Objectivity and subjectivity in the analysis of 3D movement data

I. The data characterization stage: a general picture of the situation

Pierre Loslever
Laboratory of Human and Industrial Automatics, Mechanics and Informatics of Valenciennes University
LAMIH, Valenciennes, France
Pierre.Loslever@univ-valenciennes.fr

***Abstract*— This part I proposes two ways for the design of the data characterization stage evaluation : 1) a way based on the computation of distances between signals and distances between indicators yielded by the characterization stage, 2) a way based on individual opinions about the signal resemblance and dissemblance. These two investigation ways are considered through a didactic example of 10 signals with simulated values. Despite the first way seems "objective", we show that it is the contrary as it is the case is in each paper about data analysis. In fact, the subjectivity is always present but, more or less, the main problem being to be aware of this !**

Keywords-component; data chacterization; data summarizing; objective data; subjective data; evaluation

I. Introduction

Thanks to the increasing performances of computers and measurement devices, more and more disciplines (biomechanics, ergonomics, psychology, physiology, medicine, …) are able to collect more and more movement data in a given study. The "relative easiness" to use measurement devices giving 3D data (imaging systems, force plates, goniometers,…) involves that these devices can be used in experimental studies, but in clinical and field studies also.

On the opposite, a "relative difficulty" is the large number of data pieces yielded by such measurement systems which involves more than one stage to get results. Up to 5 stages can be present:

1) a stage that characterizes the data through either more or less synthetic indicators (overall indicators such as the arithmetic mean or RMS value, energies from Fourier decomposition, frequencies associated to magnitude intervals...) or local indicators (maximum and/or the time corresponding to this value, …) [1]. The indicators become the data analysis variables;

2) a stage that transforms the data so that the analysis variables become compatible with a specific method (e.g. a quantitative scale can be changed into an ordinal scale in order to use rank based methods) and/or with each other in the perspective of a multivariate analysis (e.g. when there are both quantitative and qualitative scales) [2} ;

3) a stage that shapes the data set so that it can be investigated using monovariate or multivariate techniques. For instance, the data is arranged in a table where the rows correspond to the statistical units (individuals, pathologies, time samples, ...) and the columns to the indicators [2];

4) a stage that yields relations (between measurement variables or between factors and measurement variables), classes (of pathologies, or behaviors, ...), references (average levels, probabilities, time excursions, ...), etc. This stages yields up to 3 kinds of models: graphical (dendrogram, factor plane, ...), verbal (the conclusion to a statistical hypothesis test) or mathematical (multivariate regression, time series, ...) [1] [2];

5) a stage that shows the results in an other way than with the previous stage. For instance, in factor analysis, the contents of the tables that aid interpretation and the factor planes are verbally expressed and easier graphic can be shown (histogram, time excursion, spectrum, scatterplot, …) [3].

In the following, a statistical analysis path p is a specific quintuplet of methods for the 5 stages (if a stage does not exist, the method is named "0") and yields the result set $\mathbb{R}_p$. Thus, given a initial data set and the combinatory of the possibilities for the 5 stages, many result sets can be obtained. In this part I, we focus on the first stage and show the high diversity and thus subjectivity in the way to built the data analysis variables. To achieve this aim we will consider a simple simulated data set In part II, we focus on the two last stages.

II. Problem statement : Characterization evaluation

An empirical study (experimental, field or clinic) yields a data set $\mathbb{E}_0$ card($\mathbb{E}_0$)= E_0 values (notations : 1) *card* function gives the number of elements in a set and 2) the same letter is used to indicate the set and its size). Let us suppose that the experimental or observation design yields N multidimensional signals obtained from a 3D measurement system with the sampling frequency fs, each multidimensional signal with P components. In most cases the data base is large: N>50, P>20, fs>10 Hz (e.g. a gait study with N=50, 25 individuals and 2 trials per individual, P=50 and fs=50Hz). Then many characterization methods can be used [4], each method c (c=1, …C) yielding a data set $\mathbb{E}_{1c}$ containing card($\mathbb{E}_{1c}$)= E_{1c} values. Two main quantitative descriptors can be introduced to characterize the method c:

- the data reduction level r_c which is the ratio between the output and input numbers of values, i.e. $r_c= E_{1c}/E_0$,

- the duration d_c, which mainly combines the time spent choosing a method (human analysis time, named da_c) and the processing time (treatment time dt_c), Fig. 1, thus $d_c=da_c+dt_c=t_{1c}$.

These two criteria (named CR_1 and CR_2) are concrete but give a poor idea about the "characterizing performance". To achieve this aim, many methods can be proposed [1] but, due to its easiness, the following procedure is considered:

1) compute the distances between the input subsets, which yields the set $\mathbb{D}_0=\{d_0(\mathbb{E}'_{0i}, \mathbb{E}'_{0j}),\ i=1,\dots, N-1,\ j=i+1, \dots, N)\}$, $D_0=card(\mathbb{D}_0)=N*(N-1)/2$,

2) for each characterizing method c (c=1, …, C), compute the distances between the output subsets, which yields the set $\mathbb{D}_{1c}=\{d_{1c}(\mathbb{E}'_{1i}, \mathbb{E}'_{1j}),\ i=1,\dots, N-1,\ j=i+1, \dots, N)\}$, $D_{1c} = D_0$,

3) for each characterizing method c (c=1, …, C), put into relation the two sets $\mathbb{D}_0$ and $\mathbb{D}_{1c}$.

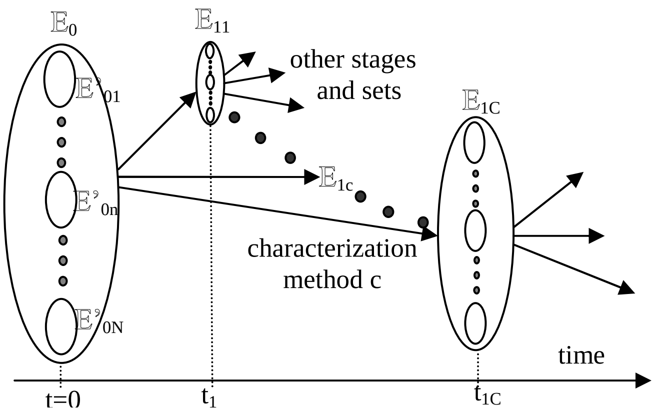

Figure 1. a way to consider the data characterization problem

We are aware that for each of the 3 steps, many choices can be made. Moreover, instead of using this quantitative approach, it is possible to perform a procedure based on expert opinions (which could give quantitative and/or qualitative data). Given that for these two kinds of approaches, the choices are highly subjective, we propose to name them objective and subjective approaches and suggest an illustration with easy data. Thus we propose to deal with simulated data instead of actual data.

III. EXPERIMENT

The data set is designed so that most of characteristic phenomena that could be present in actual time data are taken into account. The main phenomena are the following: stationary, energy, monotony, periodicity, variation around a central tendency, class presence. The set of potential phenomena is considered through 10 signals, Fig. 2. For reasons of place limit and given that 1) there is a large number C (C>>10) of possible methods for characterizing each signal, 2) we wish to consider individual opinions and 3) we have to compare objective and subjective results, only 4 methods are considered here (see [4][5] for other characterizing methods with either simulated or actual data),

A. "Objective" evaluation

Each of the 10 signals, Fig. 2, is characterized using :

- the arithmetic mean, yielding a data reduction level percentage r_1=1%,

- the RMS value. $r_2= r_1$ and dt_2 is quite identical to dt_1 (with Matlab). The human analysis time being much more complex to be appreciated, it will be discussed in part II,

- the 3 membership values averages (mva) from the space windowing shown Fig.3. $r_3=3^* r_1$ and dt_3 is around 40^*dt_1,

- the 3 membership values averages (mva), from the space windowing shown Fig.3, but with 3 time windows with identical width, $r_3=9^* r_1$ and dt_4 is around 120^*dt_1.

With our 3 stages evaluation procedure, we choice distance as "natural" as possible, i.e. based on the notion of absolute value difference (see [6] for some comments about other choices). For instance, the distance between two signals is:

$$d_0(n,n') = \sum_{k=1}^{100} \left| x_{nk} - x_{n'k} \right| \qquad (1)$$

where x_{nk} is the k[th] value of the signal n. Thus with N=10 signals, Fig. 2, 45 distance values are so computed and placed within the set $\mathbb{D}_0$. The distance between two indicator sets yielded by the characterizing method c is computed in the same way. For instance, the distance with the method 4 is computed as following:

$$d_{14}(n,n') = \frac{1}{9} \sum_{w=1}^{9} \left| mva_{nw} - mva_{n'w} \right| \qquad (2)$$

where w indicate a space-time window. With N=10 signals, 45 distance values are so computed and placed within the set $\mathbb{D}_{14}$.

For the final stage, we suggest, for each characterization method c, to simply build a scatterplot of 45 points where the coordinates correspond to $\mathbb{D}_0$ for the horizontal axis and to $\mathbb{D}_{1c}$ for the vertical axis. A graphic criteria for choosing a method can be that "the points are more or less positioned along an line with a positive slope" (CR_3). More globally, we can characterize the connection between the two sets by the classic linear correlation index (CR_4). The results are shown in Fig. 5. Obviously, given both criteria, the best characterizing method is the fourth one. This result will be discussed in section IV.

B. Subjective evaluation

A questionnaire is used with 24 participants with a background equal or higher than an engineer's, all of them work in the laboratory (LAMIH) and have to deal with data coming from human and/or technologic component systems (electrician, ergonomist, biomechanician…). The questionnaire starts with : “Here are 10 signals supposed to be recorded onto 10 individuals in identical conditions (same task and environment)”. Then the main questions are:

Q1. Among the 10 signals, indicate the two signals A and B which **resemble each other the more**:

A (from 1 to 10) : [ ]

A (from 1 to 10) : [ ]

Q2. Apart from these 2 signals, indicate the two signals C and D which resemble each other the more:

C (from 1 to 10) : [ ]

D (from 1 to 10) : [ ]

Q3. Among the 10 signals, indicate the two signals E and F which **resemble each other the less**:

E (from 1 to 10) : [ ]

F (from 1 to 10) : [ ]

Q1. Apart from these 2 signals, indicate the two signals G and H which resemble each other the less:

G (from 1 to 10) : [ ]

H (from 1 to 10) : [ ]

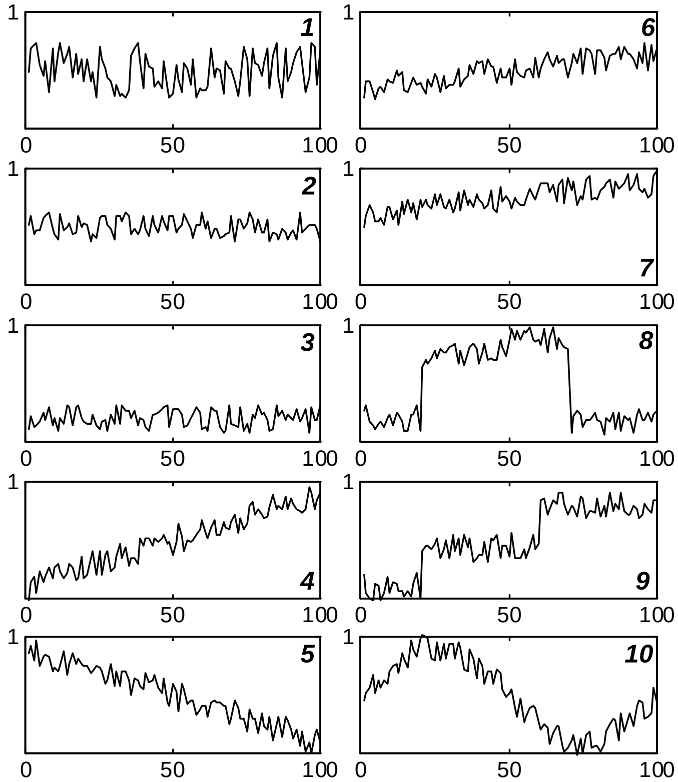

Figure 2. 10 signals (one can imagine that each signal is recorded at fs=1Hz, thus 100 values by signal)

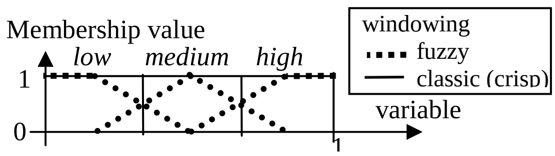

Figure 3. fuzzy space windowing for the characterization

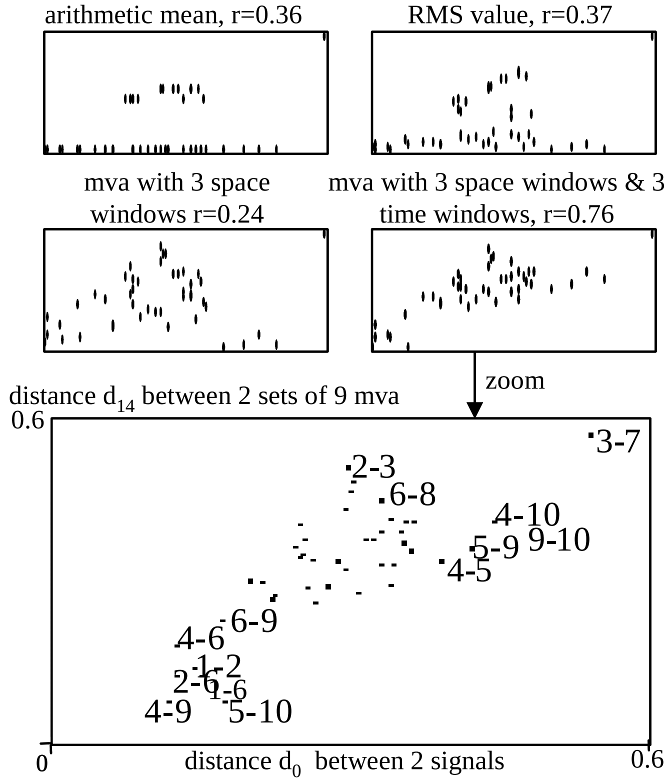

Figure 4. 45 distances between the indicators coming from the data characterization stage vs. 45 distances between the 10 signal for each of the 4 characterizing methods. R is the linear correlation coefficient.

With the subjective data, here again several statistical analysis paths can be designed. Focusing here mainly on the frequencies of the cited pairs, only Fig. 5, left column, will be considered. The set of 4 histograms shows that the cited pairs are less numerous with similarity than with dissimilarity opinions. The most often cited pair with the highest similarity is (A,B)=(1,2), namely two signals with the same continuous component, the dispersion varying a little bit. The most often cited pair with the lowest similarity is (E,F)=(4,6),

C. *Connction beween objective and subjective data*

Fig. 5, right column, puts into relation, for each of the 4 questions, the response frequencies and the distances. Even though the 4 scatterplots show large differences, it seems to have some coherence between objective and subjective data. For instance, the top scatterplot highlight that the pair (1,2), which is the most often cited pair for the question Q1, has a lowest distance d_0.

IV. DISCUSSION

The main aim of this empirical study based on a didactic example was to show that:

1) even though the data set is very simple, and may be simplistic, the characterizing methods are numerous. Furthermore one can imagine that two main cases can be envisaged and always tested:

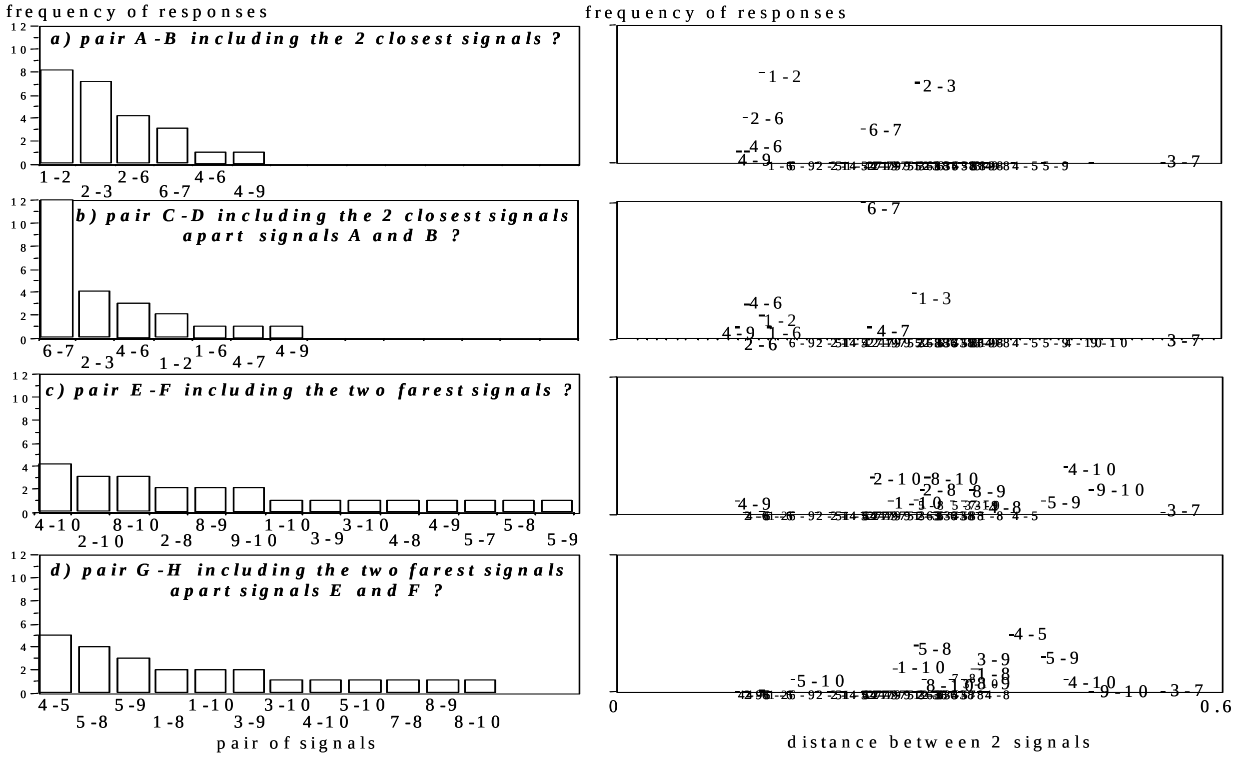

Figure 5. on the left side: frequencies of the cited pairs, for each of the 4 questions ; on the right side: frequencies vs. distances betwwen the 10 signals

* choosing a characterization method without taking into account opinions about the signal resemblance and/or dissemblance. In such a case, there are two main discourses: a) "the time data are characterized using the method CH_1", i.e. without stating the data characterization problem (the most often encountered in the literature) and b) "the data can be characterized using methods CH_1, CH_2,…, CH_C; the criteria for the choice are CR_1, CR_2,…, CR_R; given these ones, the selected method is CH_r.".

It is worth noting that with the first approach, there is not so much questions (in most cases), but with second one, there are often many questions: "Why not methods CH_{C+1}, CH_{C+2},…, criteria CH_{R+1}, CH_{R+2},… ? ". We prefer the second approach, even though many critics could be put forward. Finally, may be that the following discourse could be always present: "Giving all the existing methods, giving all the methods I have heard about, giving the methods I really know, the chosen method is …". Such a discourse should show that a part of subjectivity remains in the choice.

* choosing a characterization method from considering results of an experimental design where expert's opinions about the signal resemblance and/or dissemblance are recorded. One of the main difficulties is in the questionnaire design (e.g. are questions about resemblance, dissemblance, relationships, connections, links ?). Besides the difficulty to apprehend these concepts (mathematically and physically), another difficulty is in the choice of the "stimuli" (a subset from the actual data set) to be presented to the experts. We let the reader imagine an experiment where 10 multidimensional signals with 3 components (only) are presented: if N>>10 and P>>3, how choosing the stimuli ?

2) a space-time windowing gives the best results, which is intuitively obvious.

Finally, giving that 1) 3D measurement systems yield a large data set (N>30 and P>10), 2) this set can present imperfections (with human and/or technical origins), 3) the part of subjectivity can not be totally reduced to zero in the way to characterize data, we suggest to always start the statistical analysis using an exploratory approach based on a space-time windowing. The problem of space-time windowing and the way to investigate membership values are studied in part II.

References

[1] S.M.. Pandit, Time series and system analysis with application, FL: Krieger, Malabar, 1993.

[2] J.D. Jobson, Applied multivariate data analysis. Vol II: Categorical and Multivariate Methods, NY: Springer, 1992.

[3] E.R. Tufte, The visual display of quantitative information, CU: Graphic Press, Cheshire, 1990.

[4] P. Loslever, "Obtaining information from time data statistical analysis in human component system studies" Information Sciences, vol. 132, pp. 133–146, 2001.

[5] P. Loslever, J.C. Popieul, P. Simon "From classic statistical characterization to fuzzy windowing based characterization for the exploratory analysis of miscellaneous time variables: example in the field of car driving studies", Fuzzy Sets & Systems, vol. 137, pp. 271–296, 2003.

[6] P. Loslever, J. Flahaut, F. Lepoutre "Caractérisation statistique des mouvements", Le Travail Humain, vol. 62(2), pp. 131–154, 1999.