of a diet can induce substances able to modify microenvironments in tissues affected by cancer cells (Diaz Ochoa, 2018).

We estimate persistent homology groups to qualitatively assess persistent incoherencies and disbalances in the sampled data associated to the trajectories $\Gamma$. Since this methodology is robust against noise (Emrani et al., 2014), it is best suited to detect persistent defects in the sampled $\Gamma's$. Such disbalances are more than errors in the sampling of data and can be identified as persistent and inherent characteristics of the trajectories $\Gamma$. Such qualitative assessment is not only relevant for the optimization of modelling methods, e.g. avoiding expensive training of models, but also to assure the safety in the use of models by recognizing when a sample of data from a biological system or organism can be represented with a common underlying model, or instead requires a customized mathematical representation, which is for instance helpful to determine if personalization of relevant mathematical models is required for diagnose and therapy in medicine .

To test this kind of stratification based on a persistent bias we use the Mhealth dataset[2], which contains data from patients wearing IoT[3] sensors connected to internet devices to measure electro cardiograms (ECGs) and acceleration while they perform physical exercise in normal and non-controlled conditions. In the next section we introduce the required mathematical background. In section three we test this concept using the MHEALTH data set, for patients wearing IoT devices measuring EKGs and acceleration when they perform physical exercise. Finally, we discuss the resent results and perspectives.

# Topological methods for the assessment of bias in the sampled data

## Definition of persistent bias

According to the bias-variance decomposition derivation, the error of a model $\hat{f}$, $Error(\hat{f})$, is composed of three terms: a bias that depends on the definitions of the modeler, a variance term, and an unavoidable *irreducible error* term and is given by[4] $Error(\hat{f}) = E\left[\left(\bar{\Gamma} - \hat{f}\right)^2\right] = Bias(\hat{f}^2) + var(\hat{f}) + \sigma^2$, where $Bias(\hat{f}^2)$ is the bias of the model $\hat{f}$.

In this concept the bias is assigned to the way how the data is collected. But the specific reaction to the environment can induce persistent bias in the data structure: the way in which the organism $\boldsymbol{k}$ autonomously explores its environment, and differs respect other organisms $\boldsymbol{l}$, has an effect in the variability of the estimated error of a model, such that[5] (see the appendix 1)

$$\Delta^{kl} Error(\hat{f}) = Bias^{kl}[\bar{\Gamma}^2] - 2\hat{f} Bias^{kl}[\bar{\Gamma}], \tag{1}$$

where $Bias^{kl}[\bar{\Gamma}]$ can be defined as (see appendix 1):

[2] http://archive.ics.uci.edu/ml/datasets/mhealth+dataset
[3] Internet of things
[4] http://scott.fortmann-roe.com/docs/BiasVariance.html

$$Bias^{kl}[\bar{\Gamma}] = E[\bar{\Gamma}^k - \bar{\Gamma}^l]; \quad (2)$$

here $E$ is the expected value of the difference of the trajectory $\bar{\Gamma}^k$ of the organism $k$ respect the trajectory $\bar{\Gamma}^l$ of the organism $l$, $\bar{\Gamma}^k - \bar{\Gamma}^l$. This basically is a perturbation of the error respect the trajectories computed by other organisms generating a response to the environment.

From equation 2 we concentrate us on the analysis of $Bias^{kl}[\bar{\Gamma}]$, since we want to analyze the relative changes of trajectories between organisms $k$ and $l$ leading to this variability of the error.

We observe two different cases:

- Considering, the trajectory $\Gamma^k$ can differ from another sampled trajectory $\Gamma^l$, the equation 2 represents a persistent bias for each sampled trajectory respect other trajectories and is a representation about how the organism $k$ sees the environment, and how this perspective differs from other organisms with their own representation, such that,

  $$Bias^{kl} > 0, \quad (3)$$

  i.e. the organism generates incoherent trajectories respect other organisms.

- On the other hand, when all the trajectories of all the organisms are coherent (for instance the trajectory of a birth in a tree labyrinth is similar to the trajectory of other birds), then

  $$Bias^{kl} \to 0; \quad (4)$$

  this implies either that the environment is static, or that under high variable environments the organism is rapidly adapting. In the first case the system can be modelled, i.e. there is a function$\hat{f}$ representing a trajectory coupled to an environment (which minimizes a given landscape $U_\lambda$). This allows the modelling of the trajectory $\bar{\Gamma}$, assuming that $\bar{\Gamma}$ is an observable that somehow represents a reaction to its environment.

Thus, this persistent bias is a natural element in the perturbation of $Error(\hat{f})$. This implies that the definition of a valid and stable model does not only depends on the ability of the modeler selecting the best possible model, but also depends on the coherence of the data structure.

## Topological persistence: separation of internal bias from statistical error and modelling

To estimate whether $Bias^{kl} \to 0$ (see equation 4) we need to assess the structure of the sampled data. The strategy we propose is to assess the topological structure of the data before a model or regression is performed, ideally combining different trajectories in a phase space.

The strategy selected in this work consists in the construction of point clouds $w_1^k = \left(\Gamma_1^k(t), \Gamma_2^k(t), \Gamma_3^k(t), \dots \dots\right)$ generated from the normalized trajectory sample $\{\bar{\Gamma}^k\}$ of the organism or system $k$. A point cloud includes a large but finite set of points sampled from a primary form. In the theory t is well known that the combination of the time series of the trajectories $\bar{\Gamma}$, including the time delay of time series, can recover the dynamics of the system (Takens, 1981). Furthermore the presence of harmonic structures in the data, related to this dynamics, can be explored analyzing homological persistence (Emrani et al., 2014). Persistent homology, a tool in algebraic topology, is particularly useful in these situations where the "scale" is not known a priori. It may be viewed as a generalization of

hierarchical clustering for higher order topological features, which provides a summary using a bar plot, which essentially validates the different groups discovered using topological invariance (Emrani et al., 2014)[6] [7].

Since $\Gamma^{\lambda}$ owns a topology that reflects the periodic behavior of a signal with Euler characteristics, this means $\Gamma^{\lambda}$ owns a function $g$ with a compact subset of $\mathbb{R}^D$ and $d_K: \mathbb{R}^D \to \mathbb{R}$ the distance function of $K$. In the following we employ a similar notation as defined by Fasy et. al. (Fasy et al., 2014) for the persistence bars. Consider the *set* $L_t = \{x: d_u(x) \leq t\}$ and $k = L_0$. As $t$ varies from 0 to $\infty$, the set $L_t$ changes. $L_t$ , defined as a persistent bar, is characterized with persistent topology and summarizes how the topological characteristics $L_t$ change as a function of $t$. Key topological features include zero and first order topology (see figure 1).

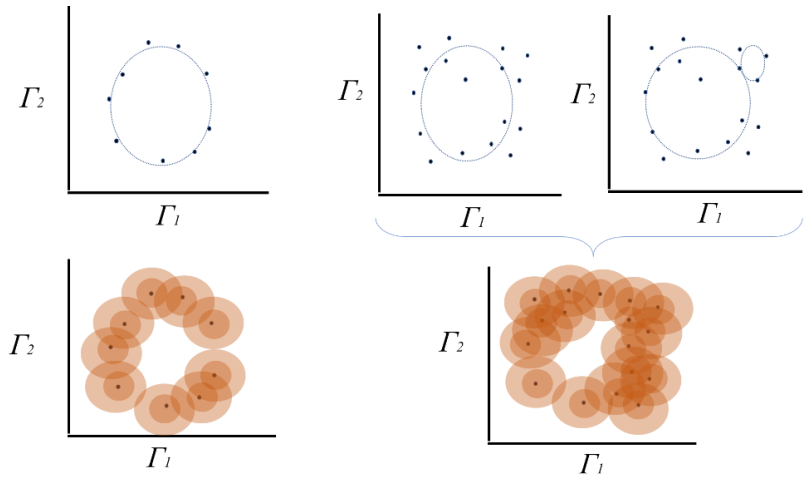



*Figure 1 Estimation of topological persistence. In the example left connected points and a surface with homology 0 can be extracted. In the right the same procedure allows the discovery of a second structure with homology 1.*

For $\lambda = 1,2,\cdots \Lambda$, if there is a sample of trajectories generating a point cloud $\{w^{\lambda}\}$ then $L_t\{w^{\lambda}\}$ are different topological persistent characteristics of the trajectory $\Gamma^{\lambda}$, such that when equation 3 is combined then,

$$\text{if } Bias^{kl}(\bar{\Gamma}^2) \geq 0, \text{then} \quad E\left(L_t\{w^k\} - L_t\{w^l\}\right) = \mathcal{M}^{kl} \geq 0, \tag{5}$$

which is equivalent to low change of $\Gamma^k$ respect $\lambda$ implying a preservation of the internal structure $\widehat{M}$ used to compute the trajectory (persistent internal bias). Otherwise,

$$\text{if } Bias^{kl}(\bar{\Gamma}^2) \to 0, \text{then} \quad E\left(L_t\{w^k\} - L_t\{w^l\}\right) = \mathcal{M}^{kl} \to 0, \tag{6}$$

i.e. organisms show similar patterns or a similar response, i.e. $\widehat{M}$ is equivalent for each organism. The matrix $\mathcal{M}^{kl}$ will be called in what follows distortion matrix.

[6] https://cran.r-project.org/web/packages/TDA/index.html
[7] https://cran.r-project.org/web/packages/TDA/vignettes/article.pdf

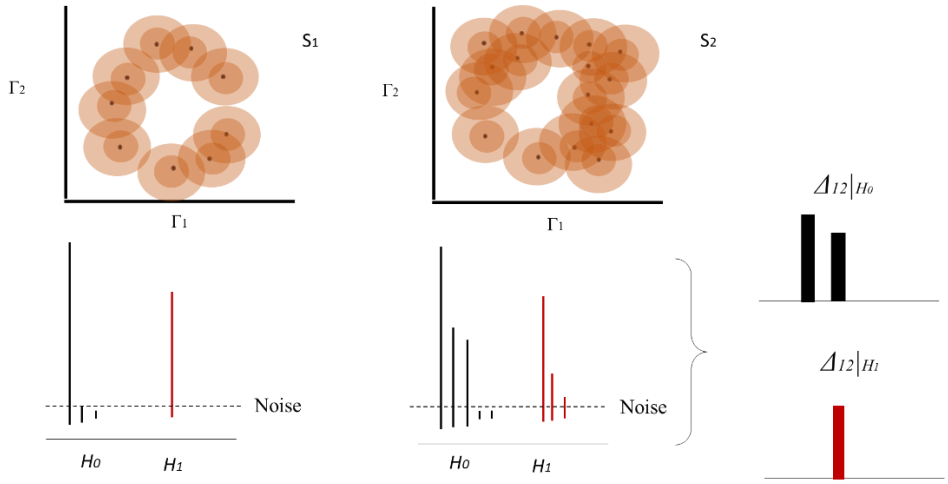



*Figure 2 Homology group and bar diagram for two different data sets based on example in figure 1. For $S_1$ the persistent dyagram delivers a homology group $H_1$, i.e. a group with genus =1. For $S_2$ the increase in the number of points allows the defintion of a homology gropup $H_0$, i.e. a surface with connected points. On the left we schematically represent the difference between both persistent diagrams (for each homology group)*

The goal is to make a relative assessment of the topological persistence of a cloud of points. Persistent characteristics as well as pure noise are not of central interest. Instead, persistent topological characteristics, computed as the difference of the persistent bars, or equivalently as the difference between homology groups, provide information about similarity between organisms as well as potential internal bias. Accordingly:

> *low relative persistence of data, i.e. $\mathcal{M}^{kl} \geq 0$, implies a high probability that a customized model $\hat{f}_k$ is required, i.e. that $\hat{f}_l$ will probably not completely fit the sampled data of k*

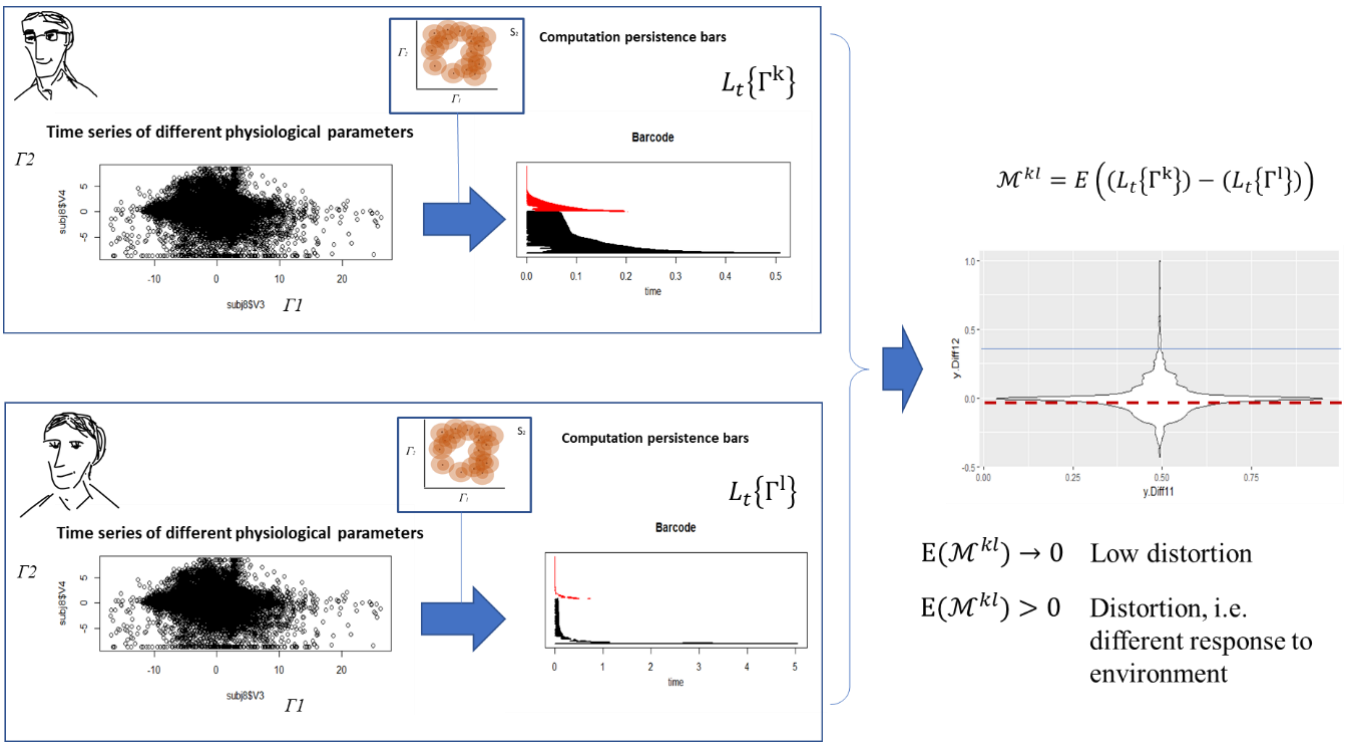



*Figure 3 Different steps in the analysis of the matrix $\mathcal{M}^{kl}$according to equation 8. For example, the data from two individuals is analyzed as a point cloud in a phase space using topological invariance. Once the persistent diagrams are obtained they are normalized respect the highest value of the persistence diagram. The final normalized distribution of the persistence diagram of each individual is subtracted. If the expected value is about 0, then the distortion is minimal, i.e. there is no significant deviation between the patterns of two individuals. Otherwise, a*

*distortion can be identified. Since we deal with the difference between distributions, we employ box plots or violin diagrams to estimate the relative difference between two individual patterns.*

In figure 3 we describe the main steps of this algorithm: each individual generates her own set of trajectories that produce a point cloud $w^{\lambda}$. Using topological homology analysis, the homology is computed for each group, and the persistence is computed using a cluster analysis for each point cloud. This computation essentially extracts the fingerprint from the data generated from each individual. Finally, the relative distance between these fingerprints, i.e. distance between homology groups, $E\left(L_t\{w^{\mathrm{k}}\} - L_t\{w^{\mathrm{l}}\}\right) = \mathcal{M}^{kl}$ is estimated. Due that we compute a difference of the silhouette of the clustering analysis, the final value $\mathcal{M}^{kl}$ is better represented by a histogram. For this reason, a box plot is used to represent the distribution of $\mathcal{M}^{kl}$ (see appendix 2). The skewness of this distribution provides relevant information about the stability of the perturbation at $\mathcal{M}^{kl}$.

## Analysis of patient data generated from IoT

We test the current method in the analysis of physiological signals of patients. As has been suggested in other studies in animals, the locomotor activity is associated to an integrated increase in the activity of different physiological signals (like heart rate, arterial pressure, etc.) (Miyata et al., 2016). Furthermore, the heart response to exercise (macroscopic scale) has origin in complex molecular mechanisms; *for instance in subjects undergoing investigation for angina, individuals with a low chronotropic index (a measure of heart rate response that corrects for exercise capacity) had impaired endothelial function, raised markers of systemic inflammation, and raised concentrations of N-terminal pro-brain natriuretic peptide (NT-proBNP) compared to those with a normal heart rate response* (Routledge and Townend, 2006).

Based on this concept we assume that each individual generates a unique pattern for this integrated response, due to the individual capacity -across several scales- to adapt and/or accommodate to changes in the environment, similar to the case presented in figure 3 (in this case the response to physical exercise).

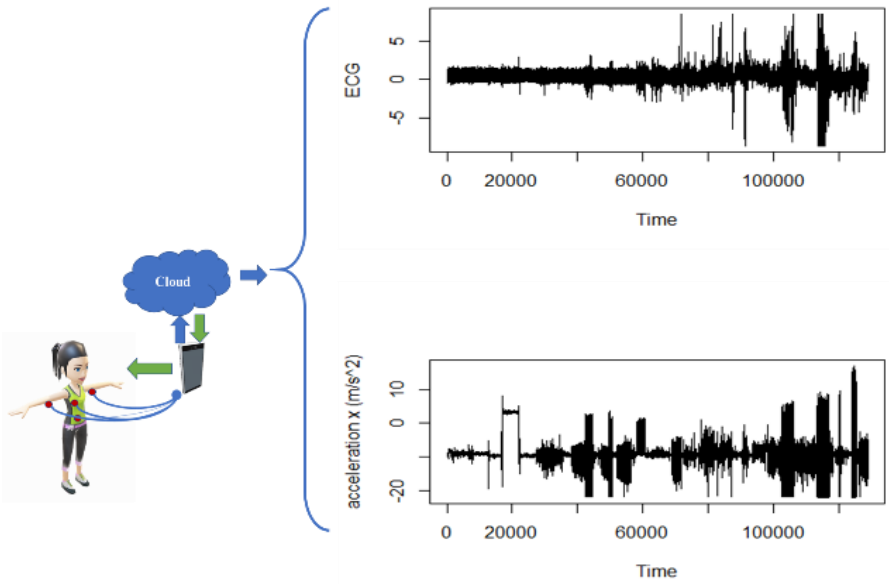
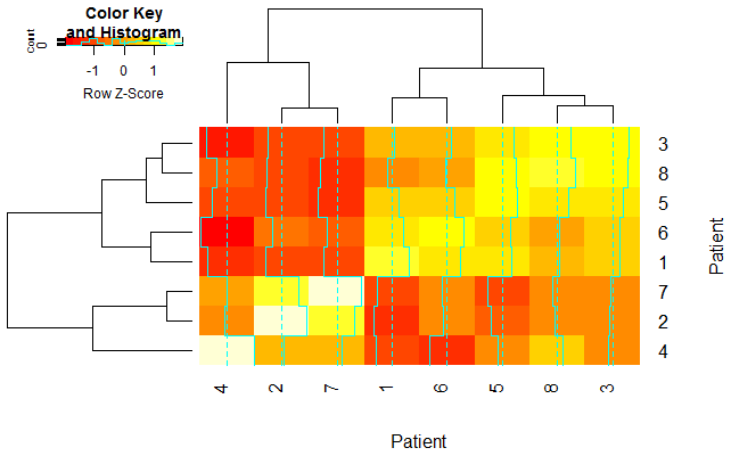



*Figure 4 Exemplary data obtained from patients wearing sensors to measure ECG and acceleration in three axes (x, y, z)*

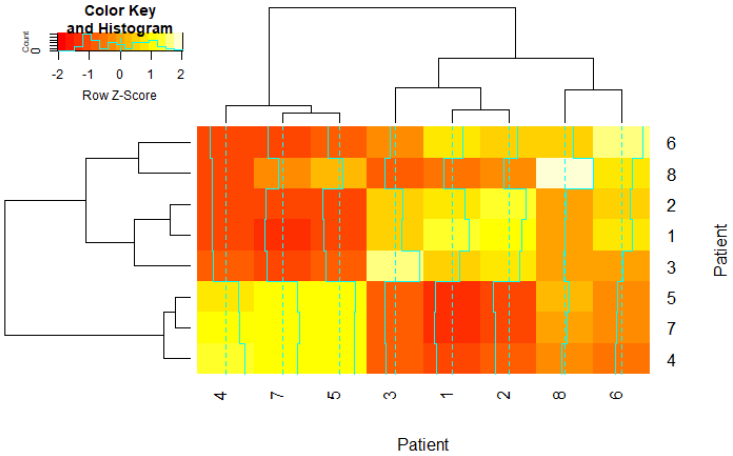
The data used in this analysis has been obtained from the MHEALTH (Mobile HEALTH) dataset, which comprises body motion and vital signs recordings for ten volunteers of diverse profile while performing several physical activities. Sensors placed on the subject's chest, right wrist and left ankle are used to measure the motion experienced by diverse body parts, namely, acceleration, rate of turn and magnetic field orientation. The sensor positioned on the chest also provides 2-lead ECG measurements, which can be potentially used for basic heart monitoring, checking for various arrhythmias or looking at the effects of exercise on the ECG. The activities were collected in an out-of-lab environment with no constraints on the way these must be executed, with the exception that the subject should try their best when executing them[8] (see figure 4). The final raw data is analyzed, i.e. features are extracted from the time series and the main qualitative features are analyzed using a heat map (see appendix 2 for an explanation about how the heat maps in figure 5 are constructed).

The relative distance between the homology groups $\mathcal{M}^{kl}$ is shown in detail in appendix 2. To facilitate the analysis of these results, the correlation between values have been estimated using a Pearson product-moment correlation coefficient (figure 5).

***X axis***




***Y axis***




***Z axis***

[8] http://archive.ics.uci.edu/ml/datasets/mhealth+dataset
Se also this reference for the obtention and use of this data
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4547155/

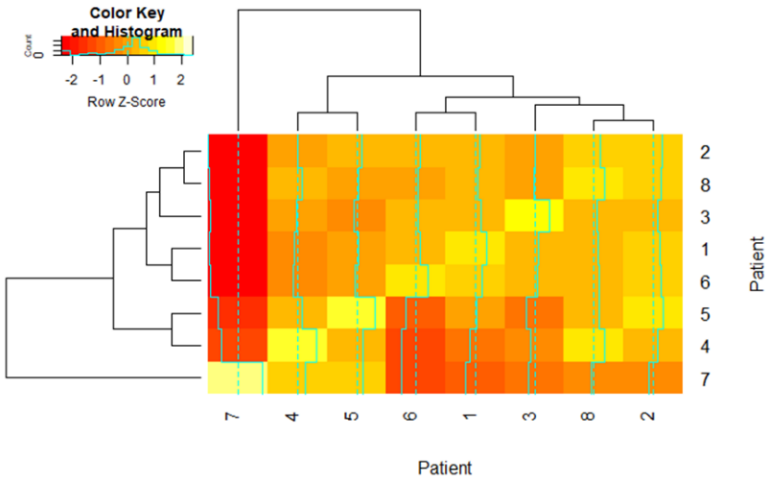



*Figure 5 Application of the methodology introduced in fig. 3 for the analysis of physiological data of different patients: once the ECG and acceleration data of a patient doing physical exercise is acquired by IoT devices and stored for instance in the cloud. The correlation of the relative difference of the homology groups (appendix 2) is analyzed for four different patients considering that the acceleration is measured on the x axis (B.), y axis (C.) and z axis (D).*

From this analysis we perform an analysis linking the acceleration measured with a chest sensor in relation to the ECG, and apply the methodology described in the previous section and in figure 3. We observe some similarities, while a rich structure is revealed when the analysis is extended to each axis where the acceleration is measured. A relevant result is that the structure of the clustering is almost equivalent for the three axes, implying that the method delivers a consistent stratification of the individuals.

With the dendrograms analysis we observe the formation of clusters in the homology groups for the patients 7 and 4, implying that there is a correlation between both probands, and in particular that in both cases the relative distance between their homology groups is low. Otherwise $\mathcal{M}^{7l} \geq 0$ as well as $\mathcal{M}^{4l} \geq 0$ (for the patient 5 this affirmation is non conclusive), or according to equation 5, $Bias^{7l}(\bar{\Gamma}^2) \geq 0$ and $Bias^{4l}(\bar{\Gamma}^2) \geq 0$. Hence, any inference of causal relations between observations must be differentiated between the patients 7 and 4 and the rest of the individuals in this dataset.

The study is highly qualitative and is reduced to the quantification of relative differences of the data structure in the heatmap, keeping in mind the use of this application for fast data assessments. However, all the assessment delivers more data than the simple qualitative result using homology groups, since it is basically a kind of unsupervised machine learning detecting structures in the time series, which can be used for the computation of prognoses. The use of this methodology will be further explored in next studies.

# Conclusion

We have introduced a qualitative method to assess persistent bias in sampled data. Instead of cumulating and managing very large data sets, it seems to be a better strategy to first recognize which data collections are appropriate and balanced to train models that can be validated and are reliable for further extrapolation. This means, an appropriate customization of models ab initio after assessing persistent bias is more efficient than the training of universal models on several data sets that will be problematic in its validation (Taniguchi et al., 2018)[9].

To deal with this problem it could be of great advantage to measure the complexity of the data to be analyzed, using for instance a Kolmogorov or a Chaitin complexity measurement (Adriaans, 2018),

[9] Therefore, we think that any research in machine learning do a better job by dealing with the natural symbiosis between information and life sciences, rather than try to simulate or imitate human cognitive capabilities.

together with the design of alternative learning architectures. We think that more comprehensive methods and techniques for machine learning are required. The problem is not only the risk that biases and disbalanced information get into big-data sets used to train machines, but it is also the ever-increasing use of resources: the growing processing and storage of information require a lot of energy and resources that end up in form of waste and green-house gases in the atmosphere[10].

Using concepts of the theoretical informatics and assuming that organisms compute their trajectories after processing information from the environment, and that they adjust to their environment and eventually evolve (Diaz Ochoa, 2018), we develop a method to qualitatively detect this complex behavior by measuring the variability of the modelling error. If the data obtained from any organism's trajectory has a persistent structure, then the variability of the modelling error is low, implying that a model can be derived and trained. Otherwise, there is not only a bias that can be assigned to the modeler, but also a persistent bias which deteriorates the quality of the statistics.

We suggest that by performing this kind of qualitative analysis before any data processing helps to recognize and introduce convenient data imbalance that not only contributes to a better understanding of the problem (recognition of entities as modelers, and not simply as a blind reaction mechanisms), but also avoids the expensive training of models.

For the detection of persistent structures in trajectories, we have implemented methods using persistent topology for the analysis of time series (Maletić et al., 2016), which have become a promising way to detect patterns in data different to entropy based methods. Despite the implementation of persistent homology is not straightforward, it shows the possibility to better select the data for training and points to the possibility to introduce clever bias in the models to reduce the amount of training data. The concrete application in the analysis of physiological data is a contribution to customize models (for diagnose and therapy) in patients. However, more work is required to standardize and test this methodology in additional datasets.

Thus, we believe that this method is a way of looking at the model through the lens of the training data[11], and helps to better assess the derivation of causal structures from data for model derivation in medicine and biology, and in general in all the field of Machine Learning.

## Acknowledgments

We are thankful for the critical feedback of Elena Ramirez, as well as useful comments provided by Bernardo Huberman to this manuscript. I also acknowledge the feedback of an anonymous referee.

## Conflict of interest

Author Juan G. Diaz Ochoa is working for PerMediQ GmbH.

# Bibliography

Adriaans, P. (2018). Information. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. (Metaphysics Research Lab, Stanford University), p.

Diaz Ochoa, J.G. (2018). Elastic Multi-scale Mechanisms: Computation and Biological Evolution. J. Mol. Evol. *86*, 47–57.

[10] Regarding increasing concentration of greenhouse gases in the atmosphere, among them 6% are generated from computation and It, it results extremely important to develop methods to reduce the ecological impact of performing Machine learning.

[11] https://arxiv.org/pdf/1703.04730v2.pdf

Emrani, S., Gentimis, T., and Krim, H. (2014). Persistent Homology of Delay Embeddings and its Application to Wheeze Detection.

Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. The Annals of Statistics *42*, 2301–2339.

Maletić, S., Zhao, Y., and Rajković, M. (2016). Persistent topological features of dynamical systems. Chaos: An Interdisciplinary Journal of Nonlinear Science *26*, 053105.

Mao, X., and Shang, P. (2017). Transfer entropy between multivariate time series. Communications in Nonlinear Science and Numerical Simulation *47*, 338–347.

Miyata, K., Kuwaki, T., and Ootsuka, Y. (2016). The integrated ultradian organization of behavior and physiology in mice and the contribution of orexin to the ultradian patterning. Neuroscience *334*, 119–133.

Paluš, M., and Vejmelka, M. (2007). Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections. Physical Review E *75*, 056211.

Pikovsky, A. (2008). Synchronization: Universal Concept: A Universal Concept in Nonlinear Sciences (Cambridge: Cambridge University Press).

Roper, M., Fernando, C., and Chittka, L. (2017). Insect Bio-inspired Neural Network Provides New Evidence on How Simple Feature Detectors Can Enable Complex Visual Generalization and Stimulus Location Invariance in the Miniature Brain of Honeybees. PLOS Computational Biology *13*, e1005333.

Rossetti, S., Esposito, J., Corlazzoli, F., Gregorski, A., and Sacchi, N. (2012). Entrainment of breast (cancer) epithelial cells detects distinct circadian oscillation patterns for clock and hormone receptor genes. Cell Cycle *11*, 350–360.

Routledge, H.C., and Townend, J.N. (2006). Why does the heart rate response to exercise predict adverse cardiac events? Heart *92*, 577–578.

Takens, F. (1981). Detecting strange attractors in turbulence. In Dynamical Systems and Turbulence, Warwick 1980, D. Rand, and L.-S. Young, eds. (Springer Berlin Heidelberg), pp. 366–381.

Taniguchi, H., Sato, H., and Shirakawa, T. (2018). A machine learning model with human cognitive biases capable of learning from small and biased datasets. Scientific Reports *8*, 7397.

# Appendix 1

The following is the detailed derivation of the expression linking the variation of error and internal bias:

$$\begin{aligned}\Delta^{kl}Error(\hat{f}) &= \Delta^{kl}E\left[\left(\bar{\Gamma}-\hat{f}\right)^2\right] = E\left[\left(\bar{\Gamma}^k-\hat{f}\right)^2\right] - E\left[\left(\bar{\Gamma}^l-\hat{f}\right)^2\right] \\ &= E\left[\bar{\Gamma}^{k^2} - 2\hat{f}\bar{\Gamma}^k + \hat{f}^2\right] - E\left[\bar{\Gamma}^{l^2} - 2\hat{f}\bar{\Gamma}^l + \hat{f}^2\right] \\ &= E\left[\bar{\Gamma}^{k^2} - \bar{\Gamma}^{l^2}\right] - 2E\left[\hat{f}\bar{\Gamma}^k - \hat{f}\bar{\Gamma}^l\right] \\ &= Bias^{kl}[\bar{\Gamma}^2] - 2\hat{f}Bias^{kl}[\bar{\Gamma}]\end{aligned}$$

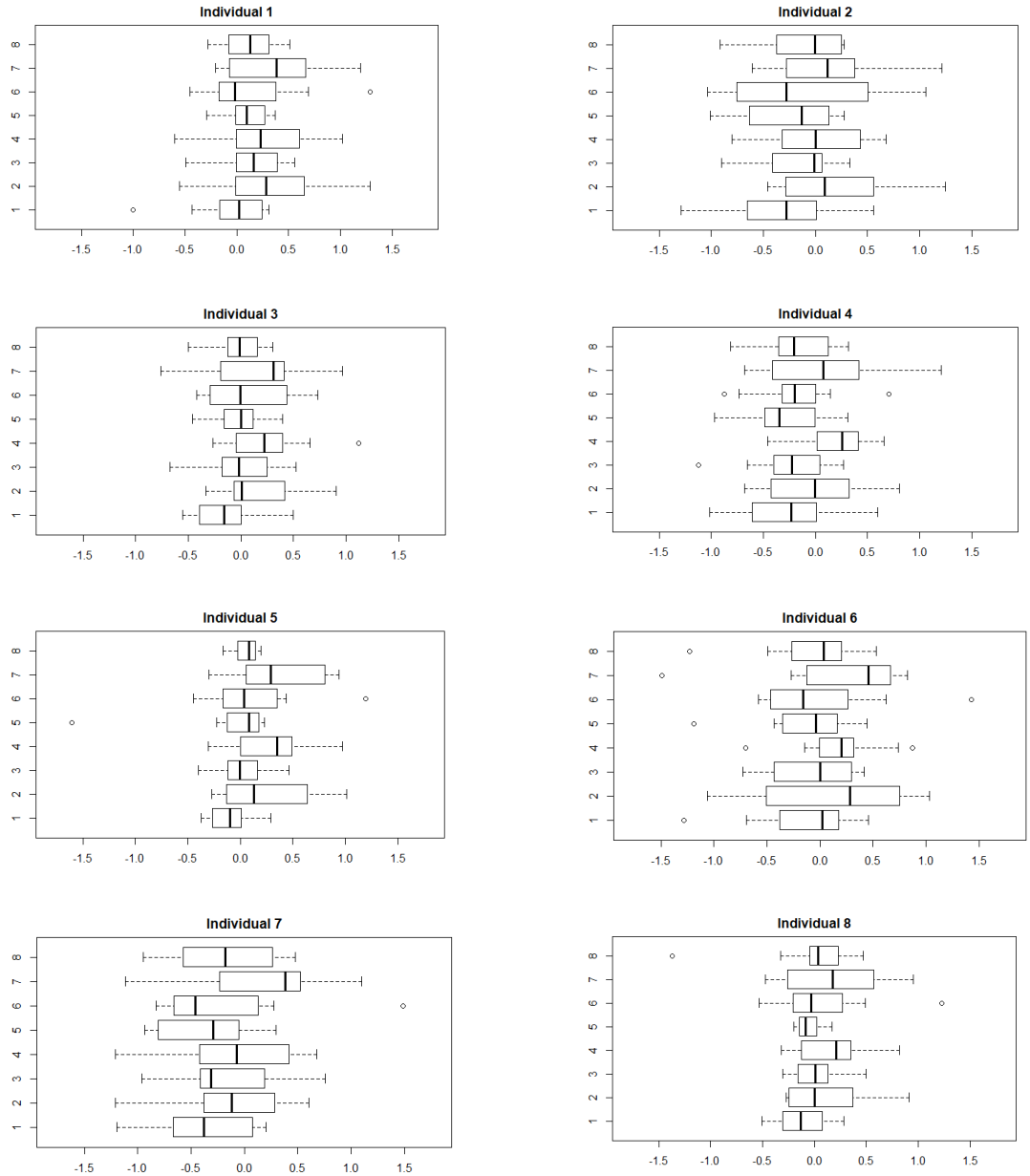
# Appendix 2

Comparation of distortion matrix: in this proposal we employ the following system

- If we compare one patient with himself to assess own changes in the own topological patterns, then we compare the persistence between two different time periods
- If we compare one system against other systems, then we compare the topological persistence of the time series of both individuals in the same time period

In this part we compute the length of the persistence bar as well as the relative difference between the persistence bars. The sampled box plots below (table 1) represents the distribution of the relative differences of persistent features between individuals when accelerations are measured on the *x* axis. Observe that values that deviate from 0 imply that data has an implicit structure that diverge for the specific individual.

*Table A2-1 Distribution of the relative distances of the persistence bars between different patients performing physical exercise. Accelerations are measured on the x axis*



A comprehensive analysis of this table is made by computing the median of the distributions in the boxplots. We use the median because it is robust against extreme outliers in the distributions presented in the box plots. The final result is presented in the graphic below. Observe that according to this result, individuals 2, 4, and 7 have a different data structure than the rest of individuals, pointing to a persistent bias.

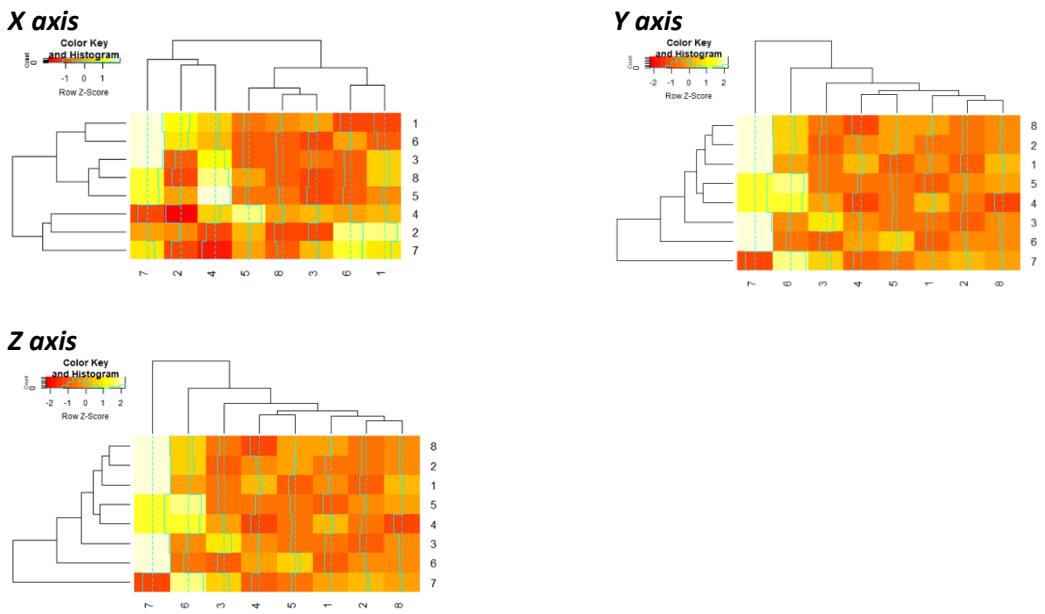



*Figure A2 – 1 Heatmap for the distribution of the medians of the boxplots from table A2-1*