Data Normalization in Signal and Pattern Analysis and Recognition: A Modeling Approach

Luciano da Fontoura Costa

Data Normalization in Signal and Pattern Analysis and Recognition: A Modeling Approach

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

25th May 2022

Abstract

Normalization constitutes an important aspect involves in supervised and unsupervised pattern recognition. In this work, we approach this relevant issue from the perspective of preliminary modeling, mainly in terms of respective formulae, the characteristics of the involved features adopted for representation of the entities to be compared and recognized. After presenting several related concepts and methods, including random vectors, densities, and modeling formulae, several normalization approaches are described. Two main methods for comparing the data (signals or sets of features), namely the Pearson correlation coefficient approach and the multiset coincidence similarity, are then presented. The interacting effects between the several described normalizations and comparison approaches, which can critically influence the comparison and classification results, is then addressed and discussed and studied respectively to three case examples involving relatively complex signals/features involving more than one level of detail. The reported concepts, methods and results led to the identification of several important issues, including the intrinsic distinctions between the Pearson correlation and coincidence similarity, with the latter being able to take into account constant or mean portions of the signals. The marked effect that different normalization approaches can have on the comparison was confirmed and discussed. The reported developments motivate an approach involving the optimization of distinct features, possibly in different manners compatible to their mathematical model, wile optimizing some required criterion has also been described.

1 Introduction

Patterns have long been important for humans, as they provide an effective means for representing all types of objects, entities and phenomena in the real-world. Indeed, recurring entities that have some special importance are often organized into *groups* or *categories* that often, but not necessarily, have mutually similar characteristics while differing from other entities. In this case, the groups can be understood as *clusters*. It is interesting to observe that not every cluster has an associated category, and not every category corresponds to a cluster, in the sense of its entities not being well-separated from the remainder entities, as is the case of e.g. two or more adjacent groups.

The assignment of categories to entities constitutes the main objective in *pattern recognition* (e.g. [1, 2, 3, 4, 5, 6, 7]), which can be of two main types: (a) supervised, where preliminary information or prototypes about the categories are available; or (b) unsupervised, when not information is known about the categories. Needless to say, the latter type of recognition is typically more challenging than the former. Observe that in the latter case, the categories are not known *a priori* and need to be inferred from the data. However, though often simpler than the unsupervised counterpart, supervised pattern recognition also represent several challenges implied by each of the involved stages and elements. These difficulties, which can interact one another while influencing the recognition, have diverse origins that include but are by no means limited to: noise and other interferences, undersampling, inconsistent categories, overlap between categories, curse of dimensionality, biased sample, as well as inadequate recognition methods (e.g. [8, 9, 7]). In addition to these issues, unsupervised pattern recognition also involves problems related to the definition of the number of categories and their delimitation.

Given the ubiquitous and critical importance of pattern recognition for most human activities, their automation through artificial means has received growing attention along the last decades. The motivation for automated pattern recognition often relates to enhancing robustness, accuracy, and speed, as well as alleviating humans from

repetitive tasks. The importance of this multidisciplinary area, extending from multivariate statistics (e.g. [10, 11]) to neuronal networks (e.g. [12]), has been corroborated by an impressive number of interesting works in the literature.

Figure 1 illustrates, in simplified manner, the main involved data and stages involving in a typical pattern recognition approach: (a) acquisition of measurements (features) of the entities to be recognized; (b) pre-processing, possibly involving *normalization* and minimization of unwanted data; and (c) the recognition proper; (d) the respective quantification of the performance of the whole system in terms of each of its involved aspects, aiming at respective validation of the approach.

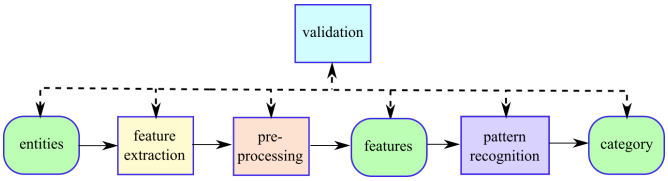

Figure 1: Typical pattern recognition pipeline, represented in simplified manner. *Feature extraction*: A set of features are measured from the entities to be recognized. *Pre-processing*: the features are handled in order to curate their quality, *normalization* purposes, as well as minimization of unwanted aspects. *Pattern recognition*: methods are applied in order to assign categories to the entities. *Validation*: The validation of the approach considering each of the involved data, stages, and results.

Each of the data and tasks in Figure 1 have their own specific effects on the recognition result, but the overall result is often a consequence also of complex interactions between those effects. Is is therefore not surprising that each of these aspects have been addressed in a substantial number of related works in the literature.

The present work focuses on the problem of *data normalization* (e.g. [13, 14, 15, 16, 17, 18, 19]), which constitutes one of the main objectives of the *pre-processing* stage. Oftentimes, once the features have been measured from the samples, they will be characterized by their own physical units, distinct value variations, including or not negative values. Since having features with markedly different ranges can influence (bias) the recognition results, means for making these measurements more commensurated often have to be devised and applied.

For generality's sake, this work henceforth takes into account that one-dimensional signal (and sets of features can be represented and handled in the same manner concerning the concepts and methods presented here (e.g. [20, 21]). In the case of signals, their values are considered as features irrespectively of the original adjacency/topology. At the same time, the values of a feature can be visualized as a function along the horizontal axes, though the order of the abscissae is immaterial and can be taken in any permutation. In other words, the approaches to signal and features normalization treat each signal or feature value independently of their position along the horizontal axis,. By adopting this approach, it becomes possible to address both signal and features normalization in an integrated manner.

We start by briefly revising some important related concepts from multivariate statistics (e.g. [10, 11]) and proceed by describing the suggested modeling of the values of the signals or sets of features in terms of respective basic mathematical formulae. The several normalization methods considered in this work, which include the standardization and minmax approaches, are subsequently presented. Then, the two main comparison approaches addressed here, namely the Pearson correlation coefficient and the multiset coincidence similarity, are described. The remainder of the work studies the effects of the several normalization and comparison approaches, with emphasis on their respective combinations, on the comparison results in terms of three case examples involving relatively complex signals/features presenting more than one level of detail, or scale. Several interesting results, as well as a respectively motivated basic optimization approach involving heterogeneous normalization of the features are also presented and discussed.

2 Random Variables, Densities, and Transformations

Science and technology are amply underlain by modeling approaches, which can be developed from a probabilistic point of view.

Any measurement, no mattering its level of randomness, can be conceptualized and modeled in terms of a respective *random variable* X. Random variables, which are intrinsically associated to random experiments, can be sampled in terms of N respective values. In case more than one measurement is taken per sample, they are often organized as a *random vector* $\vec{X}$.

Random variables and random vectors can be fully characterized, from the probabilistic point of view, in terms of the respective *probability density functions*, *probability densities*, or even simply *density* for short.

In the case of a random variable X, we would have $p(X)$; while $p(\vec{X})$ would apply for a random vector. Henceforth, the set of all values of X or $\vec{X}$ for which a probability density is assigned will be understood as the respective *support* of that density. To be statistically well-posed, a density needs to have: (a) all its values being non-negative; (b) the integral of the density along the

support needs to be identical to one.

Uniform probability density functions are characterized as:

$$p(s) = c, \quad or \tag{1}$$
$$p(\vec{x}) = \vec{c} \tag{2}$$

The standard deviation σ_X of a random variable X is equal to the positive root of the respective variance.

The $1-$dimensional and $N-$dimensional normal densities can be expressed as:

$$g(x, \mu, \sigma) = \frac{1}{2\pi\sqrt{\sigma}} e^{-0.5 \left(\frac{(x-\mu_X)}{\sigma_X}\right)^2} \tag{3}$$
$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^M |K|}} e^{-0.5[\vec{x}-\vec{\mu}]^T K^{-1} [\vec{x}-\vec{\mu}]} \tag{4}$$

where $\vec{\mu}$ and K are the average vector and covariance matrix of $\vec{X}$, and $|K|$ is the determinant of the latter matrix.

Given the standard deviation of a random variable X, which corresponds to the positive square root of the respective variance, the *coefficient of variation* (also *relative standard deviation*) of X is defined as:

$$cv(X) = \frac{\sigma_X}{\mu_X} \tag{5}$$

Though the mean and standard deviation of a random variable are, in principle, independent, relationships between these two statistical measurements can be found relatively often in practice, being of special interest while normalizing data. For instance, in the *binomial*, *log-normal*, and *exponential* densities, the mean is proportional to the standard deviation, implying in constant coefficient of variation. In these cases, taking the mean and standard deviation of a measurement as features inherently implies in redundancy. Poisson densities, the mean is proportional to the variance.

Given $1D$ density $p(x)$, $x_m \leq X \leq x_M$, its respective *cumulative distribution funtion*, *cumulative distribution*, or simply *distribution*, can be defined as:

$$P(x) = \int_{-\infty}^{x} p(x)dx = \int_{x_m}^{x} p(x)dx \tag{6}$$

Random variables not only have diverse units and choices, but can also be *statistically transformed*, or *data transformed* in a virtually infinite number of ways. For instance, given a random variable X, it is possible to consider respectively *linear transformations* of the type:

$$\tilde{X} = aX + b \tag{7}$$

These transformations typically change the respective density, which can be found by several methods, including those based on the respective density or distribution functions. In the present work, we focus our attention on the *Jacobian method* summarized in the following.

Let $p_X(\vec{x})$ be a multivariate (N dimensions) density on the random vector $\vec{x}$. Let also $q(\vec{x})$ be a function on $\vec{x}$ used to transform that original random vector into a new random vector $\vec{y}$. The *Jacobian* of this transformation can be placed as:

$$J = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_M} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_M} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial y_M}{\partial x_1} & \frac{\partial y_M}{\partial x_2} & \cdots & \frac{\partial y_M}{\partial x_M} \end{bmatrix} \tag{8}$$

As an example, in case the above Jacobian exists, the density $p_Y(\vec{y}$ corresponding to the new random vector $\vec{y}$ can then be expressed as:

$$p_Y(\vec{y}) = \frac{p_X(\vec{x})}{|J|} \tag{9}$$

In the case of the above linear transformation, in case X were originally described by $p(x)$, we would have:

$$J = a \tag{10}$$

Hence:

$$p_Y(y) = \frac{p_X(x)}{a} \tag{11}$$

Considering a random vectors $\vec{x}$ and respective density $p(\vec{x}$ undergoing the same linear transformation, it would follow that:

$$J = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & a \end{bmatrix} \tag{12}$$

from which:

$$p_Y(\vec{y}) = \frac{p_X(\vec{x})}{a^N} \tag{13}$$

When applied to features, which are random variables, statistical transformations yield new features that, though somewhat related to the original ones, typically possess distinct statistical properties. Therefore, features transformations can actually though of as constituting an approach to obtain new, or alternative features in a pattern recognition application.

A particularly important type of statistical transformation consists in the respective *standardization*. More specifically, given a set of N samples observed from a random variable X with average μ_X and standard deviation σ_X, these values can be normalized by, for each sample x_i, subtracting the average and then dividing by the standard deviation, i.e.:

$$\tilde{x_i} = \frac{x - \mu_X}{\sigma_X} \tag{14}$$

In addition to becoming dimensionless, the new values $\tilde{X}_i$ have null mean and unit standard deviation, and most of then result comprised in the interval $[-2, 2]$.

Observe that the standardization corresponds to a linear transformation of the original random variable, being expressable through Equation 7 with $a = 1/\sigma_X$ and $b = -\mu/\sigma_X$.

3 Modeling Formulae

Real world and analytical signals have specific intrinsic characteristics that define them. Even when the original formulae of these signals are not available, which is often the case, they can still be hypothesized in terms several mathematical manners, e.g. by using polynomials or Fourier series, e.g.:

$$f(x) = a_2 x^2 + a_1 x + a_0 \tag{15}$$

$$g(x) = b_3 \cos(3x) + b_2 \cos(2x) + b_1 \cos(x) + b_0 \tag{16}$$

where $a_2, a_1, a_0, b_3, b_2, b_1$ and b_0 are generic real values.

Even when the formula is available, it is often interesting to decompose it in terms of some expansion such as the Fourier series in case one is interested in investigating the effect of each of respective terms on the recognition results.

The *mean value* of a function $f(x)$, which appears as a term in several expansions including Fourier series, can be defined as:

$$\langle f \rangle = \frac{1}{x_M - x_m} \int_{x_m}^{x_M} f(x) dx \tag{17}$$

It is important to bear in mind that the mean of a function or signal does not necessarily coincides with its constant term, which is the case in Equation 16, where the constant term b_0 is also the average of the overall signal. This is so because all other terms have null mean values. However, this is not so in the case of Equation 16, where the constant term a_0 does not correspond to the overall average as a consequence of the other two terms not having null mean.

It follows from the above reasoning that, given a generic signal representation in the form:

$$f(x) = a_E f_E(x) + \ldots + a_1 f_1(x) + a_0 \tag{18}$$

the respective mean value will correspond to:

$$\langle f \rangle = a_E \langle f_E(x) \rangle + \ldots + a_1 \langle f_1(x) \rangle + a_0 \tag{19}$$

So, a_0 will correspond to the overall mean of the signal if and only the sum of the means of all other terms results zero.

The above approach to composed signals immediately extends to noise, for instance:

$$h(x) = a_2 \; u(x) + a_1 \; g(x, \mu, \sigma) + a_0 \tag{20}$$

where $u(x)$ and $g(x, \mu, \sigma)$ are uniform and normal noise distributed along the respective support.

When all (or a subset) of the available features have the same nature and units, such as the pixels of images, it is also possible to consider normalizing the whole set (or subset) of uniform features along not only the respective samples, but also among all those features.

Another situation deserving particular attention is when *outliers* (e.g. [22, 23, 24]) are present in the data. Basically and informally speaking, an outlier is a pattern which has features too distinct from all others. One of the main implied problems concerns the fact that the outlier samples can strongly influence the normalization, imposing a substantial bias on the values of that features for all other samples. One possibility to deal with outliers is to identify and remove them from the dataset before normalization, or treat them separately and them compare with the results obtained for the other samples.

As it will be discussed in the present work, the formula of the available signals or features to be recognized plays a critical role respectively to the recognition results. As we will see in this work, even the simple constant and average terms can have dramatic effects Indeed on the recognition results. The signals and features formulae can be understood as *mathematical models* of the each signal or features to be characterized, and then classified.

Though the form of the involved signals/features is rarely known for certain, it is still interesting to hypothesize them. As an example, let us suppose that we have a metal plate containing three points of interest. We may consider each three points on the plate as a category, and then take N samples of a measurement of interest (e.g. temperature) along time for each of those points. Though these samples could be understood as being related to a single same feature (temperature), it is also possible to understand each of the samples for each point as an individual feature with the same nature and unit. In this case we would have 3 categories, each one represented by N respective features, as illustrated in Figure 2. How could these features be normalized?

Another possibility would be to have 3 features taken from 20 distinct points (samples) on 5 distinct plates (entities, possible organized in categories). In case we group the features subsequently, we would have three groups of 100 elements, which could be represented by the same type of diagram as in Figure 2. How could these features be normalized?

Observe that the former of the above approaches can be understood as the analysis of three *signals*, each repre-

sented by 100 samples that are here treated as individual features. Given this representation, we may be interested in comparing or classifying these signals. The latter approach can be understood as a unsupervised *pattern recognition problem* involving 5 entities, each represented by 20 samples with 3 features each.

Observe that another approach to address the former situation would be to consider *stochastic processes* (e.g. [25, 26, 27]) taking place along time, but this alternative is not considered in the present work for simplicity's sake.

The approach presented in this work applies to both cases, though with quite distinct interpretations and treatments while of the respective analysis and/or recognition.

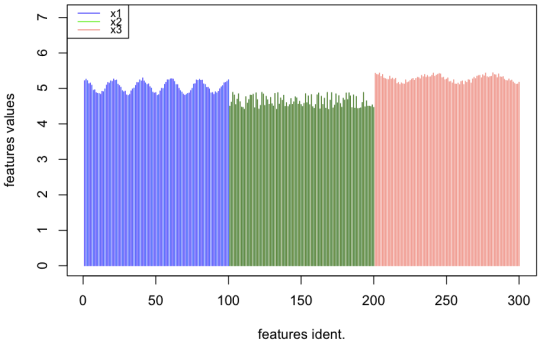

Figure 2: Three categories, each of which represented in terms of 100 features of the same nature and unit (e.g. temperature). In this case, the colors indicate the respective categories. How can we normalize this feature? Alternatively, we could understand this diagram as corresponding to 3 features, each represented by 20 samples taken from 5 different plates. In this case, the colors represent each of the three features. How could we normalized these features?

Figure 3 illustrates another hypothetical possible result, characterized by much more substantial variations of the measurements along the observations, possibly as a consequence of overall heating reflects on features 1 and 3, while point 2 would be subjected to other environmental conditions (e.g. receiving some oscillatory air convection).

These two examples will be further considered along this work in order to illustrate the effect of normalization choices on the recognition results.

It could be hypothesize that the temperatures above follow the following mathematical formula:

$$x(t) = v(t) + n(t) + \ell(t) + a \tag{21}$$

where $a_3 v(t)$ is variation term such as a sinusoidal or a power of t (as in a polynomial), $n(t)$ is a noise term, $\ell(t)$ is a linear variation term, and a is a constant term. Observe that this formula is specific to our examples. Many

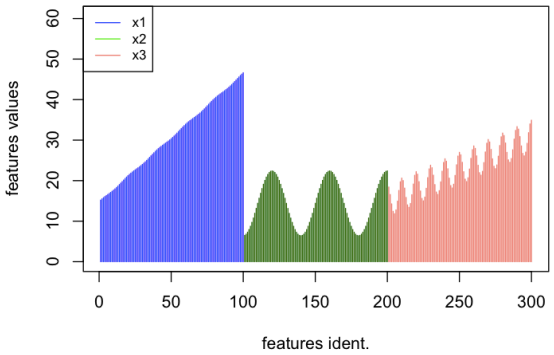

Figure 3: Another hypothetical set of the three features, characterized by more substantial variations along the observations. Groups 1 and 2 present a progressive increase that could be related to overall temperature variation. Observe also the tiny variations along the slanted profile of feature 1.

other formulae can be adopted concerning other datasets or research questions.

Given that the signals in Figure 2 and 3 were synthesized, their respective formulae are known. In the former case, they are:

$$x_1 = 0.2\ \cos(t) + 0.1\ u(t) + 5 \tag{22}$$
$$x_2 = 0.5\ u(t) + 4.4 \tag{23}$$
$$x_3 = 0.1\ \cos(0.5t) + 0.15\ u(t) + 5.2 \tag{24}$$

where $u(t)$ corresponds to uniformly random noise within the interval $[0, 1]$ and, in the second case example:

$$x_1 = 0.2\ \cos(t) + t + 15 \tag{25}$$
$$x_2 = 8\cos(0.5t) + 14.4 \tag{26}$$
$$x_3 = 4\cos(2t) + 0.5t + 15.2 \tag{27}$$

Observe that, for simplicity's sake, we have avoided having formulae with more than three terms. More specifically, the first case does not involve linear variations, and the latter does not incorporate noise. So, the formula to be considered henceforth in the subsequent examples can be summarized as:

$$x(t) = v(t) + [n(t)\ \ or\ \ell(t)] + a \tag{28}$$

with the term within brackets being referred to as the *'linear'* term. More generally, formulae with less or more terms of different types will apply to signals and features. Of particular relevance is the consideration of eventual mutual interrelationships between the involved features, which can be approached in terms of combined multidimensional formulae.

The activity of trying to develop a model respectively to each considered feature is important not only for its potential value while defining the normalization, comparison

and recognition approaches to be adopted, but also provides a motivation for considering in a more careful, comprehensive and systematic manner the nature and properties of each of the features, leading to an enhanced overall understanding of each specific problem and dataset.

4 Other Normalization Approaches

Having discussed the modeling of signals and features in terms of respective putative formulae, we now proceed to discussing the other normalization approaches to be considered in the present work. We will present these normalizations respectively to the formulae in Equation 28.

Given a non-null signal or features set x, it can be *minusmin* normalized by subtracting its minimum value, i.e,:

$$\tilde{x_i} = x_i - \min\{x\} \tag{29}$$

with $i = 1, 2, \ldots, N$. The new variable $\tilde{x}$ therefore will have its minimum value equal to 0, while nothing else can be said about its other properties.

The *minmax* normalization of that same original variable is implemented as:

$$\tilde{x_i} = \frac{x_i - \min\{x\}}{\max\{x\} - \min\{x\}} \tag{30}$$

Now, we have that $0 \leq \tilde{x} \leq 1$.

Another interesting possibility consists of transforming the original signal or feature set into a respective probability density function, which can be obtained as:

$$\tilde{x_i} = \frac{x_i - \min\{x\}}{\sum_{i=1}^{M}(x_i - \min\{x\})} \tag{31}$$

The new random variable $\tilde{x}$ will have minimum value equal to zero and area equal to one.

Provided we have the signal or features set represented in terms of the formula in Equation 28, we can perform normalizations respective to each of the involved terms.

For instance, the constant term can be removed as:

$$\tilde{x_i} = x_i - a \tag{32}$$

which will be henceforth referred to as the *minusconst* normalization.

Similarly, it is possible to remove the linear term:

$$\tilde{x_i} = x_i - \ell(t) \tag{33}$$

leading to the *minuslin* normalization.

One potential problem when using extrema (minimum or maximum) values in normalizations is that the results can become strongly affected by outliers. An interesting alternative that reduces this potential effect consists of dividing the feature values by their respective mean, i.e.:

$$\tilde{x_i} = \frac{x_i}{\langle x \rangle} \tag{34}$$

This is scheme, which will be henceforth called *mean*, yields results directly proportional to those obtained by the *density* normalization approach.

It is also possible to consider only the constant terms, in the so-called *constant* normalization, implying:

$$\tilde{x_i} = a \tag{35}$$

or only the linear (or noise) term, yielding the *linear* normalization:

$$\tilde{x_i} = \ell(t) \ \ or \ n(t) \tag{36}$$

It is also possible to consider only the varying terms, but this will not be considered here for simplicity's sake.

Another interesting possibility regarding the normalization of a signal or set of features consists in combining two or more normalization schemes. For instance, it is possible to apply standardization after any of the above normalization possibilities.

Comparisons are the basic component involved in most pattern recognition approaches, as they are required both to estimate relationships between the entities and to provide subsidies for deciding on the respective separation (e.g. [8]).

Having discussed several normalization possibilities, it is now time to proceed to addressing the two main types of *comparisons* between signals or feature sets adopted in this work, namely by Pearson correlation (next section) and coincidence similarity (Section 6).

5 The Pearson Correlation Coefficient

Given two random variables X and Y represented by N paired samples x_i, y_i, their *covariance* can be estimated as follows:

$$cov(x, y) = \frac{1}{N-1}\sum_{i=1}^{N}[x_i - \mu_X][y_i - \mu_Y] \tag{37}$$

The *Pearson correlation coefficient* between these two variables can be expressed as:

$$ccoef(x, y) = \frac{1}{N-1}\sum_{i=1}^{N}\frac{[x_i - \mu_X]}{\sigma_X}\frac{[y_i - \mu_Y]}{\sigma_Y} \tag{38}$$

In case the two variables X and Y are presented being already standardized, their Pearson correlation can

be more directly estimated as:

$$ccoef(x,y) = \frac{1}{N-1}\sum_{i=1}^{N} \tilde{x}_i\ \tilde{y}_i = \frac{\vec{\tilde{x}}^Y \vec{\tilde{y}}}{N-1} \qquad (39)$$

We have that $-1 \leq ccoef(x,y) \leq 1$, with the value 1 meaning maximum joint variation and -1 corresponding to the maximum opposite variation.

A critically important characteristic of the Pearson correlation coefficient that is not often realized corresponds to the fact that it *intrisic* and *unavoidably* implements the standardization of both random variables (Equation 38). As a direct consequence, it makes no difference whatsoever to apply or not to have the variables standardized. In addition, any preliminary normalization that leaves a constant term will be modified in the sense that this term will not be taken into account by the Pearson correlation analysis. The same applies to transformations that multiply the features by a constant factor.

As with every comparison approach, the Pearson correlation coefficient has relatively advantages and shortcomings. Its main advantage consists of being intrinsically linked to the concept of *joint variation* between a pair of random variables, being particularly effective and suited to that finality. On the other hand, the fact that this coefficient removes the average level from features and signals may be suitable or unsuitable depending on each specific case. In particular, situations where the average level is important and needs to be taken into account may not be effectively treated by using the Pearson correlation approach. In addition, this method can also amplify unwanted or irrelevant tiny variations along the signal as a consequence of its normalization of the signal magnitude after mean removal, as illustrated in Figure 4.

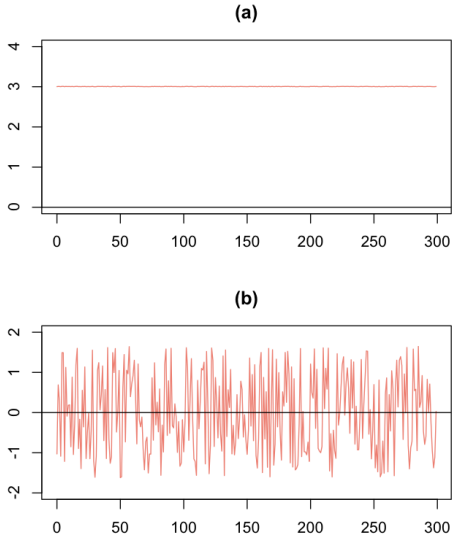

Figure 4: As a consequence of the standardization intrinsically implemented by the Pearson correlation coefficient, which removes the mean level and normalize the values magnitude dispersion to standard deviation of one, tiny oscillations (a) may get substantially amplified (b). This could be wanted or not depending on the importance of the osciallations for each specific analysis or recognition.

Observe that the small oscillations can be or not important for a specific analysis.

Another aspect of the Pearson approach that demands special attention is when the features or signals can be divided into two parts: one with well-defined joint variation, and another involving variation of only one of the variables while the other remains at particularly low values. Figure 5 illustrates this effect respectively to two signals or sets of features (a) and (b), with the former being composed by two peaks corresponding to normalized gaussians, therefore having unit area. The comparison value obtained by using the coincidence similarity (to be described in the next) section, also shown in (c), provides a more effective quantification of the relationship between these two signals, be it regarding joint variation or shared graph areas.

The consequence of this normalization is a Pearson correlation coefficient of 0.84, which is possibly too high given that half of the area of the signals is unrelated. The coincidence similarity also shown in (c) provides a more compatible quantification of the interrelationship between these two signals regarding their joint variation and shared areas below the graphs of the signals.

Yet another situation that has to be carefully considering when using standardization and Pearson correlation concerns the present of possible outliers, as illustrated in Figure 6.

Observe the significant amplification of the value of the outlier in (c), implying a more substantial reduction of the value obtained for the Pearson correlation coefficient than for the coincidence similarity. Not that, except for the single outlier point, the two signals would be identical. These situations can be dealt with by removing the outliers prior to the respective normalization and comparison of the signals and/or features. At the same time, the coincidence comparison will lead to a markedly accurate result even without outlier removal in the case of the above example.

6 Multiset Coincidence Similarity

Similarity indices have been extensively employed in several areas as means for comparing sets and quantities (e.g. [28, 29, 30, 31, 20, 32, 33, 17]). In particular, the Jaccard similarity index has been systematically used since

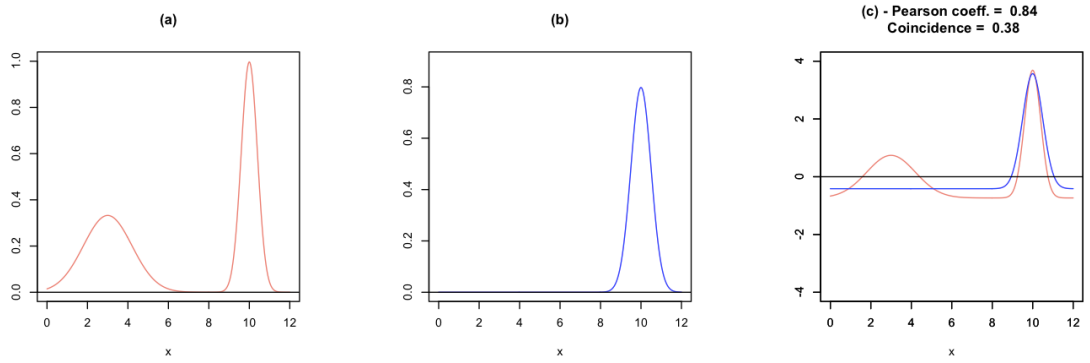

Figure 5: The values of two signals (a) and (b), and respective standardization (c). Observe that this particular normalization scheme aligns almost completely the two highest peaks, while the smaller peak in (a) becomes paired with a constant small value counterpart in (b). The result of the respective standardizations are shown jointly in (c).

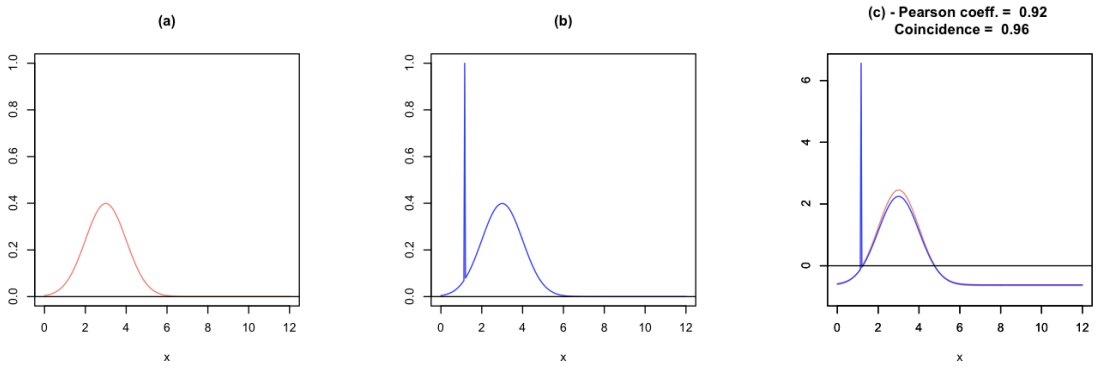

Figure 6: The effect of outliers on the respective standardization and Pearson correlation coefficient respectively to a signal (a) and this signal in presence of a single outlier (b). The respective standardization results are shown in (c).

its description by Paul Jaccard in 1901 [34, 35]. Basically, it provides an effective means for quantifying the similarity between two sets A and B as stated in Equation 40. It can be readily verified that $0 \leq \mathcal{J}(A,B) \leq 1$. However, as shown in [20], this interesting index does not take into account the relative interiority of the two sets being compared. This property can be gauged by another index given in Equation 41, known as overlap (eg. [28]), but here referred to as *interiority*. The *coincidence similarity* index has been introduced [20, 21] as a means to combine the comparisons implemented by the Jaccard and interiority index, so that a more strict — and therefore *selective* – measurement could be obtained. More specifically, as indicated in Equation 42, the coincidence similarity corresponds to the product between the Jaccard and interiority indices.

$$\mathcal{J}(A,B) = \frac{A \cap B}{A \cup B} \tag{40}$$

$$\mathcal{I}(A,B) = \frac{A \cap B}{\min\{|A|,|B|\}} \tag{41}$$

$$\mathcal{C}(A,B) = \mathcal{J}(A,B)\ \mathcal{I}(A,B) \tag{42}$$

with $0 \leq \mathcal{I}(A,B), \mathcal{C}(A,B) \leq 1$.

In multiset theory (e.g. [36, 37, 38, 39, 40, 41]), the union and intersection of two multisets $\tilde{X}$ and $\tilde{Y}$ corresponds to the maximum and minimum between the respective non-negative multiplicities $\tilde{x}$ and $\tilde{y}$, i.e. the number of times that each element appears in each multiset. This allows the previous indices to be expressed as:

$$\mathcal{J}(\tilde{X},\tilde{Y}) = \frac{\min\{\tilde{x},\tilde{y}\}}{\max\{\tilde{x},\tilde{y}\}} \tag{43}$$

$$\mathcal{I}(\tilde{X},\tilde{Y}) = \frac{\min\{|\tilde{y}|,|\tilde{y}|\}}{\min\left\{\int_S |\tilde{x}|d\tilde{x}, \int_S |\tilde{y}|d\tilde{y}\right\}} \tag{44}$$

$$\mathcal{C}(\tilde{X},\tilde{Y}) = \mathcal{J}(\tilde{x},\tilde{y})\ \mathcal{I}(\tilde{x},\tilde{y}) \tag{45}$$

with $0 \leq \mathcal{J}(\tilde{X},\tilde{Y}), \mathcal{I}(\tilde{X},\tilde{Y}), \mathcal{C}(\tilde{X},\tilde{Y}) \leq 1$.

Though the above equations adopt the function representation of multisets [20, 21], for simplicity's sake, it can be readily adapted to vectors, matrices, graphs, etc.

More recently [20, 21, 42, 43], the above indices were verified to be expressable in terms of multiset operations

adapted to take into account negative multiplicities:

$$\mathcal{J}(X,Y) = \frac{s_x\ s_y\ \min\{|x|,|y|\}}{\max\{|x|,|y|\}} \tag{46}$$

$$\mathcal{I}(X,Y) = \frac{\min\{|x|,|y|\}}{\min\left\{\int_S |x|dx, \int_S |y|dy\right\}} \tag{47}$$

$$\mathcal{C}(X,Y) = \mathcal{J}(x,y)\ \mathcal{I}(x,y) \tag{48}$$

with $0 \leq \mathcal{I}(x,y), \leq 1$ and $-1 \leq \mathcal{J}(x,y), \mathcal{C}(x,y) \leq 1$.

The index in Equation 46 had been described previously in the context of analogy to $L1$ norm in [31], being also related to another index with similar motivation [44].

Interestingly, the coincidence index has been found to present some particularly interesting characteristics including enhanced selectivity and sensitivity while comparing similar patterns, as well as robustness to localized perturbation of the features being compared [43]. Allied to other desirable properties, these have paved the way to a number of successful applications to several problems in diverse areas, including template matching (e.g. [21]) and translation of datasets into respective networks [42].

As with the Pearson correlation coefficient, the coincidence similarity presents features that can be wanted or not depending on each specific problem. One of the potential advantages of the coincidence approach is that it does not inherently remove the mean level of features or signals, allowing this potentially important information to be considered in the comparison. If required, the mean level can be removed during normalization.

The preservation of the mean level paves the way to considering many alternative normalization schemes that preserve this information, as well as regarding the choice of methods to be subsequently applied for respective analysis and recognition.

One aspect of the coincidence similarity that should receive special attention regards the fact that the relative mean levels can impact the gauged similarity, as illustrated in Figure 5. For instance, consider the two following signals:

$$x(t) = \cos(t) + a \tag{49}$$

$$y(t) = \cos(2t) + b \tag{50}$$

with $x(t), y(t) \geq 0$ and $a > b + 1$.

The respective real-valued Jaccard similarity, which is one of the terms in the coincidence similarity, can be written as (e.g. [20, 21]):

$$mathcal J(x,y) = \frac{\min\{x,y\}}{\max\{x,y\}} = \frac{\cos(2t)+b}{\cos(t)+a} \tag{51}$$

It follows that distinct ratios a/b will imply in different Jaccard similarity values. In particular, we will have similarity value equal to one whenever $a = b$, and smaller values otherwise. However, this aspect of the coincidence similarity could actually be understood as being useful in case the mean levels are important for the analysis or recognition.

It should also be kept in mind that, even though the coincidence focuses on relationships between the shared areas of the signals graphs, being not intrinsically related for quantifying the joint variation between two signals or feature values, its version catering for possibly negative values (Eq. 46) can also provide indication about the joint variation in a similar, but at the same time distinct, way to that provided by the Pearson correlation coefficient.

7 Normalization and Comparison Interrelationships

We have so far addressed several types of normalization and two main approaches to comparing the random variables, allowing several respective combinations. Interesting and importantly, each of these combinations can lead to distinct comparison and classification results, so that it is useful to better understand these several possibilities.

Figure 7 illustrates the possible combinations of all the above presented normalization schemes followed by respective correlation or coincidence analyses.

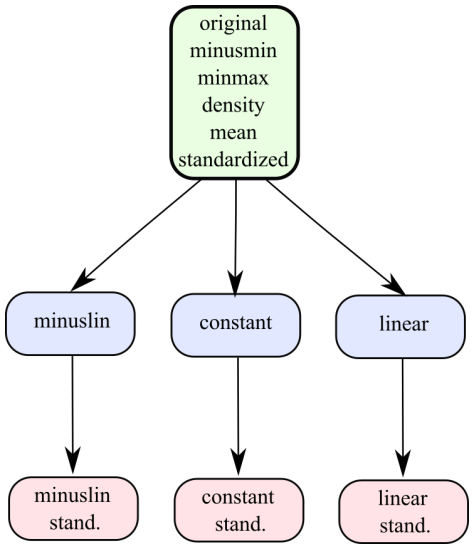

Figure 7: The combinations of the several described normalization schemes, followed by standardization, to be followed by Pearson correlation analysis.

Observe that all normalizations in the green box result identical as a consequence of the Pearson correlation coefficient implementing the standardization of both involved variables. As a consequence, we end up with only the four approaches illustrated in Figure 7.

The relationships between normalizations and coincidence similarity comparison are presented in Figure 8.

Two main groups of interrelationships can be observed. One of them involve the *minusmin*, *minmax*, *density* and *minusconst* normalizations, all of which can result distinct in the case of subsequent coincidence analysis. That is so because that comparison approach, in its non-negative form, can take into account non-null constant and mean terms in the respective signals or features sets. However, the results of all these normalizations will become equal one another and also to the standardization of the original data after they are respectively standardized. The other main group in Figure 8 is analogous to that discussed in the case of Pearson correlation analysis.

All in all, we have that the coincidence approach allows the effective consideration of many more normalizations than the Pearson correlation counterpart.

8 Case Example 1: Nearly constant averages

We are now in a better position to approach the comparison of the signals in our first example in Section 3.

Figure 9 illustrates the original groups (which may correspond to categories or features depending on how a problem is approached) and their respective normalization by employing the considered approaches described previously.

The results of the comparison between the groups of values by using the Pearson correlation coefficient are presented in Figure 10. Recall that the features now incorporate slant variations instead of noise terms.

As a consequence of this comparison approach incorporating standardization of both variables, the six first networks, as well as the *mean* case, resulted identical. Though smaller correlation values have been obtained in the cases *minuslin* and *minuslin* standardized, the relative interactions are similar to those obtained for the previous cases. The results of the last *const*, *linear* and *linear standardized* cases indicate almost no interrelationships between the three groups. All in all, the Pearson comparison of the three patterns suggest almost no interrelationship between them even though the have markedly distinct constant and/or mean terms, which is reasonable since the Pearson approach focuses on joint *variations* between the features or signal values.

Figure 11 depicts the interactions between the three groups in our first case examples as quantified by the coincidence similarity.

Remarkably, each of the networks in Figure 11 resulted mutually distinct. That is a direct consequence of the coincidence similarity being able to take into account the constant and/or mean terms in the original features or signals. The original, as well as the normalizations *density*, *constant* and *mean* led to similar results indicating that the three groups are mostly similar one another as a consequence of their comparable mean and/or constant terms. The approaches *minusmin* and *minmax* also suggest that the three groups are similar, though in a less intense degree. The results obtained for *minusconst* and *minuslin* are more asymmetric and indicate a stronger similarity between groups 2 and 3, and 1 and 3, respectively. Little similarity has been quantified in the other cases in Figure 11.

The results suggesting distinctions between the groups are mostly a consequence of the amplification of the small scale respective variations as implemented by the *minconst* and *minulsin* standardization.

It is important to keep in mind that neither of these results can be deemed to be absolutely better or correct, because each of the obtained comparisons take into account specific hypothesis about the features and their characteristics. These results always need to be further evaluated from the perspective of each specific problem and dataset. In addition, the trends observed respectively to our first example are specific to this dataset and cannot be generalized respectively to other datasets. For instance, much different results could have been obtained in case the signal oscillations were larger, or the constant and/or mean terms were smaller in absolute or relative terms.

9 Case Example 2: Diverse patterns

In order to complement our investigation of the possible effects of combinations of several normalization and comparing approaches, we now consider our second case example, corresponding to the second example in Section 3.

Figure 12 presents the original dataset, involving 100 samples of three groups, as well as its several normalizations.

The results of the Pearson correlation analysis of the datasets in Figure 12 are shown as networks (or graphs) in Figure 13.

Similarly to what happened when of applying the Pearson correlation comparison, most of the results suggest

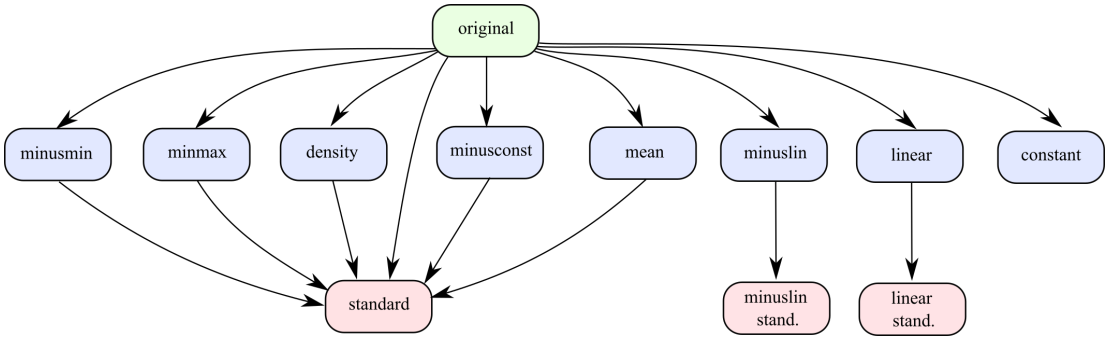

Figure 8: The relationship between the several considered normalizations respectively to coincidence similarity comparision.

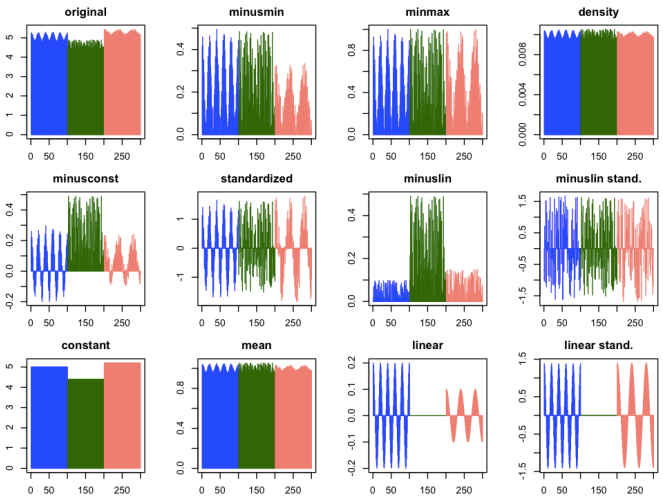

Figure 9: The original groups of our first example and the results of its normalization by applying the several described approaches. Surprisingly distinct results can be readily observed, corroborating the potentially critical influence of normalization on patterns comparison and classification.

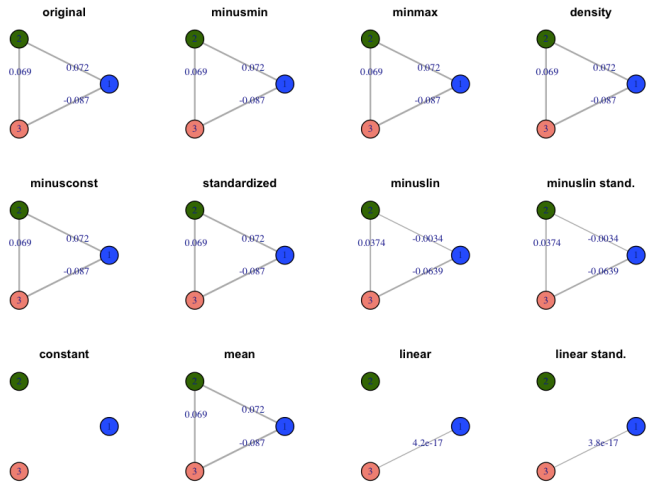

Figure 10: The interrelationships identified by application of the Pearson correlation analysis to several normalizations of the first dataset. All obtained networks suggest that the groups are weakly interrelated or similar.

lack of interrelationship or similarity between the three groups, other than between the pair 1 and 3. Again, this is a consequence of the Pearson approach eliminating the constant and/or mean terms in the features.

The networks resulting from the comparison of the three groups in our second dataset are presented in Figure 14.

As could be expected by now, the coincidence similarity analysis led to distinct results respectively to each of the considered normalization approaches, reflecting the specific hypothesis and characteristics in each case. Interestingly, strong indication about mutual symmetry of similarities between the three groups in this dataset have been obtained only respectively to the *minuslin*, *minuslin* standardized, and *constant* normalizations, with the former two cases involving small comparison values. The result obtained by the latter case is consistent with the substantial similar values of the constant/mean terms of the values in the three considered categories.

Most of the other results indicate a stronger similarity between groups 1 and 3 in this dataset, though with varying values being obtained concerning the interrelationships involving group 2. For instance, the *density* normalization case suggests a stronger similarity between that group and the others, which is compatible with the distributions in Figure 14. All cases involving subsequent standardization yielded almost negligible coincidence values, which is expected given the almost orthogonal nature of the small scale oscillations (sinusoidals with multiple frequencies). Indeed, as a more careful study of the results in this figure will indicate, the coincidence similarity comparison results tend to reflect in a mostly objective and accurate manner the interrelationships between the three categories respectively to the considered feature normalizations.

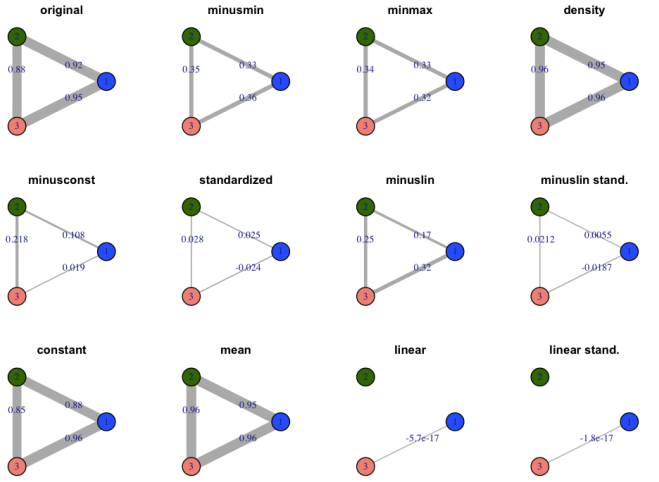

Figure 11: The interrelationships identified by application of the coincidence similarity analysis to several normalizations of the first dataset. Quite diverse results have now been obtained, each of them providing a distinct indication about the interrelationships between the three groups according to specific hypothesis and features characteristics.

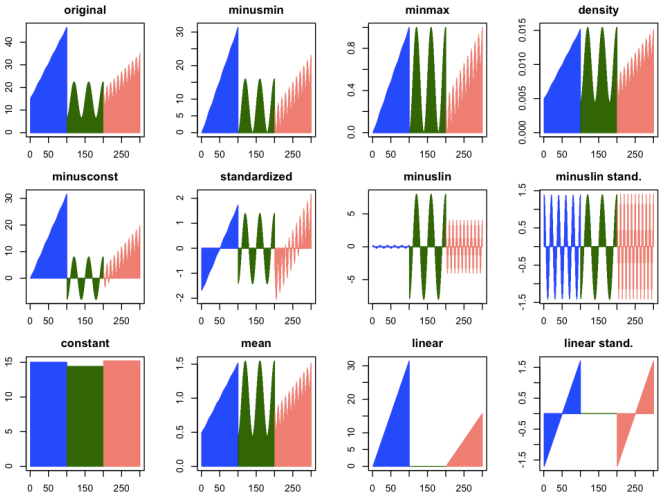

Figure 12: The original set of features of our second case example and the results of its normalization by applying the several described approaches. Even more distinct results have been again observed, corroborating the potentially critical influence of normalization on patterns comparison and classification.

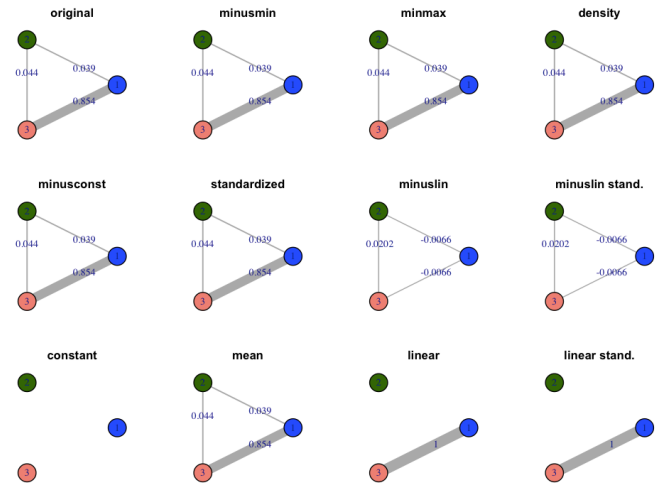

Figure 13: The interrelationships identified by application of the Pearson correlation analysis to several normalizations of the second case example. Except for the normalization approaches *minuslin* and *minuslin* standardized, the obtained networks otherwise suggest that the signals are weakly interrelated or similar.

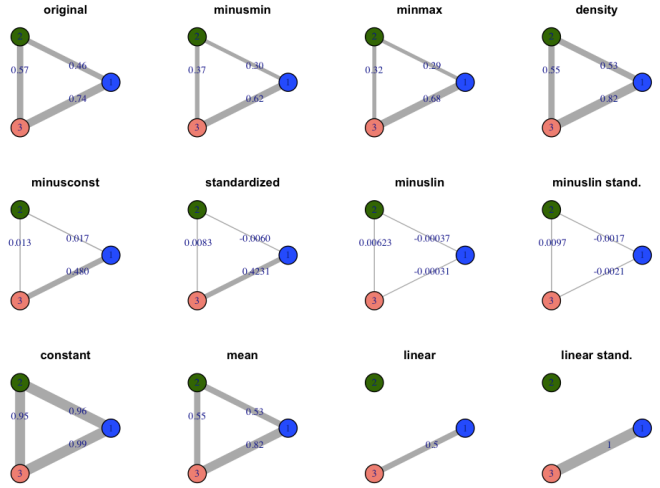

Figure 14: The interrelationships identified by application of the coincidence similarity analysis to several normalizations of the second dataset. Markedly diverse results have again been obtained, each of them providing a different indication about the interrelationships between the three categories according to specific hypothesis and data characteristics.

10 Case Example 3: Real-World Data

In order to complement our analysis of normalization effects, we now approach a real-world dataset consisting of 3 types of handwritten characters ('c', 'e', and 'o'), each represented by 50 respective samples and 4 features corresponding to geometric properties of the characters [42]. These four measurements correspond to: (1) total area (in $pixels^2$); (2) width (in *pixels*); (3) height (in *pixels*); and (4) perimeter (in *pixels*) of each handwritten character. Observe that these measurements , used mostly for didactic's sake, are not particularly effective for revealing specific geometrical properties of the characters, and therefore unlikely to lead to substantial separation between the groups.

Figure 15 illustrates the four features involved in this dataset.

The values shown in this figure immediately indicates

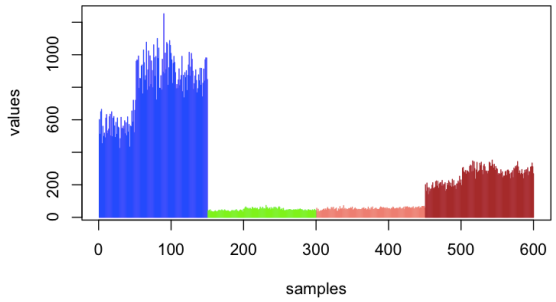

Figure 15: The distribution of the features in the handwritten character dataset. Each type of character has been represented in a specific color. Each group contains the values of the four features, organized in respective sequence. Observe the great variations of magnitudes presented by each of the four features, which can imply biases in subsequent analysis or recognition.

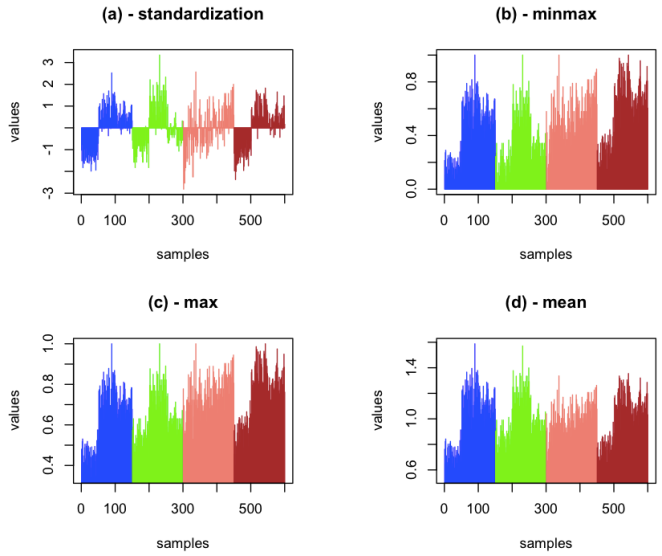

Figure 16: The handwritten character four features as normalized by: (a) *standardization*; (b) *minmax*; (c) *max*; and (d) *mean* normalization schemes. All normalizations have been implemented along each of the features taken separately.

that, though all features are positive, they have markedly distinct respective magnitudes, with the first involving the largest values, following by the forth, third and second. Another important property of these features that can readily appreciated from Figure 15 concerns the fact that all the four features have non-null respective mean values, which is a consequence of all features being positive and presenting variations.

The first decision regarding the normalization dataset actually regards whether to normalize or not. In case the original magnitudes are considered essential for the categorization of the samples, no normalization should be applied. In that case, a heavy influence of the first, and then the second features can be expected.

However, in case the features are to be taken in a relative manner while comparing the samples, it is possible to apply several normalization schemed leading to comparable magnitudes in all four cases. In the case of the present example, we consider the respective standardization as well as *minmax*, *mean* as well as dividing the features by the respective maximum values, referred to as *max*. We observe that the later normalization possibility was not previously considered in this work as it is likely to be susceptible to instabilities caused by sample outliers.

A distinct result has been produced by each of the four considered normalization approaches. As could be expected, the features standardization (a) led to negative values, as well as null means and unit standard deviation for each feature. It is the latter characteristic that implements the magnitudes leveling. The *minmax* approach (b) rolls the values of each feature from 0 to 1, therefore also making them more comensurated. Observe the critical influence of the maximum value within each feature group on the resulting normalization. Instabilities can therefore be caused by outliers with large magnitudes.

The *max* normalization (c), which suffers from this same potential problem, resulted slightly distinct from the *minmax* result as a consequence of the minima of the features not being identical to zero. The normalization implemented by the *mean* scheme, shown in (d), is likely to be relatively less affected by outliers with large values, as the respective averages in each feature group is here taken into account instead of the respective maxima.

Though additional considerations about the specific type of patterns and features in this dataset could be taken into account in order to narrow down on the five possibilities (four normalizations plus the unnormalized values), here we show the effect of all these possibilities while obtaining respective coincidence networks by using the methodology described in [42]. Figure 17 depicts the five respectively obtained coincidence similarity networks. The links correspond to pairs of samples whose coincidence is equal or larger than a respectively adopted threshold.

Several interesting effects can be observed from the obtained results. First, we have that each of the approaches led to markedly distinct coincidence results. In particular, the net in (a) presents the greatest separation between the blue group from the other two types of characters. That is a direct consequence of the fact that a substantial difference can be observed in Figure 15 between the first category and the other two in the case of the first feature (blue). Given that no normalization was applied in (a), this difference predominated while of the comparison

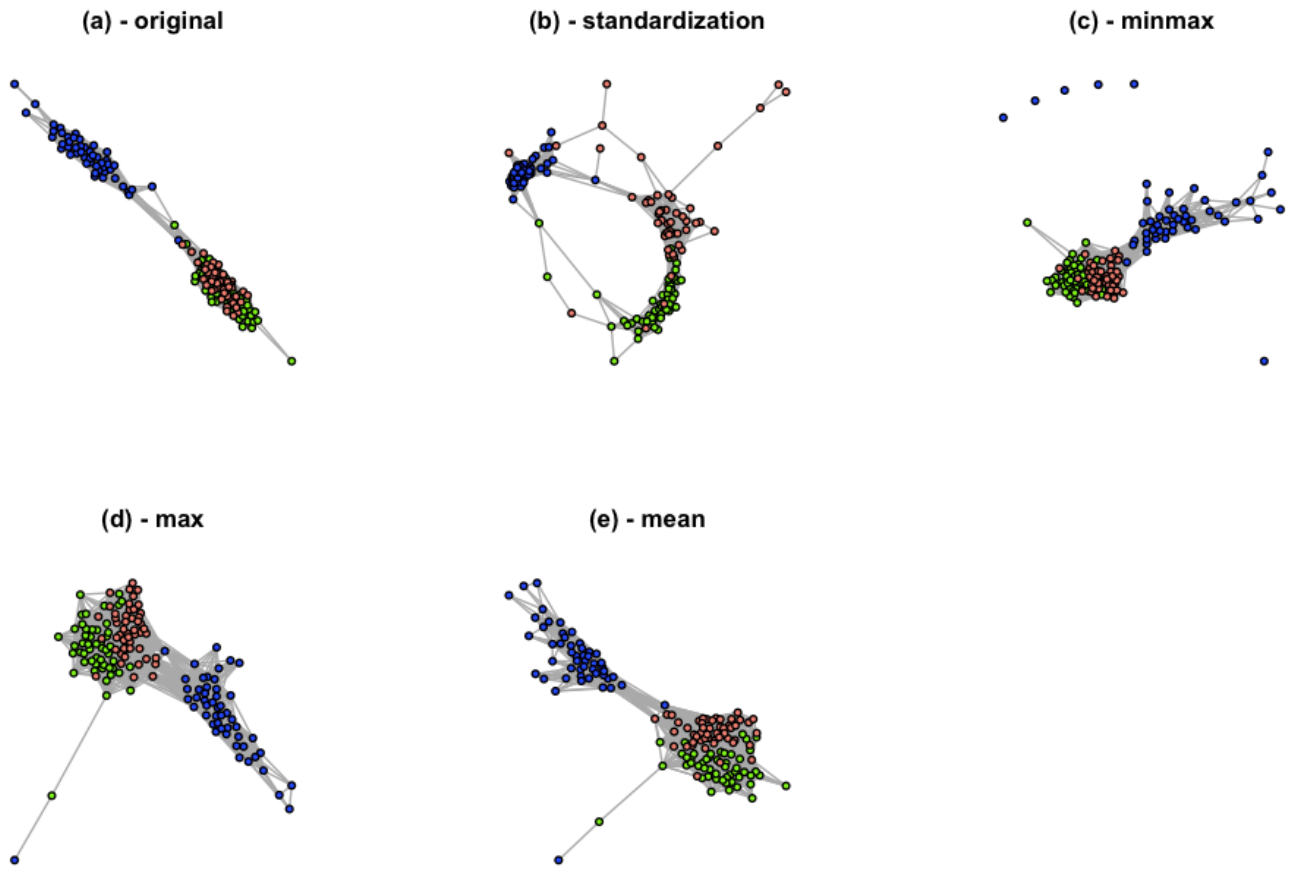

Figure 17: The coincidence similarity networks obtained for the handwritten characters dataset obtained respectively to five normalization possibilities. Remarkably distinct results can be observed for each of these alternatives. The decision on the most appropriate required additional considerations about the data, research interest and specific properties of the adopted features, as well as additional verifications involving validation schemes such as k-fold cross validation approaches. Different coincidence values were chosen so as to end up with most of the samples interconnected. The colors indicate the three types of characters.

between the samples.

The coincidence network obtained for the features standardization, shown in Figure 17(a), presents a the three types of characters defining a ring, implied by similarities identified in this case between the blue and green groups. It is also interesting to observe the satellite branch involving four samples of the third group of characters.

The network obtained respectively to the *minmax* and *max* normalizations are mostly similar, which is a consequence of the minimum values of each feature group not to be much larger than zero. A moderate separation between the blue group from the other two groups can be observed. Interestingly, three points resulted isolated in the case of the former normalization.

The *mean* normalization implied the coincidence network shown in Figure 17(e), which is similar to the nets obtained respectively to the *minmax* and *max* normalizations, though with the difference that the blue group resulted more separated in the case of the *mean* normalization.

The greatest simultaneous separation between all the three groups can be observed respectively to the network obtained by features standardization shown in Figure 17(b). However, this result needs to be considered with special attention and caution because, by ignoring the mean values of the features, the standardization may actually have implied some *bias* on the features that does not necessarily reflects the relationship between the original categories. Cross-validation approaches can be considered for better understanding the effect of the standardization (as well as all other normalization) on the separation between the categories.

It is also interesting to keep in mind that the above analysis could aim not necessarily at the separation between the groups, but as an investigation of the original data as represented by specific normalization on themselves, chose while take other considerations into account.

11 Case Example 4: Known Model

Having discussed normalization respectively to two hypothetical cases, as well as a real-world dataset, we now approach a situation in which everything is know about the possible characteristics of the datasets.

The problem concerns the study of simulated 3D parallelepiped objects with dimension $A \times B \times C$, as illustrated in Figure 18. The values A, B and C of each object is to be considered as a respective feature.

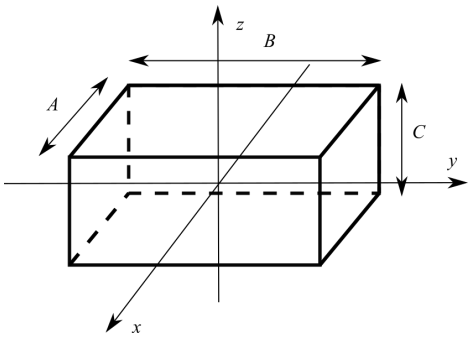

Figure 18: The objects in our forth case example constitute of parallelepipeds as illustrated in this figure scales at the same proportion a.

In principle, these objects are supposed to scale at the same proportion along all the three dimensions respectively to a reference $\tilde{A} \times \tilde{B} \times \tilde{C}$, i.e.:

$$\begin{cases} A = a\ \tilde{A} \\ B = a\ \tilde{B} \\ C = a\ \tilde{C} \end{cases} \tag{52}$$

As a consequence, the coefficients of variation of every group are mutually identical to a specified constant cv, so that:

$$\sigma_A = cv\ \mu_A \tag{53}$$

We henceforth assume the following refence:

$$\tilde{A} = 3; \quad \tilde{B} = 4; \quad \tilde{C} = 2 \tag{54}$$

Observe that the above equations specify the relationship between the three features, but not the distribution of features within each possible type of objects. The latter controls the way in which the objects are associated to the groups. As an example, let us suppose that there are 4 groups of objects, each consisting of the above specified parallelepipeds whose sizes follow a normal density with mean μ_1, μ_2, μ_3, and μ_4, and respective standard deviations σ_1, σ_2, σ_3, and σ_4. Samples are extracted from these groups with probabilities values according to the respective densities. Therefore, we now have a complete *statistical model* of all the aspects regarding the four categories and possible samplings.

Let us make:

$$a_1 = 1;\ cv = 0.5 \Longrightarrow \begin{cases} \mu_{A,1} = A = 3; & \sigma_{A,1} = 1.5 \\ \mu_{B,1} = B = 4; & \sigma_{B,1} = 2.0 \\ \mu_{C,1} = C = 2; & \sigma_{C,1} = 1.0 \end{cases}$$

$$a_2 = 9 \Longrightarrow \begin{cases} \mu_{A,2} = 27; & \sigma_{A,2} = cv\ \mu_A = 1.35 \\ \mu_{B,2} = 36; & \sigma_{B,2} = cv\ \mu_B = 1.8 \\ \mu_{C,2} = 18; & \sigma_{C,2} = cv\ \mu_C = 0.9 \end{cases}$$

$$a_3 = 5 \Longrightarrow \begin{cases} \mu_{A,3} = 15; & \sigma_{A,3} = cv\ \mu_A = 0.75 \\ \mu_{B,3} = 20; & \sigma_{B,3} = cv\ \mu_B = 1 \\ \mu_{C,3} = 10; & \sigma_{C,3} = cv\ \mu_C = 0.5 \end{cases}$$

$$a_4 = 8.5 \Longrightarrow \begin{cases} \mu_{A,4} = 25.5; & \sigma_{A,4} = cv\ \mu_A = 1.275 \\ \mu_{B,4} = 34; & \sigma_{B,4} = cv\ \mu_B = 1.7 \\ \mu_{C,4} = 17; & \sigma_{C,4} = cv\ \mu_C = 0.85 \end{cases}$$

Given that the means of these groups are relatively well separated, except possibly between the groups 2 an 4, they provide a particularly valuable resource for recognition, provided they can be estimated from the available samples with reasonable accuracy. Figure 19 shows the estimated means for successively larger sets of samples with size N respectively to each feature and each of the four types of patterns. It can be readily verified that excellent separation is observed between the means of groups 1, 3, and 2/4, though the separation between groups 2 and 4 is not reliable given the respective statistical oscillations.

Now, let us suppose a recognition problem in which we receive N samples, all from a same group, and the objective is to know which group they belong to. Figure 20(a) illustrates the sets of features respectively to one such sample.

Given that the means provide an interesting resource for group identification, they are addressed first. More specifically, we normalize the original features by removing the oscillating portion while retaining only the estimated mean, yielding the results shown in Figure 20(b). As the means estimated from the 30 samples — namely 14.86, 19.81 and 9.91 are close to the respective reference values in group 3, which is well-separated from the others, we conclude that the 30 supplied samples are of type 3. This example illustrates a situation in which the oscillating part of a set of feature values (or signal) can be disregarded while attention is focused on the mean value.

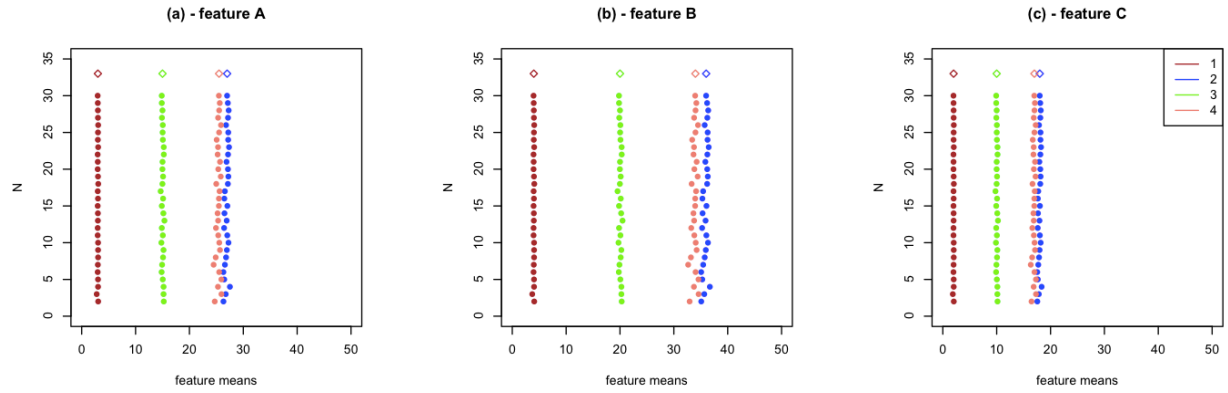

Figure 19: Values of the means of features A, B and C within each of the 4 groups as estimated from successively larger number of samples N. The original mean values are shown as diamonds in the upper portion of the plots. An excellent separation can be observed between the means of categories 1, 3 and 2/4. The means of groups 2 and 4 are very close one another respectively to the respective statistical oscillations, which could lead to identification mistakes.

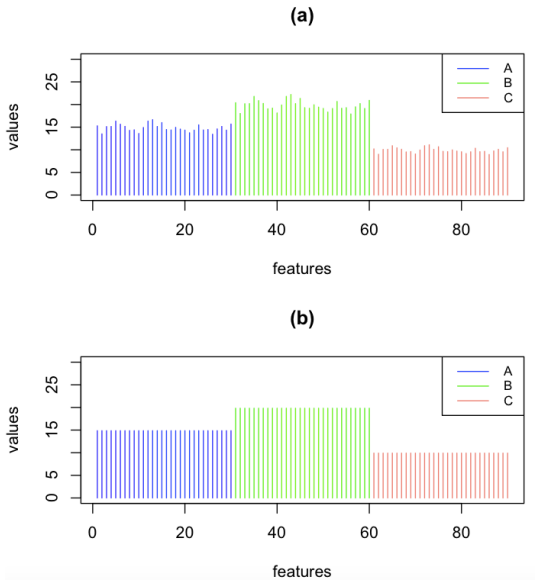

Figure 20: (a) The sets of features A, B and C respective to 30 samples of one of the groups. The problem consist of identifying to which group these samples belong. The normalization by removing the oscillation, leaving only the estimated mean values (b), provides excellent indication about the identity of the group, which can be determined by comparing these averages with those of the groups, indicating that the samples should be categorized as type 1. The removal of the oscillations could only be performed because we know the samples are from a same group, the groups have relatively separated means, and that there are enough samples to provide a reasonably accurate estimation of the means.

Other situations can lead to different choices, such as focusing on the oscillations while removing the mean value, or taken both into account.

It is interesting to observe that any normalization that would imply in removing the means would completely undermine the recognition because, despite their intrinsic dispersion, most of the groups are well characterized by the respective means.

Now, let us suppose that we receive a large number of samples, namely $N = 120$, as shown in Figure 21.

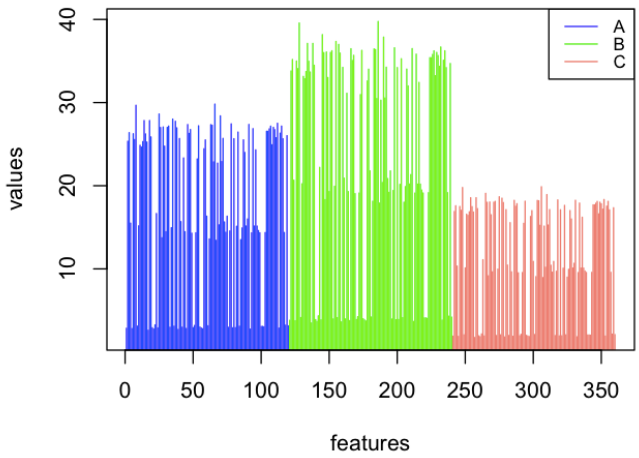

Figure 21: The three sets of features A, B, and C supplied for classification. All features have non-null means.

Now, we aim at normalizing so as to enhance the chances of correct recognition of each of these samples represented by respective features A, B and C. Given that the three sets of features have distinct ranges of magnitudes, it is interesting to verify the possible effect by some normalization scheme implementing some magnitude leveling with each of the three features. Figure 22 illustrates the result of the application of the *mean* normalization described in Section 4.

This normalization reveals that the three sets of features are identical, so that only one of these sets can be retained for subsequent recognition. Given the relatively

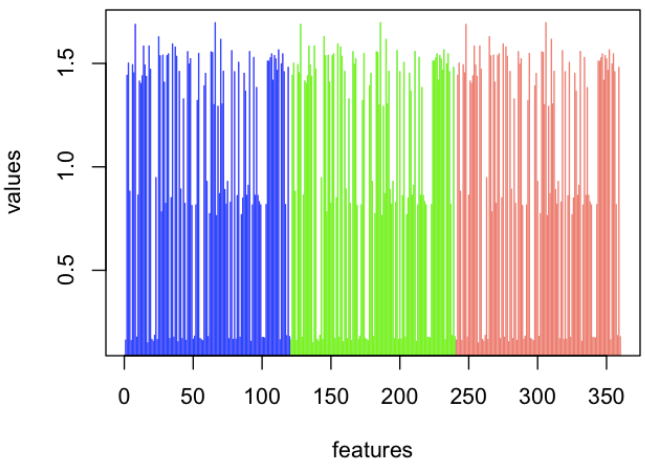

Figure 22: The result of *mean* normalization applied separately to each of the three sets of features in the dataset shown in Fig. 21. It can be readily verified that the three groups are identical, which allows us to use only one feature for subsequent recognition, e.g. by using Bayesian decision theory.

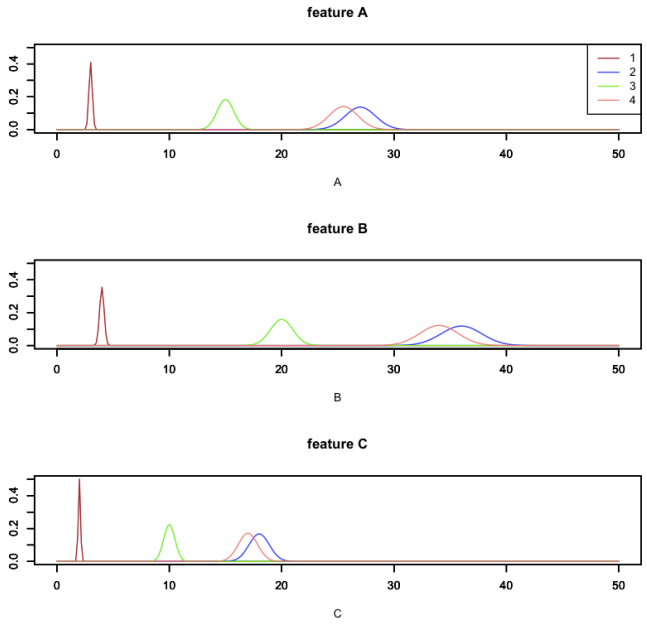

Figure 23: The reference sets of probability functions for our forth case example, respective to each of the three involved features and four categories. The intersection between these density functions define decision regions that can be considered for classifying the samples in Figure 21. Observe that the densities in each of the three plots are mutually proportional one another through the constants a_1, a_2, a_3 and a_4.

good separation of most involved groups, we could resource to Bayesian decision theory by using the decision regions indicated by the intersections between the reference normal densities shown in Figure 23. Exceptionally, as a consequence of the identical scaling of the means and standard deviations underlying the respective reference model, only one of the set-ups in this figure will need to be applied in the case of this specific example.

Let us now assume that other features are available about the parallelepiped objects, such as their weight and color, represented by three components such as R (red), G (green), and B (blue). These additional variables are rather unlikely to present a proportional relationship with the dimensions A, B, C described above. Indeed, they can follow distinct statistic models and incorporate noise or other distortions that should be modeled by different densities and approaches. These cases would possibly require what we shall call *heterogeneous normalization* of the available features, in the sense that each feature or subset of features would be treated in a possibly distinct and respectively more appropriate manner. This possibility is important enough two be highlighted by the following snippet:

When the features in a dataset have distinct natures and characteristics, it may be interesting to implement respective *heterogeneous normalization* of their values. This allows each feature to be pre-processed in specific means so as to enhance their respective important aspects while minimizing unwanted parts or components. This specific, heterogenous normalization requires the separated and joint study and modeling of all involved features, while also taking into account the recognition demands.

Despite its simplicity, the case example discussed in this section illustrates several important aspects involved in features normalization. In particular, the knowledge about the model defining the samples and categories provided valuable subsidies for taking several important decisions, therefore illustrating the importance of having as accurate as possible models of the data, sampling, as well as eventual perturbations as noise and error involved in the data analysis and recognition. It should be observed that the procedure adopted for the analysis of this case example is completely specific and unlikely to work when transported directly to other problems and datasets.

Another interesting observation is that several real-world situations may be underlain by joint scaling of several measurements as in the case of the situation discussed above. For instance, the features 1 and 4, and to some

smaller extent also relatively to feature 3 in the handwritten character database example (see Figure 15) present a relationship that resembles those observed in our forth case example.

Other interesting possibilities relating scaling relationships between measurements are addressed in the area of *allometry* (e.g. [45, 46, 47, 48]), namely the existence of interrelationships between pairs of features X and Y of the type:

$$Y = aX^b \tag{55}$$

with $a, b \in \mathcal{R}$.

12 Concluding Remarks

With the continuously increasing applications and importance of data analysis and respective supervised and unsupervised recognition, the issue of how to normalize the respectively involve data, represented in terms of sets of samples respective features, becomes more and more relevant.

The present work addressed this important issue from the perspective of modeling the involved signals and sets of features in terms of mathematical formulae that can accommodate the respective decomposition of each type of signal in terms of terms of particular relevance specifically to each problem and dataset.

After presenting the several involved concepts and methods — including random variables, transformations, modeling formulae, several standardization approaches, as well as the Pearson correlation and coincidence similarity approaches to patter comparison, we proceeded to studying two case examples involving three categories synthesized through respective formulae involving three terms, as well as a real-world dataset involving three types of handwritten characters represented by four fetures. This study involved taking each of the considered normalizations followed by comparison by the Pearson correlation and coincidence similarity approaches.

In addition to the presented concepts and methods related to the data normalization, the experimentally obtained results highlighted several critically important related aspects, especially the critical influence that distinct normalizations can have on the overall analysis and recognition results.

Another point of special relevance is that the above observed effect is often strongly modulated by the respective choices of combinations between normalizations and respective comparison methods. For instance, the use of several normalization schemes discussed in this work have no impact on the Pearson correlation analysis, as this comparison approach involves the removal of constant and/or mean respective feature values. However, those same normalizations can have a strong impact on the results while of application of the coincidence similarity approach.

In addition, it has been verified that the small scale component of features may lead to comparison interrelations that are completely distinct. This is not an artifact, but actually a reflection of the fact that the similarity between signals and sets of features depends on the scale of the analysis, as well as the selection of specific combinations of terms in their original formulae to be used in the comparison.

Among other observed results and effects, the identified effects of normalization yield some important conclusions. First, we have that normalization has to be chosen carefully while taking into account existing or putative models of the signals and sets of features in terms of respective formulae involving possible decomposition in terms of special signification for each application, is necessary. Then, we also have that oftentimes each of the signal or feature that constitute the patterns representation may have distinct nature or mathematical model, therefore implying that they should be normalized accordingly to possibly *distinct* schemes and manners, in a heterogeneous feature-by-feature way.

Another important conclusion that can be reached from the reported developments concerns the fact that the choice of normalization can be performed while taking into account distinct requirements, including enhanced robustness to noise and perturbation in the data, optimal separation between the groups while of respective recognition and clustering, or so as to emphasize some specific aspects of interest implied by each problem and application, among other possibilities. However, once an optimization parameter has been set, it is possible to perform an optimization between the several normalization alternatives respectively to each of the involved features, therefore suggesting a *normalization selection* research area analogous to that of *feature selection* (e.g. [49, 50, 51, 52]). This is particularly reasonable given that, as discussed in the present work, features normalization can actually be understood as a manner to derived new features from the original set of measurements.

When aiming at a specific optimum criterion, such as maximum separation, special attention and care need to be invested respectively to these efforts leading to biased sampling, which can undermine the data analysis and recognition. Therefore, cross-validation and other performance approaches (e.g. [8]) need to be incorporated into each approach.

All in all, it can be inferred from the several concepts, methods and results presented and discussed in this work that *pattern recognition* is actually closely related to scientific modeling, sharing several aspects. First, we have

that both these areas aim at developing models of phenomena/data as a means to better understand and make predictions about the studied problems. Then, we have that both rely on quantification of properties of the phenomena under study, which are understood as variables in scientific modeling and features in pattern recognition. Both these approaches often involve selecting and preprocessing the features, which are critically important for the modeling. In both cases, normalization has great importance, providing a bridge between the raw dataset and the subsequent analysis and modeling of the problems.

In particular, what has been shown in particular is that more effective approaches to pattern recognition may benefit from a more comprehensive understanding of the chosen features and their properties, which can be achieved by developing respective statistical and/or other types of modeling. Thus, in a sense, pattern recognition could be also understood as an important case of scientific modeling characterized by approaches that involve concepts and methods that are more general than those in scientific modeling, where often highly specific resources from related areas are often required.

As with most approaches to data analysis and pattern recognition, the choices of methods and respectively obtained results should not be understood as having absolute implications, or being relatively better or correct. These choices results need to be taken relatively to the type of data and questions of interest in each specific problem, and further validated through several means.

The several concepts, methods, and results presented in this work paves the way to a wide range of possible developments. For instance, it would be interesting to consider larger number of features and patterns, leading to more elaborate network representations. In addition, other comparison and recognition indices and approaches could be evaluated by using the suggested concepts and methods. Another issue of particular interest would be to try to identify how each of the combinations of normalization and comparison approaches result coherent with the way in which humans visually perceive and compare patterns.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

Note:

As all other preprints by the author, this work is possibly being considered by a scientific journal. Respective modification, commercial use, or distribution of any of its parts are not possible. Many of the preprints by the author are also available in HAL and arXiv. This work can also be cited by using the DOI number or article identification link. Thanks for reading.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.

[2] K. Fukunaga. *Statistical Pattern Recognition*. Morgan Kaufmann, San Diego, 1990.

[3] B. K. P. Horn. *Robot Vision*. McGraw Hill, Cambridge, 1986.

[4] L. da F. Costa. *Shape Classification and Analysis: Theory and Practice*. CRC Press, Boca Raton, 2nd edition, 2009.

[5] E. R. Davies. *Machine Vision*. Morgan Kaufmann, Amsterdam, 2005.

[6] K. Koutrombas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2008.

[7] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Trans. PAMI*, 22:4–37, 2000.

[8] L. da F. Costa. Supervised and unsupervised pattern recognition and their performance. https://www.researchgate.net/publication/360936159_Supervised_and_Unsupervised_Pattern_Recognition_and_their_Performance, 2022.

[9] J. L Horner. Metrics for assessing pattern-recognition performance. *Applied Optics*, 31:165–166, 1992.

[10] R. A. Johnson and D.W. Wichern. *Applied multivariate analysis*. Prentice Hall, 2002.

[11] N. Mukhopadhyay. *Probability and Statistical Inference*. CRC Press, New York, 2000.

[12] S. Haykin. *Neural Networks And Learning Machines*. McGraw-Hill Education, 9th edition, 2013.

[13] John Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32(4):496–501, 2002.

[14] J. Sola and J. Sevilla. Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on nuclear science*, 44(3):1464–1468, 1997.

[15] D. Singh and B. Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, 2020.

[16] S. Patro and K. K. Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[17] L. Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2008.

[18] L. Al Shalabi and Z. Shaaban. Normalization as a preprocessing engine for data mining and the approach of preference matrix. In *2006 International conference on dependability of computer systems*, pages 207–214. IEEE, 2006.

[19] L. da F. Costa. Revisiting colocalization from the perspective of similarity. https://www.researchgate.net/publication/360777098_Revisiting_Colocalization_from_the_Perspective_of_Similarity, 2022.

[20] L. da F. Costa. Further generalizations of the Jaccard index. https://www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index, 2021. [Online; accessed 21-Aug-2021].

[21] L. da F. Costa. On similarity. https://www.sciencedirect.com/science/article/pii/S037843712200334X, 2022. Physica A: Statistical Mechanics and its Applications, 127456.

[22] D. Ghosh and A. Vogt. Outliers: An evaluation of methodologies. In *Joint statistical meetings*, volume 2012, 2012.

[23] D. M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[24] C. H. Comin, J. Tejada, M. P. Viana, A. C. Roque, and L. da F. Costa. Archetypes and outliers in the neuromorphological space. In *The Computing Dendrite*, pages 41–59. Springer, 2014.

[25] Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

[26] Emanuel Parzen. *Stochastic processes*. SIAM, 1999.

[27] S. M. Ross, J. J. Kelly, R. J. Sullivan, W. J. Perry, D. Mercer, R. M. Davis, T. D. Washburn, E. V. Sager, J. B. Boyce, and V. L. Bristow. *Stochastic processes*, volume 2. Wiley New York, 1996.

[28] M. K. Vijaymeena and K. Kavitha. A survey on similarity measures in text mining. *Machine Learning and Applications*, 3(1):19–28, 2016.

[29] M. Brusco, J. D. Cradit, and D. Steinley. A comparison of 71 binary similarity coefficients: The effect of base rates. *PLOS One*, 16(4):e0247751, 2021.

[30] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Ketters, H. Rousseau, and A. Vanhoutte. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing and Management*, 25(3):315–318, 1989.

[31] C. E. Akbas, A. Bozkurt, M. T. Arslan, H. Aslanoglu, and A. E. Cetin. L1 norm based multiplication-free cosine similiarity measures for big data analysis. In *IEEE Computational Intelligence for Multimedia Understanding (IWCIM)*, France, Nov. 2014.

[32] K. Kavitha, B. Sandhya, and B. T. Rao. Evaluation of distance measures for feature based image registration using Alexnet. *International Journal of Advanced Computer Science and Applications*, 9(10), 2018.

[33] T. Ibrikci, M.E. Brandt, G. Wang, and M. Acikkar. Mahalanobis distance with radial basis function network on protein secondary structures. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society. Engineering in Medicine and Biology Society, Proceedings of the Annual International Conference of the IEEE.*, pages 2184–2185, Houston, USA, Jan. 2003.

[34] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société vaudoise des sciences naturelles*, 37:241–272, 1901.

[35] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37:547–549, 1901.

[36] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.

[37] D. E. Knuth. *The Art of Computing*. Addison Wesley, 1998.

[38] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30:36—66, 1989.

[39] W. D. Blizard. The development of multiset theory. *Modern Logic*, 4:319–352, 1991.

[40] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.

[41] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.

[42] L. da F. Costa. Coincidence complex networks. https://iopscience.iop.org/article/10.1088/2632-072X/ac54c3, 2022. J. Phys.: Complexity, (3): 015012.

[43] L. da F. Costa. Multiset neurons. https://www.researchgate.net/publication/356042155_Multiset_Neurons, 2021.

[44] B. Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic Publisher, Dordrecth, 1996.

[45] J. Gayon. History of the concept of allometry. *American zoologist*, 40(5):748–758, 2000.

[46] C. P. Klingenberg. Multivariate allometry. In *Advances in morphometrics*, pages 23–49. Springer, 1996.

[47] T. Kohyama. Significance of architecture and allometry in saplings. *Functional Ecology*, pages 399–404, 1987.

[48] K. J. Niklas. *Plant allometry: the scaling of form and process*. University of Chicago Press, 1994.

[49] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[50] V. Kumar and S. Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

[51] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[52] K. Kira and L. A. Rendell. A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier, 1992.