A Generalized Methodology for Data Analysis

Plamen Angelov, *Fellow, IEEE*, Xiaowei Gu, and Jose Principe, *Fellow, IEEE*

***Abstract*—Based on a critical analysis of data analytics and its foundations, we propose a functional approach to estimate data ensemble properties, which is based entirely on the empirical observations of discrete data samples and the relative proximity of these points in the data space and hence named empirical data analysis (EDA). The ensemble functions include the non-parametric square centrality (a measure of closeness used in graph theory) and typicality (an empirically derived quantity which resembles probability). A distinctive feature of the proposed new functional approach to data analysis is that it does not assume randomness or determinism of the empirically observed data, nor independence. The typicality is derived from the discrete data directly in contrast to the traditional approach where a continuous probability density function (pdf) is assumed a priori. The typicality is expressed in a closed analytical form that can be calculated recursively and, thus, is computationally very efficient. The proposed non-parametric estimators of the ensemble properties of the data can also be interpreted as a discrete form of the information potential (known from the information theoretic learning theory as well as the Parzen windows). Therefore, EDA is very suitable for the current move to a data-rich environment where the understanding of the underlying phenomena behind the available vast amounts of data is often not clear. We also present an extension of EDA for inference. The areas of applications of the new methodology of the EDA are wide because it concerns the very foundation of data analysis. Preliminary tests show its good performance in comparison to traditional techniques.**

***Index Terms*—data mining and analysis, machine learning, pattern recognition, probability, statistics.**

I. Introduction

CURRENTLY, there is a growing demand in Machine Learning, Pattern Recognition, Statistics, Data Mining and a number of related disciplines broadly called Data Science, for new concepts and methods that are **centered on the actual data,** the evidence collected from the **real world** rather than at **theoretical *prior* assumptions** which need to be further confirmed with the experimental data (e.g the Gaussian assumption). The core of the statistical approach is the definition of a random variable, i.e. a functional measure from the space of events to the real line, which defines the probability law [1]–[4]. The probability density function (*pdf)* is, by definition, the derivative of the cumulative distribution function (*cdf).* It is well known that differentiation can create numerical problems in both practical and in theoretical aspects and is a challenge for functions which are not analytically defined or are complex. In reality, we usually do not have independent and identically distributed (*iid)* events, but we do have correlated, interdependent (albeit in a complex and often unknown manner) data from different experiments which complicates the procedure.

The appeal of the traditional statistical approach is its solid mathematical foundation and the ability to provide guarantees of performance, when data is plenty ($N\rightarrow\infty$), and created from the same distribution that was hypothesized in the probability law. The actual data is usually discrete (or discretized), which in traditional probability theory and statistics are modeled as a realization of the random variable, but one does not know *a priori* their distribution. If the *prior* data generation hypothesis is verified, good results can be expected; otherwise this opens the door for many failures.

Even in the case that the hypothesized measure meets the realizations, one has to address the difference of working with realizations and random variables, which brings the issue of choosing estimators of the statistical quantities necessary for data analysis. This is not a trivial problem, and is seldom discussed in data analysis. The simple determination of the probability law (the measure of the random variable) that explains the collected data is a hard problem as studied in density estimation [1]–[3]. Moreover, if we are interested in statistical inference, for instance, similarity between two random variables using mutual information, the problem gets even harder because different estimators may provide different results [5]. The reason is that very likely the functional properties of the chosen estimator do not preserve all the properties embodied in the statistical quantity. Therefore, they behave differently in the finite (and even in the infinite) sample case. An alternative approach is to proceed **from the realizations to the random variables**, which is the **reverse** direction of **the statistical approach**. The literature has several excellent examples of this approach, in the area of measures of association. For instance, Pearson's correlation coefficient is perfectly well defined in realizations, as well as in random variables. Likewise, Spearman's ρ [6], Kendal's τ [7], are other examples of measures of association well defined in both the realization and the random variables. However, the problem with this approach is that the statistical properties of the

Manuscript received July 2016. This work was partially supported by The Royal Society grant IE141329/2014 "Novel Machine Learning Paradigms to address Big Data Streams".

Plamen P. Angelov and Xiaowei Gu are with School of Computing and Communications, Lancaster University Lancaster, LA1 4WA. (e-mail: {p.angelov, x.gu3}@lancaster.ac.uk)

José C. Príncipe is with Computational NeuroEngineering Laboratory, Department of Electrical and Computer Engineering, University of Florida, USA. (e-mail: principe@cnel.ufl.edu)

measures in the random variables are not directly known, and may not be easily obtained. A good example of the latter is the generalized measure of association, which is well defined in the realizations, but not all of the properties are known in the random variables [8]. Therefore, there are advantages and disadvantages in each approach, but from a practical point of view, the non-parametric approach is very appealing because we can go beyond the framework of statistical reasoning to define new operators and still cross-validate the solutions with the available data using non-parametric hypothesis tests. A good example is least squares versus regression. One can always apply least squares to any data type, deterministic or stochastic. If the data is stochastic the solution is called regression, but the result will be the same, because the autocorrelation function is a property of the data, independent of its type. The difference shows up only in the interpretation of the solution; most importantly, the statistical significance of the result can only be assessed using regression.

A more recent alternative is to approximate the distributions using non-parametric, data-centered functions, such as particle filters [9], entropy-based information-theoretic learning [5], etc. On the other hand, partially trying to address the same problems, in 1965 L. Zadeh introduced fuzzy sets theory [10], which completely departed from objective observations and moved (similarly to the belief-based theory [8] introduced a bit later) to the subjectivist definition of uncertainty. A later strand of fuzzy set theory (data driven approach developed mainly in 1990s) attempted to define the membership functions based on experimental data. It stands in between probabilistic and fuzzy representations [11], however, this approach requires an assumption on the type of membership function. An important challenge is the posterior distribution approximation. Approximate inference can be done employing maximum *a posteriori* criteria which requires complex optimization schemes involving, for example, the expectation maximization algorithm [1]–[3].

In this paper, we present a systematic methodology of non-parametric estimators recently introduced in [12]–[15] for discrete sets using ensemble statistical properties of the data derived entirely from the experimental discrete observations and extend them to continuous spaces. These include the *cumulative proximity* (q), *centrality* (C), *square centrality* (q^{-1}), *standardized eccentricity* (ε), *densit*y (D) as well as ***typicality***, (τ) which can be extended to continuous spaces, resembling the information potential obtained from Parzen windows [1]–[4] in Information Theoretic Learning (ITL) [5]. Its discrete version sums up to 1 while its continuous version integrates to 1 and is always positive; however, its values are always less than 1 unlike the *pdf* values that can be greater than 1. Additionally, the *typicality* is only defined for feasible values of the independent variable while the *pdf* can extend to infeasible values, e.g. negative height, distance, weight, absolute temperature, etc. unless specifically constraint [12]–[15]. We further consider *discrete local* (τ) and *global* (τ^D) versions. Then, we introduce an automatic procedure for identifying the local modes/maxima of τ^D as well as a procedure for reducing the amount of the local maxima/modes and extend the non-parametric estimators to the continuous domain by introducing the *continuous global density,* D^G *and typicality*, τ^G, which further involves integral for normalization. Furthermore, we demonstrate that the *continuous global typicality* does integrate to 1 exactly as the traditional *pdf* (while being free form the restrictions the latter has). This is a new and significant result which makes *continuous global typicality* an alternative to the *pdf*. This strengthens the ability of the **empirical data analysis (EDA)** framework for objectively investigating the unknown data pattern behind the data and opens up the framework for inference. The methodology is exemplified with a Naïve EDA classifier based on τ^G.

II. Theoretical Basis - Discrete Sets

In this section, we start by presenting EDA foundations in discrete sets [12]-[15] for completeness and further clarity. Firstly, let us consider a real metric space $\mathbf{R}^K$ and assume a particular data set or stream $\{\boldsymbol{x}\}_N = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\} \in \mathbf{R}^K$; with $\boldsymbol{x}_i = \left[x_{i,1}, x_{i,2}, ..., x_{i,K}\right]^{\mathrm{T}}$; $i = 1, 2, \ldots, N$, where subscripts denote data samples (for a set) or the time instances when they arrive (for a stream). Within the data set/stream, some data samples may repeat more than once, namely, $\exists \boldsymbol{x}_i = \boldsymbol{x}_j, i \neq j$. The set of sorted unique data samples, denoted by $\{\boldsymbol{u}\}_{L_N} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_{L_N}\}$ (where $\{\boldsymbol{u}\}_{L_N} \subseteq \{\boldsymbol{x}\}_N$, $1 < L_N \leq N$) and the number of occurrence, denoted by $\{f\}_{L_N} = \{f_1, f_2, ..., f_{L_N}\}$ can be determined automatically based on the data. With $\{\boldsymbol{u}\}_{L_N}$ and $\{f\}_{L_N}$, the primary data set/stream $\{\boldsymbol{x}\}_N$ can be reconstructed. In the remainder of this paper, all the derivations are conducted in the n^{th} time instance except when specifically declared otherwise. The most obvious choice of $\mathbf{R}^K$, is the Euclidian space with the Euclidean distance, but we can also extend EDA definitions to Hilbert spaces, and Reproducing Kernel Hilbert spaces. We can, moreover, consider different types of distances within these spaces motivated by the purposes of the analysis that exploit information available from the source that generated the samples or definitions that are appropriate for data analysis. Within EDA, we introduce:

a) *cumulative proximity*, q [12]–[15];
b) *square centrality,* q^{-1} ;
c) *eccentricity,* ξ [12]–[15];
d) *standardized eccentricity,* ε [12]–[15];
e) *discrete local density,* D [12]–[15];
f) *discrete local typicality*, τ [14], [15];
g) *discrete global typicality,* τ^D [14], [15];
h) *continuous local density,* D^L ;
i) *continuous global density,* D^G, and
j) *continuous global typicality,* τ^G.

The *discrete global typicality*, τ^D addresses the global properties of the data and will be introduced in the next section. For inference, the *continuous local (D^L), global density (D^G)* and the *continuous global typicality, (τ^G)* will be described in detail in section IV.

A. Cumulative Proximity and Square Centrality

For every point $\boldsymbol{x}_i \in \{\boldsymbol{x}\}_N$; $i = 1,2,...,N$ one may want to quantify how *close* or *similar* this point is to all other data points from $\{\boldsymbol{x}\}_N$. In graph theory, ***centrality*** is used to indicate the most important vertices within a graph. A measure of ***centrality*** [16], [17] is defined as a sum of distances from a point x_i to all other points:

$$c_N(\boldsymbol{x}_i) = \frac{1}{\sum_{j=1}^{N} d(\boldsymbol{x}_i, \boldsymbol{x}_j)}; \quad \boldsymbol{x}_i \in \{\boldsymbol{x}\}_N; \quad 1 \le i \le N; \quad L_N > 1 \quad (1)$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the distance/similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which can be, but not limited to Euclidean, Mahalanobis, cosine, etc.

Its importance comes from the fact that it provides centrality information about each data sample in a scalar or vector form. We previously defined [12]–[15] the *cumulative proximity* $q_N(\boldsymbol{x}_i)$ as,

$$q_N(\boldsymbol{x}_i) = \sum_{j=1}^{N} d^2(\boldsymbol{x}_i, \boldsymbol{x}_j); \quad \boldsymbol{x}_i \in \{\boldsymbol{x}\}_N; \quad L_N > 1 \quad (2)$$

which can be seen as inverse centrality with a square distance

Cumulative proximity [12]–[15] is a very important association measure derived empirically from the observed data without making any *prior* assumptions about their generation model and plays a fundamental role in deriving other EDA quantities. The complexity for computing the *cumulative proximities* of all samples in $\{\boldsymbol{x}\}_N$ is $O(N^2)$. As a result, the computational complexity of other EDA quantities for $\{\boldsymbol{x}\}_N$, which can be derived directly from *cumulative proximity* is $O(N)$. For many types of distance/similarity, i.e. Euclidean distance, Mahalanobis distance, cosine similarity, etc., with which the *cumulative proximity* can be calculated recursively [14], the complexity for calculating the *cumulative proximities* of all the samples in $\{\boldsymbol{x}\}_N$ is reduced to $O(N)$ as well.

In a very similar manner, we can consider *square centrality* as the inverse of the *cumulative proximity,* defined as follows:

$$q_N^{-1}(\boldsymbol{x}_i) = \frac{1}{\sum_{j=1}^{N} d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}; \quad L_N > 1 \quad (3)$$

B. Eccentricity

The *eccentricity*, ξ_N, defined as a normalized *cumulative proximity*, is another very important association measure derived empirically from the observed data without making any *prior* assumptions about their generation model [12]–[15]. It quantifies data samples away from the mode, useful to represent distribution tails and anomalies/outliers. It is derived by normalizing q_N and taking into account all possible data samples. It plays an important role in anomaly detection [14], [15] as well as for the estimation of the *typicality* as it will be detailed below. The *eccentricity* (ξ_N) of a particular data sample $\boldsymbol{x}_i$ in the set $\{\boldsymbol{x}\}_N$ ($L_N > 1$) is calculated as follows [12]–[15]:

$$\xi_N(\boldsymbol{x}_i) = \frac{2q_N(\boldsymbol{x}_i)}{\sum_{j=1}^{N} q_N(\boldsymbol{x}_j)} = \frac{2\sum_{j=1}^{N} d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sum_{h=1}^{N}\sum_{j=1}^{N} d^2(\boldsymbol{x}_j, \boldsymbol{x}_h)}; \quad L_N > 1 \quad (4)$$

where the coefficient 2 is included to normalize *eccentricity* between 0 and 1, i.e.:

$$0 \le \xi_N(\boldsymbol{x}_i) \le 1 \quad (5)$$

Here, we also introduce *standardized eccentricity, ε,* which does not decrease as fast as eccentricity with the increase of the amount of data, *N* and is calculated as follows:

$$\varepsilon_N(\boldsymbol{x}_i) = N\xi_N(\boldsymbol{x}_i) = \frac{2q_N(\boldsymbol{x}_i)}{\frac{1}{N}\sum_{j=1}^{N} q_N(\boldsymbol{x}_j)}; \quad L_N > 1 \quad (6)$$

Based on the expression of the *standard eccentricity* (namely, equation (6)) one can see that the data samples which are far away from the majority tend to have higher *standard eccentricity* values compared with others. Thus, the *standard eccentricity* can serve as an effective measure of the ***tail*** of data distribution without the need of clustering the data in advance. Combining the *standard eccentricity* with the well-known *Chebyshev inequality* [18], which discribes the probability that certain data sample $\boldsymbol{x}$ is more than $n\sigma$ (σ denotes the standard deviation) distance away from the mean, we get the EDA version of the Chebyshev inequality as follows [12], [14]:

$$P\left(\varepsilon_N(\boldsymbol{x}) \le n^2 + 1\right) \ge 1 - \frac{1}{n^2} \quad (7)$$

The *Chebyshev inequality* expressed by the *standard eccentricity* provides a more elegant form for anomaly detection. For example, if $\varepsilon_N(\boldsymbol{x}) > 10$, $\boldsymbol{x}$ has exceeded the 3σ limitation, and can be categorized as an anomaly.

C. Discrete Local Density

Discrete local density is defined as the inverse of *standardized eccentricity* and plays an important role in data analysis using EDA ($i = 1,2,...,N$; $L_N > 1$):

$$D_N(\boldsymbol{x}_i) = \varepsilon_N^{-1}(\boldsymbol{x}_i) = \frac{\sum_{j=1}^{N} q_N(\boldsymbol{x}_j)}{2Nq_N(\boldsymbol{x}_i)} = \frac{\sum_{j=1}^{N}\sum_{l=1}^{N} d^2(\boldsymbol{x}_j, \boldsymbol{x}_l)}{2N\sum_{l=1}^{N} d^2(\boldsymbol{x}_i, \boldsymbol{x}_l)} \quad (8)$$

For example, if the Euclidean distance is used, the *density* can be expressed as ($i = 1,2,...,N$; $L_N > 1$):

$$D_N\left(\boldsymbol{x}_i\right)=\frac{1}{1+\frac{\left\|\boldsymbol{x}_i-\boldsymbol{\mu}_N\right\|^2}{X_N-\boldsymbol{\mu}_N^{\mathrm{T}}\boldsymbol{\mu}_N}} \tag{9}$$

where $\boldsymbol{\mu}_N$ is the mean of $\{\boldsymbol{x}\}_N$; X_N is the mean of $\left\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right\}_N$; $\boldsymbol{\mu}_N$ and X_N can be updated recursively using [19]:

$$\begin{cases}\boldsymbol{\mu}_k=\frac{k-1}{k}\boldsymbol{\mu}_{k-1}+\frac{1}{k}\boldsymbol{x}_k; \quad \boldsymbol{\mu}_1=\boldsymbol{x}_1; \\ X_k=\frac{k-1}{k}X_{k-1}+\frac{1}{k}\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k; \quad X_1=\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1;\end{cases} \quad k=1,2,...,N$$

As we can see from equation (9), the *discrete local density* itself can be viewed as a univariate **Cauchy function** while **there is no assumption or any pre-defined parameter involved in the derivation** besides the definition of the distance function (Euclidean distance used here).

D. Discrete Local Typicality

Discrete local typicality was firstly introduced in [13], and called unimodal typicality. In this paper, it is redefined as the normalized *local density* ($i=1,2,...,N$; $L_N>1$):

$$\tau_N\left(\boldsymbol{x}_i\right)=\frac{D_N\left(\boldsymbol{x}_i\right)}{\sum_{j=1}^{N}D_N\left(\boldsymbol{x}_j\right)}=\frac{q_N^{-1}\left(\boldsymbol{x}_i\right)}{\sum_{j=1}^{N}q_N^{-1}\left(\boldsymbol{x}_j\right)} \tag{10}$$

The *discrete local typicality* resembles the traditional unimodal probability mass function (*pmf*), but it is automatically defined in the data support unlike the *pmf* which may have non-zero values for infeasible values of the random variable unless specifically constraint.

The *discrete local density* resembles membership functions of a fuzzy set having value of 1 for $\boldsymbol{x}=\boldsymbol{\mu}$ while the *discrete local typicality* resembles *pmf* with the sum of $N\tau_N$ values being equal to *1* and values for both *D* and τ being from the interval [0,1].

As an example, the *square centrality*, *standardized eccentricity*, *discrete local density* and *typicality* of real climate dataset (wind chill and wind gust) measured in Manchester, UK for the period 2010-2015 [20] are presented in Supplementary Fig. 1. In these examples, Euclidean distance is used.

III. Theoretical Basis: Discrete Global Typicality

In this section, we will consider the more realistic case when data distributions are multimodal. Traditionally, this requires identifying local peaks/modes by clustering, expectation maximization, optimization, etc. [1]–[3], [21]–[23]. Within EDA, the *discrete global typicality* (τ^D) is derived automatically from the data with no user input and can quantify multimodality. It is based on the *local cumulative proximity*, *square centrality*, *eccentricity* and *standardized eccentricity*. The only requirements to define the *discrete global typicality* are the raw data and the type of distance metric (which can be any).

A. Discrete Global Typicality

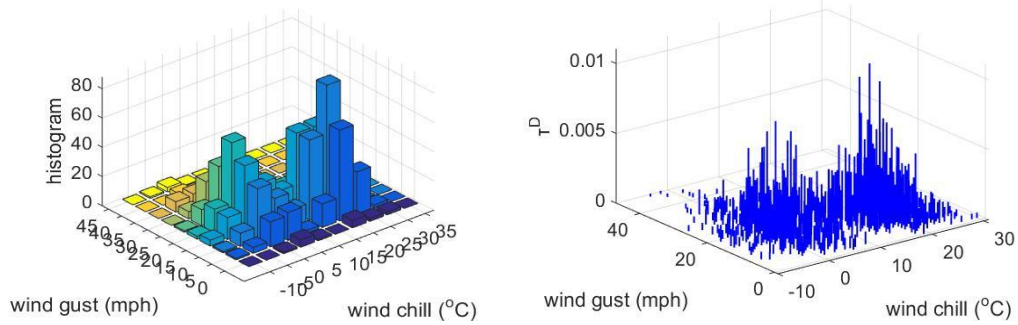
Expressions (9)-(10) provide definitions of local operators that are very appropriate to quantify the peak point ($\boldsymbol{x}^*$) of unimodal discrete functions. Moreover, if the peak coincides with the global mean $\boldsymbol{\mu}_N$ ($\boldsymbol{x}^*=\boldsymbol{\mu}_N$), then the value of the *local density* is equal to 1: $D_N\left(\boldsymbol{\mu}_N\right)=1$. A similar property having a maximum, though its value is <1, is also valid for the traditional probability by definition and according to the central limit theorem [1]–[3]. In reality, data distributions are usually multimodal [21]–[24], therefore the local description should be improved. In order to address this issue, the traditional probability theory often involves mixture of unimodal distributions, which requires estimation of number of modes and it is not easy [24]. Within the EDA framework, we provide the *discrete global typicality,* τ^D, directly from the dataset, which provides multimodal distributions automatically without the need of user decisions and only requires a threshold for robustness against outliers.

The *discrete global typicality* of a unique data sample is expressed as a combination of the *normalized discrete local density* weighted by the corresponding frequency of occurrence of this unique data sample ($i=1,2,...,L_N; L_N>1$):

$$\tau_N^D\left(\boldsymbol{u}_i\right)=\frac{f_iD_N\left(\boldsymbol{u}_i\right)}{\sum_{j=1}^{L_N}f_jD_N\left(\boldsymbol{u}_j\right)}=\frac{f_jq_N^{-1}\left(\boldsymbol{u}_i\right)}{\sum_{j=1}^{L_N}f_jq_N^{-1}\left(\boldsymbol{u}_j\right)} \tag{11}$$

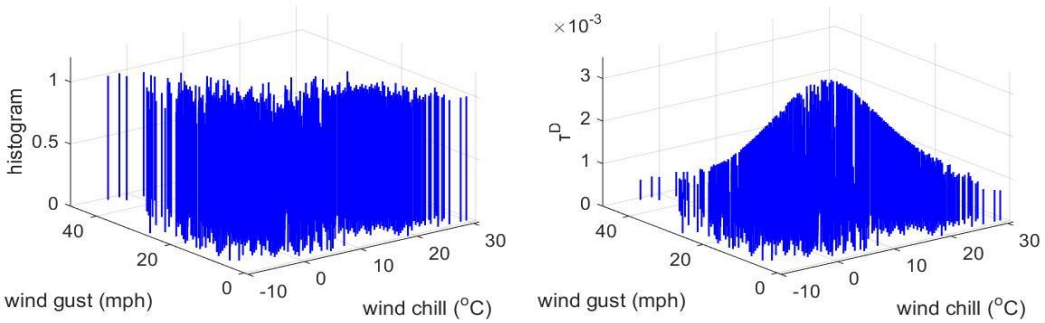
where $q_N^{-1}\left(\boldsymbol{u}_i\right)$ and $D_N\left(\boldsymbol{u}_i\right)$ are the *square centrality* and the *discrete local density* of a particular data sample, $\boldsymbol{u}_i$ calculated from $\{\boldsymbol{u}\}_{L_N}$ only.

This expression is very fundamental, because, in fact, it combines information about repeated data values and the scattering across the data space, and resembles the well-known membership functions of fuzzy sets. We further explain this link in a publication that is currently under review [25].



a) Histogram b) *Discrete global typicality* τ_N^D

Fig.1. Histogram and *discrete global typicality* τ_N^D of the real climate data [20] using Euclidean distance



(a) Histogram with very small quantization (b) *Discrete global typicality,* τ_N^D

Fig. 2 Histogram and *discrete global typicality* for the unique data samples

One can easily appreciate from Fig. 1, the differences between the τ_N^D and histogram with a quantization step equal to 5 for both dimensions. Note that, the histogram requires the selection of one parameter (the quantization step) per dimension, while none is needed for the *discrete global typicality*. For large dimensions (*D*), this can be a big problem. The size of the grid/axis is a user-specified parameter. The histogram takes only values from a finite set $\left\{0;\frac{1}{N};\frac{2}{N};\cdots;1\right\}$, while τ_N^D can take any real value.

The *discrete global typicality* has the following properties:

i) sums up to 1;

ii) the value is within $[0,1]$;

iii) provides a closed analytic form, equation (11);

vi) there is no requirement for *prior* assumptions as well as any user- or problem-specific threshold and parameters;

v) is free from some peculiarities of traditional probability theory (its value never gets >1 and non-zero positive for infeasible values [14], [15]) ;

vi) can be recursively calculated for various types of metrics.

When all the data samples in the dataset have different values ($f_i = 1;\ \ \forall i$), and the histogram quantization step parameter is not properly set, the histogram is unable to show any useful information, while the *discrete global typicality* can still show the mutual distribution information of the dataset, see Fig. 2 (a) and (b). This is a major advantage of *discrete global typicality* because it is parameter free. Here the figures are based on the unique data samples of the same climate dataset. As we can see, the data samples which are closer to the mean of the dataset will have higher value of *global typicality* and vice versa.

It is also interesting to notice that for equally distant data, the *discrete global typicality*, τ_N^D is ***exactly the same*** as the frequentistic form of probability. Then equation (11) reduces to $\tau_N^D(\boldsymbol{u}_i) = f_i \Big/ \sum_{j=1}^{L_N} f_j$. Supplementary Fig. 2 shows a simple example of the *discrete global typicality* τ_N^D and *pmf* of an artificial climate dataset $\{\boldsymbol{x}\}_{50}$ with only data of wind chill, which have *2* unique data samples, $\{\boldsymbol{u}\}_{50} = \{10;20\}$ (^oC), while $\{f\}_{50} = \{20;30\}$.

Obviously, $q_2(\boldsymbol{u}_1) = q_2(\boldsymbol{u}_2) = d^2(\boldsymbol{u}_1, \boldsymbol{u}_2)$, and $\tau_2^D(10^oC) = 0.4$; $\tau_2^D(20^oC) = 0.6$. Indeed, if 20 times we observe wind chill is 10^oC and 30 times 20^oC the likelihood for wind chill of 10^oC will be 40% and for wind chill of 20^oC will be 60%, respectively.

The *discrete global typicality* τ_{100}^D of the outcome of throwing dices for 100 times is presented in Supplementary Fig. 3 as an additional illustrative example. In this experiment, for 1, we can use $[1;0;0;0;0;0;]^{\mathrm{T}}$, for 2, we can use $[0;1;0;0;0;0;]^{\mathrm{T}}$, etc. Let the outcome of throwing dices 100 times be $\{f\}_6 = \left\{\frac{17}{100};\frac{14}{100};\frac{15}{100};\frac{15}{100};\frac{21}{100};\frac{18}{100}\right\}$, the values of the *discrete global typicality* τ_N^D of the six outcomes are equal to their corresponding frequencies, see the Supplementary Fig. 3.

B. Identifying Local Modes of Discrete Global Typicality

In this sub-section, an automatic procedure for identifying all local maxima of the *discrete global typicality*, τ_N^D defined in the previous sub-section will be described. It results in the formation of *data clouds* (samples associated with the local maxima) [19], [26]. Data clouds are free shape while clusters, are usually hyper-spherical, hyper-ellipsoidal. This data partitioning resembles Voronoi tessellation [27]. They are also used in the *AnYa* type neuro-fuzzy predictive [19], [26], classifiers and controllers.

The illustrative figures in this section are based on the same climate dataset [20] that was used earlier in Fig. 1, which has two features/attributes: wind chill (^oC) and wind gust (*mph*). In all cases, the Euclidean distance is used, though, the principle is valid for any metric.

The proposed algorithm can be summarized as follows:

Step 1: Identifying the global maximum of the discrete global typicality τ_N^D

For every unique data sample of the dataset $\{\boldsymbol{x}\}_N$, its *discrete global typicality* $\tau_N^D(\boldsymbol{u}_i)$ ($i = 1,2,...,L_N$) can be calculated using equation (11).

The data sample with the highest τ_N^D is selected as the reference data sample in the ranked collection $\{\boldsymbol{u}^*\}_{L_N}$:

$$\boldsymbol{u}^{*(1)} = \arg\max_{j=1,2,...,L_N}\left(\tau_N^D(\boldsymbol{u}_i)\right) \tag{12}$$

where $\boldsymbol{u}^{*(1)}$ is the data sample with the highest value of *discrete global typicality* (in fact, the global maximum), and we set $\boldsymbol{u}^{*m} \leftarrow \boldsymbol{u}^{*(1)}$. In case when there are more than one maxima, we can start with any one of them.

Step 2: Ranking the discrete global typicality τ_N^D

Then, we find the unique data sample that is nearest to $\boldsymbol{u}^{*m}$ denoted by $\boldsymbol{u}^{*(2)}$ and put it into $\{\boldsymbol{u}^*\}_{L_N}$, meanwhile, remove it from $\{\boldsymbol{u}\}_{L_N}$. $\boldsymbol{u}^{*(2)}$ is set to be the global maximum $\boldsymbol{u}^{*m} \leftarrow \boldsymbol{u}^{*(2)}$.

The ranking operation continues by finding the next data sample, which is closest to $\boldsymbol{u}^{*m}$, putting it into $\{\boldsymbol{u}^*\}_{L_N}$, removing it from $\{\boldsymbol{u}\}_{L_N}$ and setting it as the new global maximum.

By applying the ranking operation until $\{\boldsymbol{u}\}_{L_N}$ becomes empty, we can finally get the ranked unique data samples, denoted as $\{\boldsymbol{u}^*\}_{L_N} = \{\boldsymbol{u}^{*(i)} \mid i = 1,2,...,L_N\}$ and their corresponding ranked *discrete global typicality* collection: $\{\tau_N^D(\boldsymbol{u}^*)\}_N = \{\tau_N^D(\boldsymbol{u}^{*(1)}), \tau_N^D(\boldsymbol{u}^{*(2)}),...,\tau_N^D(\boldsymbol{u}^{*(L_N)})\}$.

Step 3: Identifying all local maxima

The ranked *discrete global typicality* is filtered using equation (13) to detect all local maxima of τ_N^D:

$$IF\left(\tau_N^D(\boldsymbol{u}^{*(j-1)}) > \tau_N^D(\boldsymbol{u}^{*(j)})\right) AND \left(\tau_N^D(\boldsymbol{u}^{*(j)}) > \tau_N^D(\boldsymbol{u}^{*(j+1)})\right)$$
$$THEN\left(\boldsymbol{u}^{*(j)} \text{ is a local maxma of } \tau_N^D\right) \quad (13)$$

We denote the set of the local maxima (can be used as a basis for forming *data clouds* and, further, AnYa type fuzzy rule-based models [19], [26]) of τ_N^D as the set $\{\boldsymbol{u}^{**}\}_{P_N} = \{\boldsymbol{u}^{**(j)} \mid j = 1,2,...,P_N\}$; P_N is the number of the identified local maxima and $P_N \le L_N$.

The ranked *discrete global typicality* is depicted in Fig. 3(a), the corresponding local maxima are depicted in Fig. 3(b).

Step 4: Forming data clouds

Each local maxima, $\boldsymbol{u}^{**(i)}$, is then set as a prototype of a *data cloud*. All other data points are assigned to the nearest prototype (local maximum) forming *data clouds* using equation (14).

$$winning\ label = \arg\min_{j=1,2,...,P_N}\left(d\left(\boldsymbol{x}, \boldsymbol{u}^{**(j)}\right)\right) \quad (14)$$

Data clouds can be used to form *AnYa* models [19], [26]. After all the data samples within $\{\boldsymbol{x}\}_N$ are assigned to the *data clouds*, the center (mean) $\boldsymbol{\mu}_N^j$, the standard deviation σ_N^j and support S_N^j ($j = 1,2,...,P_N$) per *cloud* can be calculated.

Step 5: Selecting the main local maxima of the discrete global typicality, τ_N^D

We then calculate τ_N^D at the *data cloud* centers, denoted by $\{\boldsymbol{\mu}\}_N$ using equation (11) with the corrpesonding supports as their frequencies. Then, we use the following operation to take out the less prominent local maxima.

For each center $\boldsymbol{\mu}_N^i$, we check the condition ($i, j = 1,2,...,P_N$; $i \ne j$):

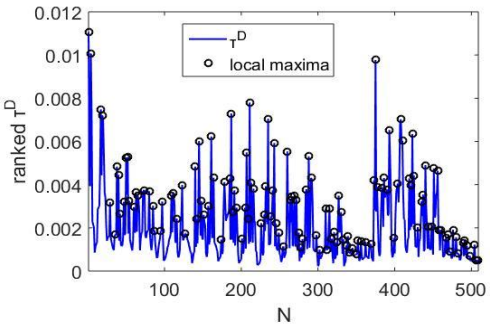
$$IF\left(\left\|\boldsymbol{\mu}_N^i - \boldsymbol{\mu}_N^j\right\| \le 2\sigma_N^i\right) AND \left(\tau_N^D(\boldsymbol{\mu}_N^i) < \tau_N^D(\boldsymbol{\mu}_N^j)\right)$$
$$THEN\left(\{\boldsymbol{\mu}\}_R \leftarrow \boldsymbol{\mu}_N^i\right) \quad (15)$$

This condition means that if there is another center with higher τ_N^D located within the $2\sigma_N^i$ area of $\boldsymbol{\mu}_N^i$, this new more prominent center replaces the existing one. This condition guarantees that the influence areas of neighboring *data clouds* will not overlap significantly (it is well known that according to the Chebyshev inequality for arbitrary distribution the majority of the data samples (>75%) lie within 2σ distance from the mean [1]–[3]).

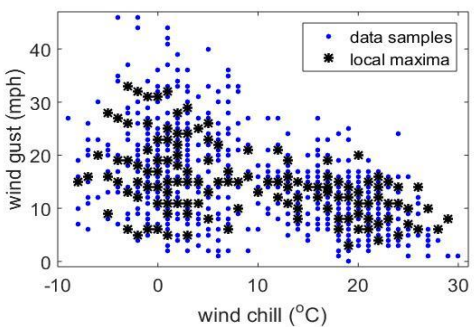
By finding out all the centers satisfying the above condition and assigning them to $\{\boldsymbol{\mu}\}_R$, we get the filtered *data cloud* centers denoted by $\{\boldsymbol{\mu}^*\}_{P_N^*} = \{\boldsymbol{\mu}_N^{*(j)} \mid j = 1,2,...,P_N^*; P_N^* \le P_N\}$ by excluding $\{\boldsymbol{\mu}\}_R$ from $\{\boldsymbol{\mu}\}_{P_N}$ ($\{\boldsymbol{\mu}^*\}_{P_N^*} \cup \{\boldsymbol{\mu}\}_R = \{\boldsymbol{\mu}\}_{P_N}$ and $\{\boldsymbol{\mu}^*\}_{P_N^*} \cap \{\boldsymbol{\mu}\}_R = \varnothing$), where P_N^* is the number of remaining centers.

After that, we set $\{\boldsymbol{u}^{**}\}_{P_N} \leftarrow \{\boldsymbol{\mu}^*\}_{P_N^*}$, $P_N \leftarrow P_N^*$ and repeat ***Steps 4-5*** until the *data cloud* centers do not change any more.

Finally, we can get the composed result, re-named as $\{\boldsymbol{\mu}^o\}$, and use the $\{\boldsymbol{\mu}^o\}$ as the prototypes to



(a) Ranked *discrete global typicality* τ_N^D



(b) Local maxima/peaks/modes of τ_N^D

Fig.3. Identifying local maxima of the *discrete global typicality*, τ_N^D

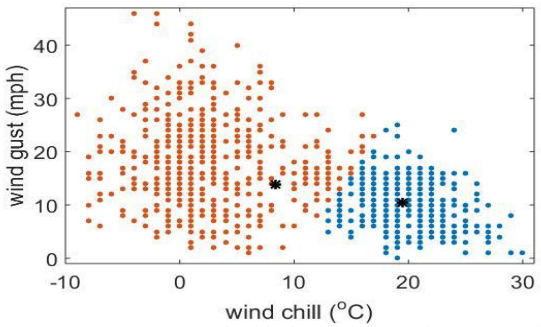


Fig.4. Final filtering result (The black '*' denotes the centers of the data clouds, the data samples from different *data clouds* are plotted with different

build *data clouds* using equation (14).

The final *data cloud* centers for each selection round is presented in the Supplementary Video, which can also be downloadable from: https://www.dropbox.com/s/kkq8xztya3u3kh1/Supplementary_Video.pptx?dl=0.

The final result is presented in Fig. 4. Compared with Fig. 3(b), in the final round, there are only two main modes left broadly corresponding to the two main seasons in Northern England and all the details are filtered out.

Even if $f_1 = 1,\ \forall i$, the *discrete global typicality* can still be extracted successfully from the data samples, despite the fact that the result may not be exactly the same because of the changing data structure, see Supplementary Fig. 4, which uses the same real climate dataset in Fig. 4.

The summary of automatic mode identification algorithm is as follows.

Automatic mode identification algorithm:

i. Calculate $\tau_N^D(\boldsymbol{u}_i)$, $i = 1,2,...,L_N$ using equation (11);

ii. Find the unique data sample $\boldsymbol{u}^{*(1)}$ with global maximum of τ_N^D using equation (12);

iii. Send $\boldsymbol{u}^{*(1)}$ into $\{\boldsymbol{u}^*\}_{L_N}$ and $\tau_N^D(\boldsymbol{u}^{*(1)})$ into $\{\tau_N^D(\boldsymbol{u}^*)\}_{L_N}$ and delete $\boldsymbol{u}^{*(1)}$ from $\{\boldsymbol{u}^*\}_{L_N}$;

iv. $\boldsymbol{u}^{*m} \leftarrow \boldsymbol{u}^{*(1)}$;

v. **While** $\{\boldsymbol{u}\}_{L_N} \neq \varnothing$

* Find the unique data sample(s) which is/are nearest to $\boldsymbol{u}^{*m}$;

* Send the data sample(s) and the corresponding $\tau_N^D(\boldsymbol{u}_i)$ into $\{\boldsymbol{u}^*\}_{L_N}$ and $\{\tau_N^D(\boldsymbol{u}^*)\}_{L_N}$;

* Delete data sample(s) from $\{\boldsymbol{u}\}_{L_N}$;

* Set the latest element in $\{\boldsymbol{u}^*\}_{L_N}$ as $\boldsymbol{u}^{*m}$;

vi. **End While**

vii. Filter $\{\boldsymbol{u}^*\}_{P_N}$ and $\{\tau_N^D(\boldsymbol{u}^*)\}_{P_N}$ using equation (13) and obtain $\{\boldsymbol{u}^{**}\}_{P_N}$ as centers of *data clouds*;

viii. **While** $\{\boldsymbol{u}^{**}\}_{P_N}$ are not fixed

* Use $\{\boldsymbol{u}^{**}\}_{P_N}$ and form the *data clouds* from $\{\boldsymbol{x}\}_N$ using equation (14);

* Obtain the new centers $\{\boldsymbol{\mu}\}_{P_N}$ standard deviations $\{\sigma\}_{P_N}$ and supports $\{S\}_{P_N}$ of the *data clouds;*

* Calculate $\tau_N^D(\boldsymbol{\mu}_N^j)$, $j = 2,...,P_N$ using equation (11);

* Find $\{\boldsymbol{\mu}\}_R$ satisfying equation (15);

* Exclude $\{\boldsymbol{\mu}\}_R$ from $\{\boldsymbol{\mu}\}_{P_N}$ and obtain $\{\boldsymbol{\mu}^*\}_{P_N^*}$;

* $\{\boldsymbol{u}^{**}\}_{P_N} \leftarrow \{\boldsymbol{\mu}^*\}_{P_N^*}$;

* $P_N \leftarrow P_N^*$;

ix. **End While**

x. $\{\boldsymbol{\mu}^o\} \leftarrow \{\boldsymbol{u}^{**}\}_{P_N}$;

ix. Build the data clouds with $\{\boldsymbol{\mu}^o\}$ using equation (14);

C. *Properties of EDA Operators*

Having introduced the basic EDA operators, we will now outline their properties.

- ✓ They are entirely based on the empirically observed experimental data and their mutual distribution in the data space;
- ✓ They do not require any user- or problem-specific thresholds and parameters to be pre-specified;
- ✓ They do not require any model of data generation (random or deterministic), only the type of distance metric used (however, it can be any);
- ✓ The individual data samples (observations) do not need to be independent or identically distributed (*iid*); on the contrary, their mutual dependence is taken into account directly through the mutual distance between the data samples;
- ✓ The method does not require infinite number of observations and can work with just a few exemplars;

Within EDA, we still can consider cross validation and non-parametric statistical tests based on the realizations of experimentally observed data similarly to the significance tests utilized on the random variable assumed in the traditional probability theory and statistics. As a conclusion, EDA can be seen as an advanced data analysis framework which can work efficiently with any feasible data and any type of distance or similarity metric.

IV. Theoretical Basis - Continuous Density and Typicality

Up to this point, all EDA definitions are useful to describe data sets or data streams made up of a discrete number of observations. However, they cannot be used for inference because they are only defined on points where samples occur (discrete spaces). In this section, we define the *continuous local and global density* and *global typicality* which can be

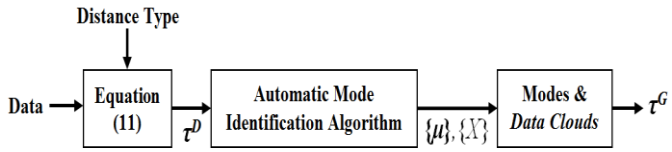


Fig.5. The process of extracting distribution from data in EDA

used for inference on the continuous domain of the variable $\boldsymbol{x}$. At this stage, we depart from the entirely data based and assumptions-free approach we used so far, however, this is done after we identified the local modes, formed *data clouds* around these focal points and obtained the support of these *data clouds.* Therefore, the extension to the continuous domain is inherently local (per data cloud). We assume that the local mode considered as the mean and the support considered as frequency plus the deviation of the empirical data do provide the triplet of parameters ($\boldsymbol{\mu}$, X, N_i). We do recognize that these triplets are conditional on the specific N_i data samples observed and associated with the particular *data cloud*, but this will be updated when new data is available. Now, having this triplet of parameters we, firstly, define the *continuous local density*, D^L as:

$$D_N(\boldsymbol{x}_i)=\frac{\sum_{j=1}^{N} q_N(\boldsymbol{x}_j)}{2Nq_N(\boldsymbol{x}_i)};\quad L_N>1 \tag{15}$$

Like equation (9), for the case of Euclidean distance, the *continuous local density*, D^L is simplified to a continuous Cauchy type function over any feasible value of the variable $\boldsymbol{x}$ with the parameters $\boldsymbol{\mu}$ and X extracted from N available data samples as described earlier:

$$D^L_{N,i}(\boldsymbol{x})=\frac{1}{1+\frac{\|\boldsymbol{x}-\boldsymbol{\mu}_{N,i}\|^2}{\sigma^2_{N,i}}};\quad i=1,2,...,C_N;\quad L_N>1 \tag{16}$$

where $\sigma^2_{N,i}=X_{N,i}-\|\boldsymbol{\mu}_{N,i}\|^2$; $\boldsymbol{\mu}_{N,i}$ and $X_{N,i}$ are the mean and the average value of scalar products of the data samples within the i^{th} data cloud; C_N is the number of data clouds; the subscript N means the local densities are derived from N observed data samples. It is obvious that with more data samples observed, the parameters will change and have to be updated regularly. Note that equation (16) is defined based on Euclidean distance. The expression of *continuous local density* D^L varies from the type of distance used. Nonetheless, in general, the *continuous local density* of the data can be expressed in the same form as the *discrete local density* but in the continuous space.

The *continuous local density* D^L is defined on the continuous space for each local maximum per *data cloud.* Furthermore, we introduce the *continuous global density* D^G as a weighted sum of the *local density* of each *data cloud* with weights being the support (number of data samples) of the respective *data cloud*. Finally, we introduce the *continuous global typicality* τ^G based on D^G. The *continuous global density and typicality* play a similar role to the mixture of *pdfs.* However, the questions "how many distributions in the mixture", "which are their parameters" and "what type of distributions" see Fig. 5 are all answered from the data directly, free from any user or problem-specific pre-defined parameters, *prior* assumptions, knowledge or pre-processing techniques like the cases of clustering, EM, etc.

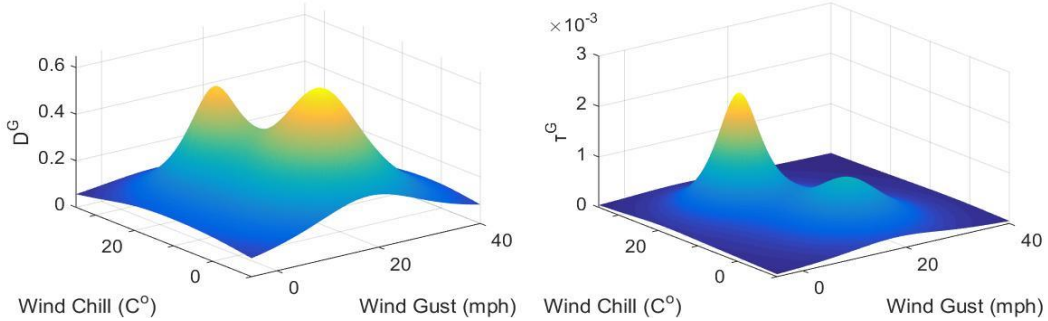
A. *Continuous Global Density*

Continuous global density is a mixture that arises simply from the metric of the space used to measure sample distance and the density of samples that exist in the space. However, it works for all types of distance/similarity metric. As we can see from equation (16) the local density is Cauchy type when the Euclidean distance is employed therefore, the simplest of the procedures is to define the *continuous global density* as a mixture of Cauchy distributions. The *continuous global density* enables inference of new samples anywhere in the space.

For any $\boldsymbol{x}$ and any type of distance used, we define *continuous global density* in a general form very much like the mixture distributions, as a weighted combination of *continuous local densities*:

$$D^G_N(\boldsymbol{x})=\frac{\sum_{i=1}^{C_N} S_{N,i}D^L_{N,i}(\boldsymbol{x})}{N};\quad L_N>1 \tag{17}$$

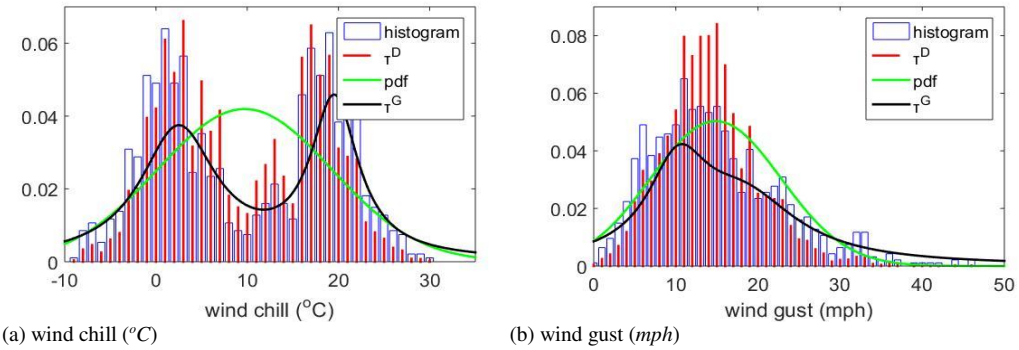
where $D^L_{N,i}(\boldsymbol{x})$ is the *local density* of $\boldsymbol{x}$ in the i^{th} *data cloud*; C_N is the number of data clouds at the N^{th} time instance; $S_{N,i}$ is the support (number of members) of the i^{th} *data cloud* based on the available experimental/actual data. For normalization, we impose the condition $\sum_{i=1}^{C_N} S_{N,i}=N$. The *continuous global density* D^G is defined non-parametrically from each of the modes of the data (D^L) and near the peaks; it is a very good approximation of D^L, but it will deviate progressively from it in trough regions. As an example, the *global density* for the same climate dataset used before [20] is presented in Fig. 6(a).



(a) *Continuous global density* (b) *Continuous global typicality*

Fig.6. *Continuous global density* and *global typicality* of the real climate dataset [20] using Euclidean type distance.

Compared with the *discrete local density* introduced in section II which is discrete and unimodal by definition, D^G is more effective to detect the natural multimodal data structure such as abnormal data samples because only the data samples that are close to the larger *data clouds*, which can be viewed as the main modes of the data patterns, can have higher values of *continuous global density*. This feature is clearly depicted by the value of D^G of those data samples located in the space between the two main modes in the figures below, while for the *local density*, see Supplementary Fig. 1(c), it is exactly the opposite case.



(a) wind chill (^oC) (b) wind gust (*mph*)

Fig.7. Comparison between the *continuous global typicality* τ^G, *discrete global typicality* τ^D, histogram and traditional *pdf*.

B. Continuous Global Typicality

Having introduced *the continuous global density*, we can also define the *continuous global typicality*, τ^G as well. It is also defined as a normalized form of the density (similarly to the weighted typicality, τ^D, equation (11)) but with the use of integral instead of the sum. As stated in section II, the *weighted typicality,* τ^D is discrete and sums to 1. The *global typicality* is expressed as follows:

$$\tau_N^G(\boldsymbol{x}) = \frac{D_N^G(\boldsymbol{x})}{\int_{-\infty}^{\infty} D_N^G(\boldsymbol{x})d\boldsymbol{x}} = \frac{\sum_{i=1}^{C_N} S_{N,i} D_{N,i}^L(\boldsymbol{x})}{\sum_{i=1}^{C_N} S_{N,i} \int_{-\infty}^{\infty} D_{N,i}^L(\boldsymbol{x})d\boldsymbol{x}} \tag{18}$$

It is important to notice that equation (18) is general and valid for any type of distance/similarity metric. For a general multivariate case, it is important to normalize the mixture of *continuous local densities* $D_{N,i}^L(\boldsymbol{x})$ to make τ^G ***integrate to* 1**. By finding out the integral of the *continuous global density* within the metric space and dividing τ^G by its integral, one can always guarantee unit integral, regardless the type of distance/similarity metric used.

As we said before, we consider the well-known expression of the multivariate Cauchy distribution [21]–[23] to transform the $D_{N,i}^L(\boldsymbol{x})$ without loss of generality.

$$f(\boldsymbol{x}) = \frac{\Gamma\left(\frac{K+1}{2}\right)}{\pi^{\frac{d+1}{2}} \sigma^K \left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu})}{\sigma^2}\right)^{\frac{K+1}{2}}} \tag{19}$$

where $\boldsymbol{x} = [x_1, x_2, ..., x_K]^{\mathrm{T}}$; π is the well-known mathematical constant and $\Gamma(\cdot)$ is the gamma function; $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{x}]$; σ is scalar parameter. This guarantees that:

$$\int_{x_K} \cdots \int_{x_2} \int_{x_1} f(x_1, x_2, ..., x_K) dx_1 dx_2, ...dx_K = 1 \tag{20}$$

Based on (17)-(19), we introduce the *normalized continuous local density* as follows:

$$\bar{D}_{N,i}^L(\boldsymbol{x}) = \frac{\Gamma\left(\frac{K+1}{2}\right)}{\pi^{\frac{K+1}{2}} \sigma_{N,i}^K} \left(D_{N,i}^L(\boldsymbol{x})\right)^{\frac{K+1}{2}} \tag{21}$$

Here $\sigma_{N,i} = \sqrt{X_{N,i} - \boldsymbol{\mu}_{N,i}^{\mathrm{T}} \boldsymbol{\mu}_{N,i}}$ for the Euclidean distance.

We can, finally, get the expression of the *continuous global typicality*, τ^G in terms of the normalized *continuous global density* as:

$$\tau_N^G(\boldsymbol{x}) = \frac{\sum_{i=1}^{C_N} S_{N,i} \bar{D}_{N,i}^L(\boldsymbol{x})}{\sum_{i=1}^{C_N} S_{N,i} \int_{-\infty}^{\infty} \bar{D}_{N,i}^L(\boldsymbol{x})d\boldsymbol{x}} = \frac{\Gamma\left(\frac{K+1}{2}\right)}{\pi^{\frac{K+1}{2}} N} \sum_{i=1}^{C_N} S_{N,i} \frac{\left(D_{N,i}^L(\boldsymbol{x})\right)^{\frac{K+1}{2}}}{\sigma_{N,i}^K} \tag{22}$$

For the Euclidean distance, equation (22) becomes

$$\tau_N^G(\boldsymbol{x}) = \frac{\Gamma\left(\frac{K+1}{2}\right)}{\pi^{\frac{K+1}{2}} N} \sum_{i=1}^{C_N} \frac{S_{N,i}}{\sigma_{N,i}^K \left(1 + \frac{\|\boldsymbol{x}-\boldsymbol{\mu}_{N,i}\|^2}{\sigma_{N,i}^2}\right)^{\frac{K+1}{2}}} \tag{23}$$

The *continuous global typicality* of the real climate dataset with Euclidean distance is presented in Fig.6(b)

The comparisons between the continuous *global typicality* (the modes are extracted by the approach introduced in section III*), discrete global typicality*, histogram and traditional *pdf* are presented in *2D* form for visual clarity in Fig. 7 using the same the real climate dataset [20].

As shown in Fig. 7, compared with the traditional *pdf* using a Gaussian model*, the global typicality* derived directly from the dataset without any *prior* assumption about the number of local modes or type of distribution represents very well the two modes in the data pattern and gives results very close to what a histogram would give and significnatly better to what a single unimodal distribution would provide.

In summary, the proposed *continuous global typicality* has the following properties, many of which it shares with the *discrete global typicality* introduced in section III*:*

i) integrates to 1;
ii) provides a closed analytic form;
iii) no requirement for *prior* assumptions as well as any user or problem-specific threshold and parameters; these are derived from the data entirely;
vi) can be recursively calculated for various types of metrics.

V. Applications

A. Examples

In this subsection, we will give several examples of the *continuous global typicality,* τ^G of different datasets extracted by the proposed automatic mode identification algorithm. The *continuous global typicality* of the Seeds dataset [28] and Combined Cycle Power Plant dataset [29] and Wine Quality dataset [30] with Euclidean distance is presented in Fig. 8. As the dimensionality of the original datasets is > 2, for a better visualization, we use the principal component analysis (PCA) method [31] to reduce the dimensionality and use the first 2 principal components in the figures as the x-axis and y-axis. Supplementary Fig. 5 (a) and (b) present the τ^G derived from the first 1/3 and the first 2/3 the Wine Quality dataset. Supplementary Fig. 5 (c) depicts the τ^G derived by scrambling the order of the data samples. The *continuous global typicality* τ^G of 2 dimensional benchmark datasets A1, S1 and S2 [32] are also presented in Supplementary Fig. 6.

If we want more details from the *continuous global typicality*, we can also stop the automatic mode identification algorithm described in section III early, i.e. before the final iteration, and build the *continuous global typicality* based on more detailed data partitioning results. The Supplementary Video referred in section III.B also depicts evolution of the *global continuous typicality* based on the results of different iteration times of the proposed mode identification algorithm.

B. Inference Primer

Assuming, there are 3 arbitrary non-integer values of wind chill data $x=\{-7.5;2.5;14.7\}$ (^oC), which does not exist in the dataset, we can quickly obtain the corresponding *continuous global typicality* using equation (18), $\{\tau^G(x)\} = \{0.0080, 0.0375, 0.0180\}$ and the inferences made are presented in Fig. 9. Here we only consider the two main modes. That means that wind chill of -7.5^oC is less likely while the wind chill of 2.5^oC is more likely.

In addition, if we want to know the *continuous global typicality* of all the values larger than $\boldsymbol{t}$, we can integrate as follows:

$$\mathrm{T}(\boldsymbol{x}\geq\boldsymbol{t})=1-\int_{x=t}^{\infty}\tau_N^G(\boldsymbol{x})d\boldsymbol{x} \tag{24}$$

For example, when Euclidean distance is used, and here we only consider one-dimensional data for simpler derivation, equation (24) can be re-written as:

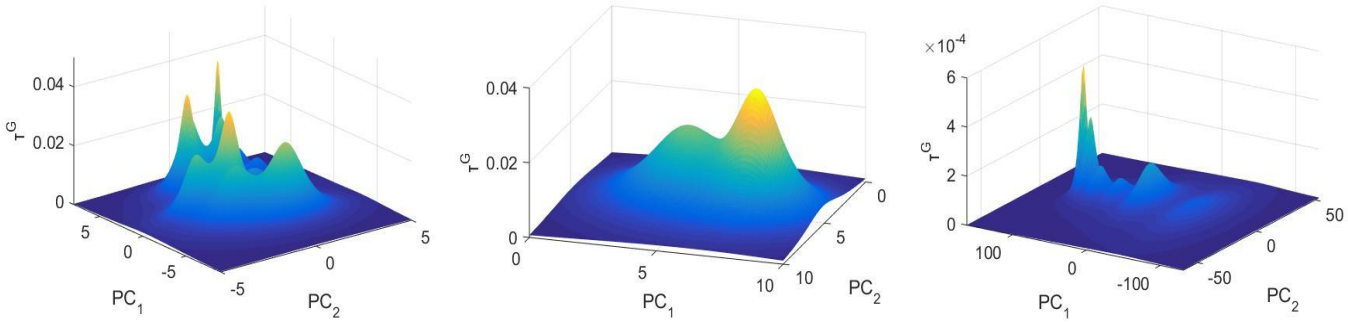
$$\mathrm{T}(x\geq t)=1-\int_{x=t}^{+\infty}\frac{\sum_{i=1}^{C_N}S_{N,i}\bar{D}_{N,i}^L(x)}{N}dx$$
$$=1-\frac{\sum_{i=1}^{C_N}S_{N,i}\left(\frac{1}{\pi}\arctan\left(\frac{t-\mu_{N,i}}{\sigma_{N,i}}\right)+\frac{1}{2}\right)}{N} \tag{25}$$

Let us continue the example in Fig. 9. If we want to know the *global continuous typicality* of all the data samples above *20 ^oC*, which is the green area of this figure, we can calculate the value using equation (25) to yield $\mathrm{T}(x>20)=0.2447$. That means that the likelihood, a value to be equal to or greater than 20 ^oC is 24.47%. One can see that the *continuous global typicality* can serve as a form of probability.

C. Naïve EDA Classifier

In this sub-section, we borrow the concept of naïve Bayes classifiers [1]–[3] and propose a new version of naïve EDA classifier. In contrast with the original naïve EDA classifier proposed in [15], which relies for inference on the *discrete global typicality* and linear interpolation and/or extrapolation, the naïve EDA classifier in this paper uses the *continuous global typicality* instead, which is based on the local modes of the *discrete global typicality* identified by an automatic procedure as described in section III.B. This procedure is more effective in reflecting the ensemble features of the distribution of the data samples of different classes in the data space.

As the proposed approach accommodates various type of distance/similarity metrics, one can use the current knowledge in the area to choose the desired distance measure for a reasonable approximation that simplifies the processing. Moreover, one can change to other distance measures easily



(a) Seeds dataset (b) Combined Cycle Power Plant dataset (c) Wine Quality dataset

Fig.8. *Continuous global typicality* of the Seeds dataset [28], Combined Cycle Power Plant dataset [29] and Wine Quality dataset [30]

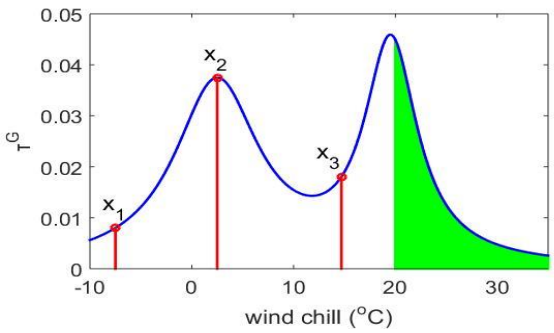


Fig.9. *Continuous global typicality* τ^G of wind chill data and simple inferences

and compare the results obtained by the classifier with different type of measures. For consistence, in the following numerical examples, we use the Euclidean distance.

Let us assume H classes at the N^{th} time instance, where some classes may have many *data clouds*. The *continuous global typicality* per class can be defined as ($i = 1, 2, ..., H$):

$$\tau_{N,i}^{G}(\boldsymbol{x}) = \frac{\sum_{j=1}^{W_j} S_{N,i,j} D_{N,i,j}^{L}(\boldsymbol{x})}{\sum_{j=1}^{W_j} S_{N,i,j} \int_{-\infty}^{\infty} D_{N,i,j}^{L}(\boldsymbol{x}) d\boldsymbol{x}} \tag{26}$$

where, W_j is the number of *data clouds* sharing the same i^{th} class label, $\sum_{i=1}^{H} W_j = C_N$; $S_{N,i,j}$ is the support of the j^{th} *data cloud* having the i^{th} class label; $D_{N,i,j}^{L}(\boldsymbol{x})$ is the corresponding *continuous local density*.

For any unlabeled data sample $\boldsymbol{x}$, its label is decided by the following expression:

$$label(\boldsymbol{x}) = \arg\max_{j=1,2,...,H} \left(\tau_{N,j}^{G}(\boldsymbol{x}) \right) \tag{27}$$

The *2D* plots (wind chill and wind gust) of the *continuous global typicality* with Euclidean type of distance of the real climate dataset are given in Supplementary Fig.7.

The performance of the proposed naïve EDA classifier is further tested on the following problems:

i) Banknote Authentication dataset [33];
ii) Pima dataset [34];
iii) Climate dataset [20];
iv) Pen-Based Handwritten Digits Recognition dataset [35];
v) Madelon dataset [36];
vi) Optical Handwritten Digits Recognition dataset [37];
vii) Occupancy Detection dataset [38].

The proposed naïve EDA classifier is compared with a SVM classifier with Gaussian radial basis function and a naïve Bayes classifier in terms of their performance. The details of the datasets used in the classification are demonstrated in Supplementary Section B.

In the experiments, PCA [31] is applied as a pre-processing step to reduce the dimensionality and balance the variances of the datasets. It has to be stressed that PCA is not a part of the proposed method and is not necessary for simpler problems.

For Banknote Authentication, Pima and Climate datasets, we randomly select 70% of the data for training and use the rest for validation. The performance is evaluated after 10-fold cross-validation. For Pen-Based Digits, Madelon, Optical Digits and Occupancy Detection datasets, we train the classifiers with the training sets and conduct the validation with the testing/validation sets.

The overall performance of the 3 classifiers is tabulated in Table I, where we consider the first 3 principal components for classification. Considering the first 5 principal components, the overall results obtained by the classifiers are tabulated in Table II.

As it is shown in Tables I and II, the proposed naïve EDA classifier outperforms the SVM classifier and naïve Bayes classifier on different problems in the majority of the numerical examples. The performance of the proposed naïve EDA classifier is the best. In addition, it is worth to note that the classification conducted by the naïve EDA classifier is totally free from unrealistic assumptions, restrictions or *prior* knowledge.

TABLE I
CLASSIFICATION PERFORMANCE- 3 PRINCIPAL COMPONENTS CONSIDERED

Dataset	Overall Accuracy		
	Naïve EDA classifier	SVM classifier	Naïve Bayes classifier
Banknote	0.9910	**0.9978**	0.9629
Pima	**0.7374**	0.6487	0.7343
Climate	**0.9777**	0.6365	0.9709
Pendigit	**0.8070**	0.2247	0.7424
Madelon	**0.6167**	0.5000	0.6083
Optdigit	0.7084	0.5442	**0.7218**
Occupancy detection testing set 1	**0.9700**	0.6735	0.9377
Occupancy detection testing set 2	**0.9532**	0.8168	0.8676

VI. CONCLUSION AND FUTURE DIRECTION

In this paper, we propose a new systematic approach to derive ensemble properties of data without any *prior* assumptions about data sources, amount of data and user- or problem- specific parameters. The EDA (Empirical Data Analytics) framework considers the relative position of data in a metric space only and extracts from the raw experimental discrete observations a series of measures of their ensemble properties, such as the *cumulative proximity* (q), *centrality* (C), *square centrality* (q^{-1}), *standardized eccentricity* (ε), *density* (D) as well as ***typicality***, (τ). The local and global versions of

TABLE II
CLASSIFICATION PERFORMANCE - 5 PRINCIPAL COMPONENTS CONSIDERED

Dataset	Overall Accuracy		
	Naïve EDA classifier	SVM classifier	Naïve Bayes classifier
Pima	**0.7391**	0.6522	0.7365
Climate	**0.9734**	0.5170	0.9578
Pendigit	**0.8190**	0.1072	0.7730
Madelon	**0.6117**	0.5000	0.5817
Optdigit	**0.8603**	0.1708	0.8436
Occupancy detection testing set 1	**0.9726**	0.6353	0.9422
Occupancy detection testing set 2	**0.9647**	0.7899	0.8654

the ***typicality***, (τ and τ^G) are both considered originally in discrete form and then in continuous form approximating the actual data-driven discrete estimators by a mixture of local functions. It was demonstrated that for the case when the distance metric used is Euclidean, the density (both in its discrete form that is exactly describing the actual data and in its continuous form which is approximating the entire data space density) takes the form of a Cauchy function. However, importantly, this is not an assumption made a *priori*, but is driven and parameterized by the data and the selected metric. Furthermore, we propose an autonomous algorithm for identifying all local modes/maxima of the *global discrete typicality*, τ^D as well as for filtering out the main local maxima based on the 2σ closeness of each local maximum. Finally, we present a number of numerical examples aiming to verify the methodology and demonstrate its advantages. We introduce a new type of classifier, which we call naïve EDA for investigating the unknown data pattern behind the large amount of data in a data-rich environment. In conclusion, the proposed EDA framework and methodology provides an efficient alternative that is entirely based on the experimental data and the evidence. It touches the very foundations of data mining and analysis and, thus, has a wide area of applications, especially, in the era of big data and data streams where handcrafting offline methods and making detailed assumptions is often not an option.

Nonetheless, we have to admit that the bottlenecks of the proposed methodology are the lack of theoretical confidence levels for the analysis and the theoretical idea of reliability and generalization, which are the inherited limitations of nonparametric approaches.

In this paper, we only provide the preliminary algorithms and results on data partitioning, analysis, inference and classification. As a future work, we will focus on developing more advanced algorithms within the EDA framework for various applications of different areas, including, but not limited to, high frequency trading data processing, foreign currency trading problem, handwritten digits recognition, remote sensing, etc.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. Burlin: Springer, 2009.

[2] C. M. Bishop, *Pattern recognition*. New York: Springer, 2006.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. Chichester, West Sussex, UK,: Wiley-Interscience, 2000.

[4] T. Bayes, "An essay towards solving a problem in the doctrine of chances," *Philos. Trans. R. Soc.*, vol. 53, p. 370, 1763.

[5] J. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer, 2010.

[6] C. Spearman, "The proof and measurement of association between two things," *Am. J. Psychol.*, vol. 15, pp. 72–101, 1904.

[7] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1, pp. 81–93, 1938.

[8] L. A. . Goodman and W. H. . Kruskal, "Measures of association for cross classifications," *J. Am. Stat. Assoc.*, vol. 49, no. 268, pp. 732–764, 1954.

[9] P. Del Moral, "Nonlinear filtering: interacting particle resolution," *Comptes Rendus l'Académie des Sci. - Ser. I - Math.*, vol. 325, no. 6, pp. 653–658, 1997.

[10] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[11] M. Chen and D. A. Linkens, "Rule-base self-generation and simplification for data-driven fuzzy models," *Fuzzy Sets Syst.*, vol. 142, no. 2, pp. 243–265, 2004.

[12] P. P. Angelov, "Anomaly detection based on eccentricity analysis," in *2014 IEEE Symposium Series in Computational Intelligence, IEEE Symposium on Evolving and Autonomous Learning Systems, EALS, SSCI 2014*, 2014, pp. 1–8.

[13] P. Angelov, "Outside the box: an alternative data analytics framework," *J. Autom. Mob. Robot. Intell. Syst.*, vol. 8, no. 2, pp. 53–59, 2014.

[14] P. Angelov, X. Gu, and D. Kangin, "Empirical data analytics," *Int. J. Intell. Syst.*, DOI 10.1002/int.21899, 2017.

[15] P. P. Angelov, X. Gu, J. Principe, and D. Kangin, "Empirical data analysis - a new tool for data analytics," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2016, pp. 53–59.

[16] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[17] L. C. Freeman, "Centrality in social networks conceptual clarification," *Soc. Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[18] J. G. Saw, M. C. K. Yang, and T. S. E. C. Mo, "Chebyshev inequality with estimated mean and variance," *Am. Stat.*, vol. 38, no. 2, pp. 130–132, 1984.

[19] P. Angelov, *Autonomous learning systems: from data streams to knowledge in real time*. John Wiley & Sons, Ltd., 2012.

[20] "Climate Dataset in Manchester," *http://www.worldweatheronline.com*.

[21] S. Nadarajah and S. Kotz, "Probability integrals of the multivariate t distribution," *Can. Appl. Math. Q.*, vol. 13, no. 1, pp. 53–84, 2005.

[22] C. Lee, "Fast simulated annealing with a multivariate Cauchy distribution and the configuration's initial temperature," *J. Korean Phys. Soc.*, vol. 66, no. 10, pp. 1457–1466, 2015.

[23] S. Y. Shatskikha, "Multivariate Cauchy distributions as locally Gaussian distributions," *J. Math. Sci.*, vol. 78, no. 1, pp. 102–108, 1996.

[24] A. Corduneanu and C. M. Bishop, "Variational Bayesian model selection for mixture distributions," *Proc. Eighth Int. Conf. Artif. Intell. Stat.*, pp. 27–34, 2001.

[25] P. P. Angelov and X. Gu, "Empirical fuzzy sets," *under review*, 2017.

[26] P. Angelov and R. Yager, "A new type of simplified fuzzy rule-based system," *Int. J. Gen. Syst.*, vol. 41, no. 2, pp. 163–185, 2011.

[27] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial tessellations: concepts and applications of Voronoi diagrams*, 2nd ed. Chichester, England: John Wiley & Sons., 1999.

[28] "Seeds Dataset," *https://archive.ics.uci.edu/ml/datasets/seeds*.

[29] "Combined Cycle Power Plant Dataset," *http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant*.

[30] "Wine Quality Dataset," *https://archive.ics.uci.edu/ml/datasets/Wine+Quality*.

[31] I. Jolliffe, *Principal component analysis*. John Wiley & Sons, Ltd., 2002.

[32] "Clustering datasets," *http://cs.joensuu.fi/sipu/datasets/*.

[33] "Banknote Authentication Dataset," *https://archive.ics.uci.edu/ml/datasets/banknote+authentication*.

[34] "Pima Indians Diabetes Dataset," *https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes*.

[35] "Pen-Based Recognition of Handwritten Digits Dataset," *http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits*.

[36] "Madelon Dataset," *http://archive.ics.uci.edu/ml/datasets/Madelon*.

[37] "Optical Recognition of Handwritten Digits Dataset," *https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits*.

[38] "Occupancy Detection Dataset," *https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+*.

Plamen P. Angelov (F'16, SM'04, M'99) is a Chair Professor in Intelligent Systems with the School of Computing and Communications, Lancaster University, UK. He obtained his PhD (1993) and his DSc (2015) from the Bulgarian Academy of Science. He is the Vice President of the International Neural Networks Society and a member of the Board of Governors of the Systems, Man and Cybernetics Society of the IEEE, a Distinguished Lecturer of IEEE. He is Editor-in-Chief of the Evolving Systems journal (Springer) and Associate Editor of IEEE Transactions on Fuzzy Systems as well as of IEEE Transactions on Cybernetics and several other journals. He

received various awards and is internationally recognized pioneering results into on-line and evolving methodologies and algorithms for knowledge extraction in the form of human-intelligible fuzzy rule-based systems and autonomous machine learning. He holds a wide portfolio of research projects and leads the Data Science group at Lancaster.

Xiaowei Gu received the B.E. and M.E. degrees from the Hangzhou Dianzi University, Hangzhou, China. He is currently pursuing the Ph.D. degree in computer science with Lancaster University, UK.

José C. Príncipe (F'00) is a Distinguished Professor of Electrical and Computer Engineering at the University of Florida. He is also the Eckis Professor and Founding Director of Computational NeuroEngineering Laboratory (CNEL), University of Florida. His primary research interests are advanced signal processing with information theoretic criteria (entropy and mutual information), adaptive models in the reproducing kernel Hilbert spaces (RKHS) and the application of these advanced algorithms in Brain Machine Interfaces (BMI). Dr. Principe is a Fellow of the IEEE, ABME and AIBME. He is the past Editor-in-Chief of the IEEE Transactions on Biomedical Engineering, past Chair of the Technical Committee on Neural Networks of the IEEE Signal Processing Society and past President of the International Neural Network Society. He received the IEEE EMBS Career Award, and the IEEE Neural Network Pioneer Award.