An Introduction to Information Theoretic Learning, Part II: Applications

Daniel G. Silva, Denis G. Fantinato, Jânio C. Canuto, Leonardo T. Duarte, Aline O. Neves, Ricardo Suyama, Jugurta Montalvão and Romis Attux

***Abstract*—This is the second part of the introductory tutorial about information theoretic learning, which, after the theoretical foundations presented in Part I, now discusses the concepts of correntropy, a new similarity measure derived from the quadratic entropy, and presents example problems where the ITL framework can be successfully applied: dynamic modelling, equalization, independent component analysis and cluster analysis.**

***Index Terms*—ITL, information theory, entropy, correntropy.**

I. Introduction

In this second half of the two-part tutorial on information theoretic learning, we analyze a number of key ITL criteria and methods and also discuss their use on some important applications. The discussion starts from the notion of correntropy, an interesting higher-order statistical extension of the classical concept of correlation, and also includes a brief survey of pertinent works and case studies. In the following, a representative selection of supervised and unsupervised applications is presented, and is used as a background for the exposition of important methods, like those based on error entropy, unsupervised kernel criteria and independence. It is our belief that this presentation will provide the reader with a complete view on canonical ITL strategies and on its potentialities.

The work is structured as follows: Section II brings the definition of correntropy and discusses several instances of application; Section III analyzes the use of ITL methods in dynamic modeling, supervised and unsupervised equalization, independent component analysis — including recent formulations concerning finite fields and clustering; finally, Section IV summarizes our conclusions and final remarks.

II. Correntropy

The great majority of classical adaptive equalization methods and also of those based on ITL consider that the available data is independently distributed, which, in many cases, is not true. Thus, Santamaria et al. [1] proposed a new measure that is able to take into account both the statistical and temporal structures of the signals. The proposed generalized correlation function was termed *correntropy*, since it is directly related to Rényi's quadratic entropy estimated using the Parzen window (see Section IV.B of Part I). Mathematically, correntropy is defined as

$$v(t_1, t_2) = E[K(x_{t_1}, x_{t_2})], \tag{1}$$

where x_t is a stochastic process and $K(\cdot)$ is a kernel function. Likewise other ITL measures (see Part I), the Gaussian function is usually employed as the kernel:

$$v(t_1, t_2) = E[G(x_{t_1}|x_{t_2}, \sigma^2)]. \tag{2}$$

In this case, for a non-zero lag, the value of correntropy asymptotically tends to the information potential [1], [2].

Correntropy can also be straightforward redefined for a pair of random variables, X and Y

$$v(X, Y) = E[G(X|Y, \sigma^2)], \tag{3}$$

which is formally denoted as the cross-correntropy between X and Y.

There are two main interpretations for correntropy. The first one associates it with a feature space interpretation that relates nonlinearly with the input space, hence, using correntropy as a Parzen kernel is equivalent to having a linear kernel in a high-dimensional space (Hilbert Space) with reproducing properties [3]. The second interpretation is that it is the integral over the line $x_{t_1} = x_{t_2}$ of the joint pdf estimated with Parzen window, which powerfully indicates that correntropy can be viewed as a measure of probability that the random variables x_{t_1} and x_{t_2} are equal. Such view supports the notion of correntropy as a generalized similarity measure.

Using a series expansion for the Gaussian kernel, equation (2) may be rewritten as:

$$v(t_1, t_2) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n \sigma^{2n} n!} E\left[\|x_{t_1} - x_{t_2}\|^{2n}\right] \tag{4}$$

which involves all the even-order moments of the random variable $\|x_{t_1} - x_{t_2}\|$. Through (4), it is possible to see that this new measure includes the information provided by the conventional covariance function. Furthermore, the authors of [1] demonstrate that, in order to obtain the property $v(t, t-\tau) = v(\tau)$, the stochastic process must be strictly

Daniel G. Silva is with the Dep. of Electrical Engineering, University of Brasília - UnB, Brasília, DF, Brazil. e-mail: danielgs@ene.unb.br.

Denis G. Fantinato and Romis Attux are with the Lab. of Signal Processing for Communications - DSPCom, University of Campinas - Unicamp, Campinas, SP, Brazil. e-mail: {denisgf, attux}@dca.fee.unicamp.br

Leonardo T. Duarte is with the School of Applied Sciences (FCA), University of Campinas - Unicamp, Limeira, SP, Brazil. e-mail: leonardo.duarte@fca.unicamp.br

Jânio C. Canuto and Jugurta Montalvão are with the Dep. of Electrical Engineering, Federal University of Sergipe, São Cristóvão, SE, Brazil. e-mail: janio.canuto@gmail.com, jmontalvao@ufs.br.

Ricardo Suyama and Aline O. Neves are with the Engineering, Modeling and Applied Social Sciences Center, Federal University of ABC, Santo André, SP, Brazil. e-mail: {ricardo.suyama, aline.neves}@ufabc.edu.br.

Digital Object Identifier: 10.14209/jcis.2016.7

stationary. In this case, the definition presented in (2) can be estimated in terms of a sample mean:

$$\hat{v}(m) = \frac{1}{N-m+1} \sum_{n=m}^{N} G(x(n)|x(n-m), \sigma^2) \quad (5)$$

where N is the number of available signal samples.

A careful analysis of (4) was considered in [4], where the authors show that the series may diverge depending on the distribution of the signal being considered. However, for shorter-tailed distributions such as the uniform, it is also possible to derive certain conditions for which the series converges [4]. Nevertheless, it is not necessary that the series exist in order that correntropy exist. As demonstrated in an equalization scenario, correntropy performs well even when the series diverges.

Moreover, the choice of the kernel size σ is crucial. If its value is too large, correntropy will basically rely on second-order properties. On the other hand, if the value is too small, an undesirable behavior can be observed, in which the correntropy is dominated by moments of extremely high-order [4]. In that sense, σ plays a different role in correntropy, being related to weights on the statistical moments, while, in the information potential, σ is closely related to the shape of the distributions. It is also worth mentioning that, although correntropy asymptotically converges to information potential when Gaussian kernels are used, the computational complexity of correntropy is one order below the cost associated with information potential, as it encompasses a single sum operator over the kernel argument.

Comparing correntropy to the conventional autocorrelation function, it is possible to observe that the mean value of the former changes for different source distributions, whereas the autocorrelation function remains basically the same. This characteristic may be useful in eliminating the bias on the estimation of entropy from finite data sets using the Parzen window method (recall Section IV.B of Part I).

If correntropy is used to design an optimal equalizer for a digital communication system, it is possible to show that the performance will be better than that of a Mean Square Error (MSE)-based receiver if the noise PDF has its global maximum at the origin [5]. Furthermore, it presents great robustness to impulsive noise. On the other hand, MSE-based estimation is biased if the noise PDF has non-zero mean, which leads to performance degradation when in the presence of impulsive noise. Thus, correntropy may be very useful in nonlinear and non-Gaussian signal processing [5]. Such characteristic also comes from the fact that correntropy may be viewed as a localized similarity measure, related to the probability of how similar two variables are in a neighborhood of the joint space controlled by the kernel size.

In the last few years, correntropy has been used successfully in a large variety of applications when compared to classical techniques. In [6], correntropy is used in a supervised scenario with impulsive noise, outperforming LMS in system identification and noise cancellation. In [1], [7], correntropy is used to perform blind equalization (which is discussed in detail in Section III-C), outperforming classical methods like the CMA [8] in the case of correlated sources. Nonetheless, CMA may have a better performance if the sources are iid [7], [9]. In [10], correntropy is used as a unifying instantaneous blind source separation criterion, capable of separating iid sources, which requires higher-order statistics (HOS), and also of separating temporally-correlated Gaussian sources with distinct spectra, which demands temporal information. In [11], it is shown that, since correntropy is capable of quantifying nonlinear statistical relationships, it is suitable as a measure for identifying nonlinear dynamic systems.

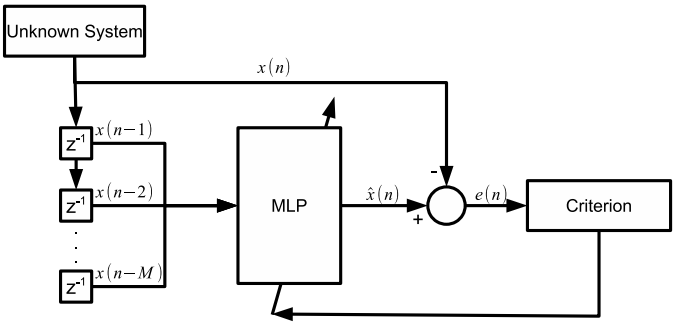

Fig. 1. Prediction scheme with a time-delayed MLP.

III. ITL Applications

ITL was proposed with a practical emphasis, aiming at the solution of complex signal processing problems that require a significant amount of information about the available data. As consequence, it is possible to form a representative set of applications of the algorithms into different tasks, employing different criteria, filtering structures and optimization procedures. In the following, some examples are provided to the reader, in order to give a general idea of the potentialities of this paradigm in dealing with modern, data-driven, engineering problems.

A. Dynamic modeling

The aim of dynamic modeling is to build a mathematical representation of the functional relationship between input and output variables. This is the case, for example, when one is interested in building a model that predicts the behavior of an unknown dynamical system. This is a problem often studied within neural network theory, for which a recent and successful approach has been the application of deep neural networks [12] or, alternatively, the popular and well-established multilayer perceptron (MLP) [13] in the role of predictor. The inputs are time-delayed measures of a state variable of the system and the model should provide an estimate of the current state value (see Figure 1).

While the most extensively used criterion in this context is the mean squared error, the typical ITL approach is based on the Minimum Error Entropy (MEE) criterion [14], which consists in the minimization of the error entropy with respect to the MLP synaptic weights. The ideal condition in this case is to have the error signal always at zero, i.e., $e(n)$ should have a distribution in the form of a delta function centered at zero.

Since the quadratic entropy, in association with the use of Gaussian kernels, yields a simple calculation of the required integral, as seen in Section IV.B of Part I, the authors argue in favor of its use in this scenario, arriving at an optimization problem with the following cost function, to be minimized via a gradient descent method[1]

$$J_{h_2} = \hat{h}_2(e) = -\log \hat{V}(e). \tag{6}$$

Recall that $\hat{V}(\cdot)$ is the information potential estimator, and we can drop the $-\log(\cdot)$ operation to simplify the expression, converting (6) into a cost function to be maximized. Hence, the gradient vector with respect to the weights is calculated and it is possible to apply the backpropagation algorithm [13]:

$$\frac{\partial \hat{V}(e)}{\partial w} = \frac{1}{2N^2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (e_j - e_i) \cdot \cdot G(e_i | e_j, 2\sigma^2) \cdot \left(\frac{\partial \hat{x}_j}{\partial w} - \frac{\partial \hat{x}_i}{\partial w} \right). \tag{7}$$

Observe that the expression has a computational cost proportional to N^2 — instead of $O(N)$, as the MSE-based training —, which is a direct consequence of adopting the information-theoretic criterion and its associated information potential estimator.

Simulations in chaotic time series prediction and nonlinear identification [14] have shown that the MEE criterion gave rise to an error PDF more concentrated around zero, while the PDF of the output signal was closer to that of the desired signals, in comparison with the corresponding results of an MSE-based neural network. Moreover, a sequence of this pioneering work indicated that the dynamic adjustment of the kernel size in the training, by means of an annealing process, increases the chance of escaping from locally optimal solutions [15].

Zupanc [16] gives additional support to the observations of previous works, through the comparative analysis between MSE and MEE in the prediction of a chaotic dynamic system and modeling of a polymer mixing process. The results indicated that the MEE criterion achieves a better generalization capability, is more robust to outliers and better approximates the PDF of the state variable under observation. However, the higher computational cost in comparison with an MSE training algorithm and the sensitivity to the kernel size adjustment are issues that the user must take into account.

Prediction of power generated by a wind park was another domain where the effectiveness of ITL criteria has been verified. This scenario is interesting because the error distribution is non-gaussian, and [17] showed that the MEE and the maximum correntropy criterion (MCC) are better than the MSE in providing accurate predictions for the offline and online training modes.

B. Classification

Classification is an important machine learning problem with a vast range of models, criteria and approaches. It can be represented in several ways, one of the most useful involves the definition of a set of discriminant function

$$g_i(\mathbf{x}), \quad i = 1, ..., c, \tag{8}$$

where $\mathbf{x}$ represents an m-dimensional feature vector of a given phenomenon to be classified into one of c possible classes. The classifier is said to assign a feature vector $\mathbf{x}$ to class ω_i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad \forall j \neq i. \tag{9}$$

One of the key aspects is, hence, the definition of the functional mapping $g_i(\cdot)$ and its optimal parameters. Consider an indicator function $\mathbf{1}_{\omega_i}(\mathbf{x})$, which is 1 if $\mathbf{x}$ belongs to class ω_i and 0 otherwise; $g_i(\cdot)$ should approximate the respective class indicator function and, hence, the error vector is defined as $\mathbf{e} = [g_1(\mathbf{x}) - \mathbf{1}_{\omega_1}(\mathbf{x}), ..., g_c(\mathbf{x}) - \mathbf{1}_{\omega_c}(\mathbf{x})]^T$. This formulation allows a straightforward application of MEE as optimization criterion, according to Shannon's

$$J_S = \hat{h}(\mathbf{e}), \tag{10}$$

or Rényi's definition:

$$J_{h_2} = \hat{h}_2(\mathbf{e}). \tag{11}$$

From the training perspective, a linear discriminant function can be adapted via the Stochastic Information Gradient algorithm [18] or artificial neural networks such as the MLP architecture can be adjusted via the gradient-based search with the backpropagation technique — recall the expressions derived in Section III-A.

But, regardless the training method, and since MSE is well-known due to the robustness and wide adoption as a criterion for classification, a fundamental question that arises when adopting MEE is: does it leads to a smaller classification error probability, for a given model?

Interestingly, whether Shannon's or Rényi's definition is employed, [19] shows that, for a perceptron-based classifier, MEE may not lead to solutions close to the minimal misclassification error. Moreover, there are theoretical situations where entropy maximization leads to the ideal configuration. Nevertheless, when the Parzen window entropy estimators are considered, their smoothing property can overcome such limitations, as long as an appropriate kernel size is defined.

This idea is reinforced in [20], where an extensive experimental simulation is performed comparing MSE, MEE (with both Shannon's and Rényi's definition), Cross-Entropy [21] and the generalized exponential risk in the context of MLP training for 35 different classification public datasets. The results indicate that MSE generally under-perform the other criteria, including MEE. Cross-Entropy and Exponential Risk achieved most of the highest classification rates among the datasets and, remarkably, MEE with Renyi's quadratic entropy obtained the poorest generalization capability.

To summarize, MEE criteria presents strong empirical evidences that is beneficial as a surrogate for MSE in classification tasks, however, there is a sensitive dependence on the database and the problem domain that suggests a careful analysis to the designer, in order to choose MEE or a different criteria.

[1]For simplification reasons, we shall also consider lowercase letters to represent the arguments of probabilistic / information-theoretic operators.

Another promising (and recent) classification criterion comes from the notion of correntropy, where the model can be adapted via the maximization of a cross-correntropy-based criterion [22]:

$$J_v = v(g_i(\mathbf{x}), \mathbf{1}_{\omega_i}(\mathbf{x})), \tag{12}$$

i.e. the functional mapping is defined in order to maximize the correntropy between the classifier output and the class labels. Recent works [23], [24] on image pattern recognition add to (12) a regularization term derived from sparsity analysis, to define a linear representation of a test image $\mathbf{y}$ such that

$$J_{v_{l_1}} = v(\mathbf{y}, \mathbf{X}\mathbf{w}) - \lambda|\mathbf{w}|_{l_1} \tag{13}$$

is maximized with respect to $\mathbf{w}$, where $\mathbf{X}$ represents the training dataset and $|\cdot|_{l_1}$ is the l_1-norm of a vector. The solution of this problem is obtained by half-quadratic optimization techniques, and it provides a vector basis for each class of objects to be recognized, which is subsequently adopted to classify the new image as belonging to the particular class basis that reconstructs the most similar (in the correntropy sense) prototype of $\mathbf{y}$.

The experimental results considering severe distortions, such as pixels occlusions and non-Gaussian noise, demonstrated a very good effectiveness of the method in such scenarios. Furthermore, the aforementioned correntropy-based criteria are being employed also in the context of deep learning and extreme learning classifiers [25], [26], with promising results as well.

C. Equalization

In bandlimited and high data rate digital communication systems, equalizers are important devices. Their function is to restore the transmitted information, i.e. the information at the channel input, mitigating or eliminating channel interference. In order to do so, a large variety of techniques have been developed in the last 70 years [27].

Equalization may be considered in two scenarios: supervised or unsupervised. Supervised methods are traditionally based on the MSE criterion, while unsupervised methods rely exclusively on HOS of the involved signals. Under the classical assumption of linearity and Gaussianity, the above mentioned methods are known to provide a reliable performance. However, with respect to non-classical scenarios, e.g., for sparse / correlated signals or even in presence of non-Gaussian noise, the same assertion cannot be hold.

In light of this, as ITL has the potential of extracting the complete statistical information present in signals, a very interesting option emerged: to employ new criteria based on this field to the problem of channel equalization, especially in non-classical scenarios. Hence, let us start by presenting the problem formulation.

Consider a source signal $s(n)$ being transmitted through a linear time-invariant channel. The channel output can be expressed as:

$$x(n) = \sum_i h_i s(n-i) + \eta(n) \tag{14}$$

where h_i are the channel coefficients and η is the additive noise.

The equalizer, designed to remove the intersymbol interference introduced by the channel, is generally modeled as a finite impulse response (FIR) filter. Its output may be written as:

$$y(n) = \sum_{i=0}^{D-1} w_i x(n-i) = \mathbf{w}^T \mathbf{x}(n) \tag{15}$$

where w_i are the values of the D filter coefficients.

In the sequel, we will present several ITL-based equalization algorithms.

1) Supervised Equalization: The application of ITL to supervised equalization started with the analysis of the use of the minimum quadratic Rényi's entropy of the error, MEE, between the desired signal and the equalizer output, instead of using the classical MSE criterion [28], [18]. Recalling that *entropy minimization* is equivalent to *information potential maximization*, as already mentioned in Section III-A, and using the Parzen window method to estimate the error PDF, the associated criterion results in:

$$J_V = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G(e_j|e_i, 2\sigma^2) \tag{16}$$

where $e_i = d(i-\tau) - y(i)$, being d the desired signal and τ the equalizer delay. The kernel size σ is a parameter to be adjusted according to the given scenario. Note that (16) considers the relationship between each pair of error samples. A gradient based method can be employed in order to maximize (16), being called stochastic information gradient for MEE (MEE-SIG) [18].

In situations where the channel is linear and the additive noise is gaussian, a linear equalizer that maximizes (16) will present a performance similar to that of an equalizer trained to minimize the MSE, since their solutions tend to be close to each other or even equivalent, under certain specific conditions [28]. However, this will not be the case for non-Gaussian noise, where the MEE-SIG algorithm tends to be more robust than that based on the MSE. Furthermore, the difference between the MEE and the MSE becomes more pronounced in nonlinear scenarios [14], [15]. As an example, when the channel is composed of a linear distortion followed by a nonlinear function and the equalizer is modeled as a multilayer perceptron neural network, by minimizing the entropy, it is possible to obtain an improved performance in equalization [28].

Another strong branch in supervised ITL criteria is that based on correntropy (Section II). Since this entity can be seen as a nonlinear similarity measure between random variables, one can apply it to the equalization problem by maximizing the correntropy between the transmitted and the equalizer output signals, giving rise to the maximum correntropy criterion (MCC) [6]:

$$J_v = \frac{1}{N} \sum_{i=n-N+1}^{n} G(d_i|y_i, \sigma^2), \tag{17}$$

where the kernel size σ, in this case, determines the length of the neighborhood of d_i to be considered. Hence, a suitable

choice of σ can improve the robustness of the MCC against outliers and impulsive noise. With respect to the MEE, the MCC presents the advantages of requiring a lower computational cost (note that there is a single summation operator for MCC) and being less sensitive to variations in σ. On the other hand, it can demand a larger number of samples N to provide a good estimate.

From (17), it is possible to derive a simple Least Mean Squares (LMS) like algorithm, called MCC-SIG [18]. In the presence of impulsive noise, such method has shown a better performance than the original LMS in system identification [6]. An interesting aspect is that, for a very large kernel size, the solution will be very close to the one obtained through the MSE criterion. An alternative algorithm for the optimization of (17) was proposed in [29], based on a fixed point solution, which presents a fast convergence when compared to the well-known Recursive Least Squares (RLS) algorithm [27], also being independent of the eigenvalue spread of the data.

2) Unsupervised Equalization: The use of the ITL framework in the task of unsupervised equalization is a very attractive possibility, in view of the natural availability of higher order statistical information required to solve the problem. In that sense, one of the first unsupervised ITL-based criteria [30] brings together Rényi's α-entropy and the idea behind the well-known blind constant modulus (CM) criterion [31], which penalizes deviations of the equalizer output from a constant modulus, to form the following criterion:

$$J_{SFA} = h_\alpha\left(|y(n)|^2 - R_2\right) = h_\alpha\left(|y(n)|^2\right) \tag{18}$$

where $R_2 = \frac{E\left[|s(n)|^4\right]}{E\left[|s(n)|^2\right]}$. The last equality comes from the fact that entropy does not depend on the mean of the signal. By assuming $\alpha = 2$ and using the IP estimator, the cost function to be maximized becomes:

$$\hat{J}_{SFA} = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} G(|y_{n-j}|^2||y_{n-i}|^2, 2\sigma^2). \tag{19}$$

The steepest descent algorithm resulting from (19) was named stochastic fast algorithm (SFA) [30]. It should be noted that some kind of constraint with respect to the equalizer taps has to be added in order to avoid the trivial solution — which can be done, for instance, by fixing one of the taps to unity or by admitting a unit norm constraint to the equalizer taps.

A parallel development based on the Benveniste-Goursat-Ruget theorem [32] — one of the milestones in blind equalization — was also reached for blind ITL criteria, which gravitates around the notion of matching the PDF of the equalizer output to that of the transmitted signal. As pointed out by [33], [34], [35], this idea can be translated into the following cost function

$$\begin{aligned} J_{QD} &= \int_{-\infty}^{\infty}\left(f_{Y^2}(v) - f_{S^2}(v)\right)^2 dv \\ &= \int_{-\infty}^{\infty}\left(f_{Y^2}^2(v) + f_{S^2}^2(v) - 2f_{Y^2}(v)f_{S^2}(v)\right) dv, \end{aligned} \tag{20}$$

where f_{Y^2} and f_{S^2} are the PDFs of the random variables Y^2 and S^2, which, in turn, are associated with the signals $|y(n)|^2$ and $|s(n)|^2$, respectively. In [35], all terms of (20) depending on the equalizer output are considered and, once again, the PDFs associated with the signals are estimated using the Parzen window method, resulting in a simplified cost function:

$$\begin{aligned} \hat{J}_{QD} = &\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} G(|y_{n-j}|^2||y_{n-i}|^2, 2\sigma^2) \\ &- \frac{2}{LN}\sum_{i=1}^{L}\sum_{j=0}^{N-1} G(|y_{n-j}|^2||s_i|^2, 2\sigma^2), \end{aligned} \tag{21}$$

where L is the cardinality of the transmitted symbol alphabet and s_i its ith symbol. The associated gradient-based algorithm was named stochastic quadratic distance (SQD). A slight modification was also proposed in [34], in which the PDF associated with the transmitted signal was evaluated in some specific target values. Still aiming at matching the PDFs, we also highlight the work of [33], where it is suggested the use of only the last term of (20) — the other terms simply work as a normalization factor between the PDFs and can be neglected. In this case, the estimation of PDFs via Parzen window results in the last term of (21), which we call $\hat{J}_{MQD}$.

Very interestingly, as presented in [36], the estimates of these blind criteria can be interrelated as:

$$\hat{J}_{QD} = \hat{J}_{SFA} + \hat{J}_{MQD}. \tag{22}$$

From this, we point out that these criteria will differ mainly in its computational complexity and robustness. While $\hat{J}_{QD}$ is considered the more complex and robust — as it encompasses a richer statistical information about both y_n and s_n —, the $\hat{J}_{MQD}$ offers a good trade-off and is an attractive option, since $\hat{J}_{SFA}$, due to the necessity of imposing a constraint to the equalizer coefficients, tends to be more susceptible to local convergence.

It is also important to indicate the points of contact that these ITL criteria establish with the classical CM criterion. J_{SFA} can be seen as a direct extension of the CM formulation to the ITL framework. $\hat{J}_{MQD}$ uses as kernel argument the deviations of the squared equalizer output from a fixed term, just like the CM criterion. Finally, since $\hat{J}_{QD}$ gathers contributions from both of these blind ITL criteria (22), it is expected that the quadratic distance criterion also preserves some elements of the CM approach. Indeed, by comparing the surface contours of $\hat{J}_{QD}$ and the CM cost, as illustrated in Figure 2 for the channel with impulse response $H(z) = 1 + 0.6z^{-1}$, there are some similarities between the minima. Besides that, for linear equalizers and under the hypothesis of Gaussianity, the blind ITL criteria behave similarly to the CM, but, for impulsive noise models, the latter loses performance. Similar ITL blind methods for deconvolution can be found in [37],[38] and [39].

Although the criteria discussed above begin with distinct hypothesis, they all assume a common feature: the transmitted signal are composed of iid samples. However, in practical scenarios, the sources may exhibit temporal dependence, in consequence, for instance, of the application of codes before signal transmission and the handling of analog discrete-time signal processing (e.g. in audio-related scenarios). In that

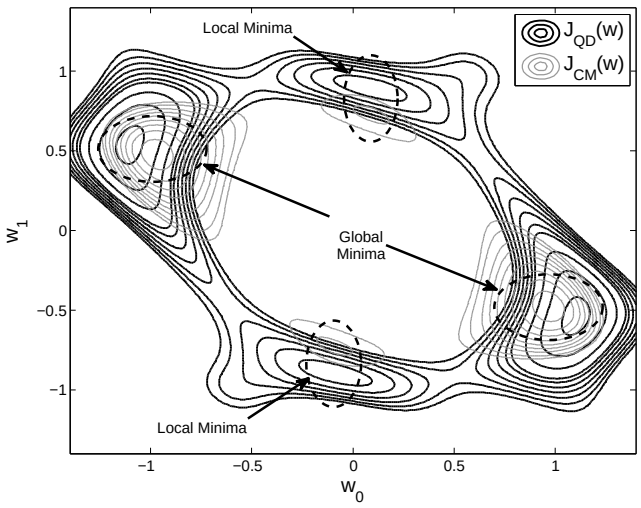

Fig. 2. Surface Contour of the QD and CM criteria

sense, as initially proposed in [1], the ITL measure of correntropy can be used to statistically evaluate the time structure of the signal in a blind context. The objective is to make the correntropy of the equalizer output as close as possible to the correntropy of the transmitted source, known *a priori*:

$$\hat{J}_{corr} = \sum_{m=1}^{P} \left(v_s(m) - \hat{v}_y(m)\right)^2 \tag{23}$$

where v_s is the correntropy of the source, $\hat{v}_y$ is the estimated correntropy of the equalizer output and P is the number of lags considered. Since correntropy takes into account the statistical and temporal structures of the signals, it has shown a good performance when treating correlated sources, a situation that classical methods fail to equalize [7]. Another advantage of the correntropy-based method is its reduced computational complexity in comparison with the PDF matching-based criteria — although it can demand an elevated number of samples for estimation.

D. Independent Component Analysis

Independent Component Analysis (ICA) has been originated as a natural extension of Principal Component Analysis (PCA), and both techniques are unsupervised signal processing paradigms. They are also very useful tools in the context of factor analysis [40].

One of the most representative problems to which ICA is applied is Blind Source Separation (BSS), a task in which information-theoretic optimization criteria have been used with success for three decades [41]. BSS, in its linear and instantaneous form, can be formulated as: consider that one observes, at a given time instant, the signal $\mathbf{x}$, m-dimensional, which is the result of the linear combination of $k \leq m$ independent signals (sources), i.e.

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{24}$$

where $\mathbf{s}$ is the source vector, k-dimensional, and $\mathbf{A}$ is the mixing matrix, $m \times k$.

Without *a priori* knowledge of $\mathbf{A}$ and $\mathbf{s}$, the problem consists in obtaining a demixing matrix $\mathbf{W}$ to estimate the output vector $\mathbf{y} = \mathbf{W}\mathbf{x}$ such that it be equal to $\mathbf{s}$ up to scale and permutation factors. Figure 3 illustrates this formulation, when $k = m$.

The connection between ICA and BSS comes from the fact that, if the solution $\mathbf{W}$ generates a set of components $\mathbf{y}$ that are independent, this ensures that the original $\mathbf{s}$ has been recovered. In this context, up to the previously mentioned ambiguities, there are several criteria to perform ICA, including (*i*) negentropy [40], a criterion to maximize non-Gaussianity; (*ii*) the Infomax principle [42], which is based on the idea of maximizing the information flow between the mixtures and the separating system outputs; (*iii*) the minimization of entropy rate [43], [44], which, similarly to correntropy, allows the exploration of both statistical and temporal structures of the signals; (*iv*) cumulants and (*v*) kurtosis [45].

In a general perspective, the ICA criteria can be related to mutual information (MI) rate minimization between the separating system outputs, where two diversity aspects may be considered, in a joint manner or independently: HOS and dependence of source samples [46]. Thus, it is convenient to directly measure the independence degree of these signals via ITL methods. As an example, if only HOS diversity is considered, mutual information rate reduces to mutual information, which yields the following criterion to be minimized:

$$J_{ICA} = \sum_{i=1}^{m} h(y_i) - \log|\det \mathbf{W}|. \tag{25}$$

Furthermore, if the observations are previously whitened (PCA) and, as consequence, the demixing matrix is a pure rotation matrix, the loss function in (25) is reduced to just the first term, the sum of marginal output entropies[2]. The Minimum Rényi's Mutual Information (MRMI) algorithm [47] applies the gradient descent method to minimize (with respect to the elements of $\mathbf{W}$) this reduced cost function, replacing Shannon's differential entropy by Rényi's definition with $\alpha = 2$. The marginal entropies are estimated with the well known non-parametric entropy estimator (recall the definition in Section IV.B of Part I), already employed in previous applications.

[2]Although this assumption brings important practical advantages, such as making second-order search methods easier for implementation in ICA algorithms, constraining $\mathbf{W}$ to be orthogonal limits the search space and, consequently, the achievable performance [46].

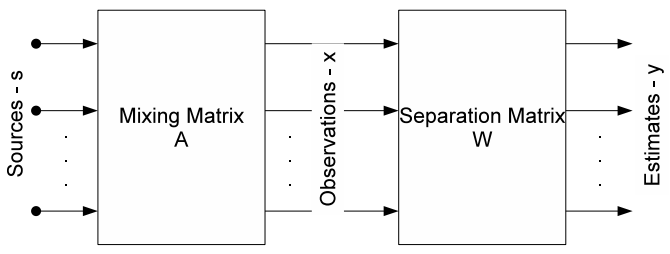

Fig. 3. Linear and instantaneous formulation of the Blind Source Separation problem.

Nonetheless, a deeper analysis of this method [48] has recently demonstrated that the maximal value of Rényi's entropy is associated with the Gaussian distribution *only when* $\alpha = 1$. In this case, a modification in the MRMI original criterion is necessary according to the estimated distribution, which can be classified as: sub-Gaussian, a distribution flatter and shorter-tailed (with kurtosis less than 3) than the Gaussian, or super-Gaussian, a distribution more peaked and longer-tailed (with kurtosis greater than 3) than the Gaussian. The new implementation was compared with other traditional ICA algorithms in separating audio recordings and the results indicated a superior performance of MRMI with the parameter $\alpha = 2$. Nevertheless, the theoretical results of modified MRMI are valid only for the generalized exponential family of distributions, and [49] demonstrated that the choice of the α value may lead to a cost function that does not satisfy the requirements of a contrast function [41]. The final conclusion is that the adoption of Rényi's entropy to perform ICA must be preceded by a careful analysis of the scenario at hand.

An extension of the BSS problem that has been recently studied is related to the post-nonlinear (PNL) model [50], which adds nonlinear, memoryless and invertible functions to the BSS linear model. These functions may represent, for example, the effect of sensors in some measurement process. To perform the separation, it is necessary to apply the nonlinear memoryless functions to each component of $\mathbf{x}$ previously to the demixing matrix $\mathbf{W}$.

One of the most robust approaches to separate PNL mixtures is also based on the independence recovery with mutual information minimization. In this direction, [51] uses a direct MI estimator based on order statistics (recall Section IV.C of Part I) as cost function to be minimized with an immune-inspired algorithm.

1) ICA over Finite Fields: Recently, ICA has been extended to the domain of finite and discrete valued signals and systems. Yeredor [52] firstly explored this idea with the development of an ICA algorithm for Boolean signals mixed in accordance with XOR and classical product operations, i.e. in the context of a Galois field of order two. The algorithm iteratively extracts the sources by searching for the linear combination of the mixtures that minimizes the entropy, followed by a deflation process [53] to remove the extracted source from the mixtures.

Afterwards, [54] extended the algorithm towards dealing with finite fields of any size. Analogously, [55] improved the pioneering algorithm, known as AMERICA, and proposed a faster (but less accurate) algorithm based on sequential reduction of the pairwise mutual information, the name of which is MEXICO. A summary of all contributions, at that point, was consolidated in [56].

All these techniques comprise parameter adaptation based on ITL cost functions — the histogram based estimator of Shannon's entropy (see Section IV.F of Part I) — and a sequential search for the separating matrix elements. Differently from this perspective, [57] proposed the application of an immune-inspired algorithm to search for the complete separating matrix. The problem was formulated as a combinatorial optimization task such that the solution was the separating matrix that led to the minimal mutual information (actually, the sum of marginal entropies — recall (25)) between extracted components.

Another related proposal was subsequently developed [58], where a more robust immune-inspired algorithm was applied in association to a *Michigan*-like approach [59] to model the population individuals. The algorithm criterion was to minimize the entropy of each extracted source, considering that the intrinsic diversity operators of the algorithm may allow that distinct independent signals are obtained, in the end.

E. Cluster Analysis

Clustering is a self-organizing process that plays an important role in a broad range of fields, from pattern recognition, signal compression [21], and knowledge discovery in databases [60] to communication channel estimation and/or equalization [61]. Roughly speaking, clustering is aimed at partitioning a set of objects into groups that share some kind of (predefined) similarity. Clearly, it is not a well-posed problem, just like the estimation of MI from finite dataset. To clarify this important point, let us consider a tiny data set of 5 points, represented in Figure 4.

For someone looking for cluster formation in observed data, a naive clustering hypothesis may promptly be raised, as illustrated in Figure 5. Nevertheless, even though a visual inspection may lead one to accept the plausibility of this first hypothesis, there is not a consensual way to measure it, unless we define/chose a numeric criterion, which is, in turn, an arbitrary decision/choice itself. Indeed, it is well known that different clustering criteria, when applied to the very same dataset, may provide different clustering hypotheses, mainly when the number of available samples under analysis is small. This means that one cannot even infer the existence of clusters from finite datasets themselves: any conclusion should be based on *a priori* information about the data source model. Note that this is implicitly true even when a simple distance measure is used in a clustering criterion.

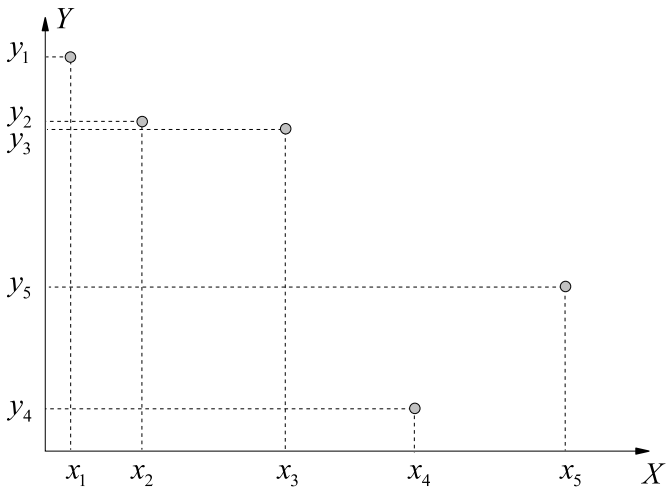

Fig. 4. Five 2D numeric samples drawn at random from a unknown source.

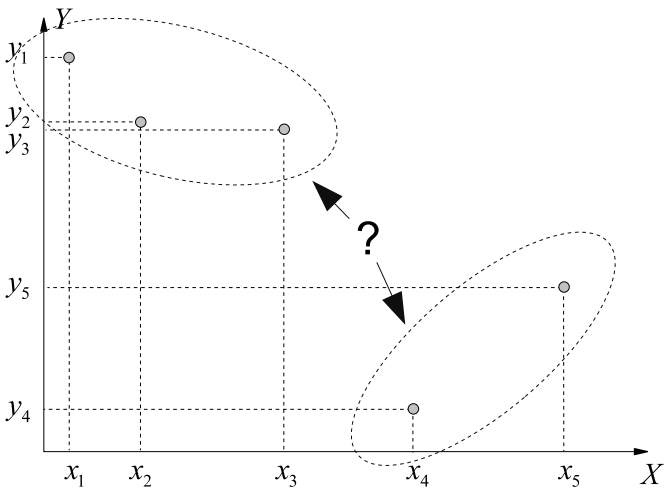

Fig. 5. Clustering hypothesis from five 2D numeric samples.

This unavoidable need for *a priori* information engenders a striking analogy between cluster analysis and mutual information estimation from finite datasets. To properly show it, we first recall that the MI between two random variables, say X and Y, is the amount of decrease in the uncertainty regarding one of them when the other is known (recall Section 2 of Part I). Strictly speaking, for finite datasets, unless there are coincident values (which have a vanishing probability for continuous variables), there is no randomness to be removed. For instance, in Figure 4, by knowing that event $X = x_3$ occurred, we conclude, deterministically, that $Y = y_3$ is to occur too. This means that, strictly speaking, for finite sets of samples generated by continuous variables, the mutual information that can be inferred without any *a priori* source model is always maximal (i.e. no randomness at all)!

Evidently, any useful analysis must consider a source model, and use data to adjust this model, as in clustering. Not surprisingly, there can be found in literature many works combining both analyses, mainly on the simplified use of MI for finding consensus among many clustering hypothesis [62], [63].

One simple and straightforward combination of MI and clustering ensemble concerns the estimation of the number of clusters formed by finite datasets in metric spaces. Again, it is an ill-posed problem in clustering analysis, and the use of many clustering hypotheses, along with an MI based criterion, may facilitate the difficult choice of a specific metric and an algorithm to this task. Indeed, in a clustering ensemble based approach, an arbitrarily large number of clustering algorithms, M, provide independent clustering hypothesis with K clusters each one. Each hypothesis yields a vector of labels (one label per pattern), which are regarded as random outputs drawn from M sources of K symbols, thus creating M random discrete variables, X_m. Figure 6 illustrates this ensembling process. Therefore, for each X_m, an entropy measure can be obtained, as follows:

$$H(X_m) = -\sum_{k=1}^{K} p_m(k) \log p_m(k), \quad 1 \leq m \leq M$$

where $p_m(k) = P[X_m = k]$ stands for the probability of randomly selecting the k-th label from the vector of labels $\mathbf{x}_m$.

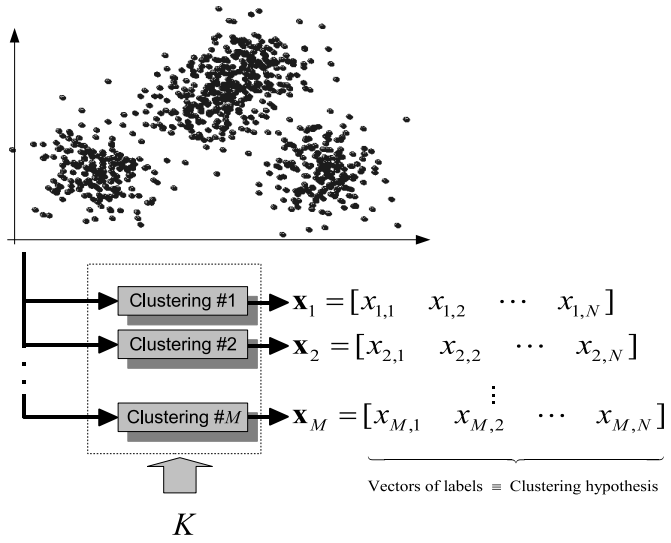

Fig. 6. Clustering ensemble of M algorithms, providing M clustering hypothesis with K clusters each one, codified as M vectors of labels.

Moreover, for each pair of the so defined random variables, it is also possible to measure their labeling agreement through MI, by computing:

$$I(X_i; X_j) = H(X_i) + H(X_j) - H(X_i, X_j)$$

where $H(X_i, X_j) = -\sum_{a=1}^{K}\sum_{b=1}^{K} p_{i,j}(a,b) \log p_{i,j}(a,b)$ and $p_{i,j}(a,b) = P[X_i = a, X_j = b]$.

Given a fixed K, one may average all pairwise quantities $I(X_i; X_j)$ as a measure of how much the imposition of K clusters corresponds to a stable configuration. Furthermore, to properly test this "natural" stability, it is also necessary to ensure a diversity of clustering hypotheses, which can be done (a) through the use of many different clustering methods, (b) through the subsampling of available data or (c) both strategies.

Finally, because the ranges of values for the entropies and the MI depend on K, one would prefer to normalize $I(X_i; X_j)$, yielding the Normalized Mutual Information (NMI). This normalization procedure is not unique. A usual choice is:

$$NMI(X_i; X_j) = \frac{I(X_i; X_j)}{max(H(X_i), H(X_j))} \tag{26}$$

and the Averaged NMI for the ensemble of clustering hypothesis is given by:

$$ANMI(\{X_k\}) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} NMI(X_i; X_j). \tag{27}$$

The ANMI is likely to be higher for values of K corresponding to stable clustering hypotheses. Therefore, by varying K, it is possible to estimate the number of clusters in a dataset. As an illustration, Figures 7 and 8 show two bidimensional

datasets whose visual inspection provides initial guesses concerning the number of clusters in each one. In both cases, we used ensembles of $M = 20$ clustering hypotheses provided by the standard K-Means algorithm. Diversity of hypotheses was induced by simple dataset subsampling. In Figure 7, the ANMI peaks at $K = 3$, whereas, in Figure 8, as the upper cluster spreads up, this peak moves to $K = 4$, but both values (i.e. 3 and 4) seem to be almost equally likely, which suits most human observer opinions.

This approach relies on a committee of clustering algorithms, whose computational complexity depends on designer choices. On the other hand, as for the remaining structure, the computational complexity is mainly dominated by the computation of the $M(M-1)/2$ pairwise information terms, $I(X_i; X_j)$, for each tested value of K.

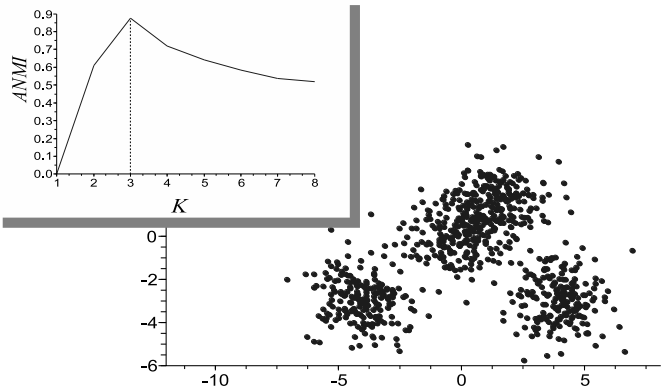

Fig. 7. Estimation of the number of clusters through Averaged Normalized Mutual Information (ANMI) – strong consensus in favor of 3 clusters.

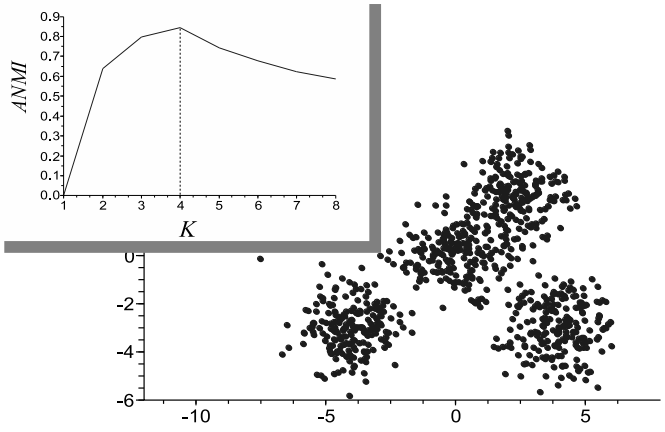

Fig. 8. Estimation of the number of clusters through Averaged Normalized Mutual Information (ANMI) - weak consensus in favor of 4 clusters.

IV. Conclusion

This two-part work presented an introduction to information theoretic learning, an emerging discipline that employs information theory for developing new machine learning criteria and algorithms.

The Part I of this tutorial was devoted to a description of fundamental concepts of information theory, from the seminal work of 1948 by Claude E. Shannon — which is considered the 'birth' of this field — to the generalized measures proposed by Alfred Rényi, which allowed, approximately three decades later, the application of definitions such as entropy and mutual information in the context of new adaptive algorithms that can effectively explore the higher-order statistical content of data. Furthermore, the problem of estimating information-theoretic measures from the available data is discussed, as, differently from classical applications of Information Theory, in ITL, as a rule, there is no prior knowledge about the probability distributions that are involved.

In Part II, a new concept that arises from Rényi's quadratic entropy and the idea of information potential is presented: correntropy, a nonlinear similarity measure that possesses several possibilities of applications, mainly in the domain of signals with a temporal structure. The rest of the paper brings a set of representative problems for which ITL provides effective solutions: dynamic modeling, classification, equalization, independent component analysis and cluster analysis. In each case, we present the main criteria that have been developed, together with the pros and cons of each methodology.

Although this tutorial does not cover the whole spectrum of applications that this research field already presents, we expect that it has provided the reader with a general understanding of the motivations and characteristics of ITL techniques. Moreover, the references employed in this work can be recommended as a basis for further study.

Acknowledgment

The authors thank FAPESP (Grant 2013/14185-2), CAPES and CNPq for the financial support.

References

[1] I. Santamaria, P. Pokharel, and J. Príncipe, "Generalized correlation function: Definition, properties and application to blind equalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187–2197, 2006. doi: 10.1109/TSP.2006.872524

[2] J. Príncipe, D. Xu, and J. Fischer, *Unsupervised Adaptive Filtering*. Wiley, 2000, vol. 1, ch. Information Theoretic Learning, pp. 265–319.

[3] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950. doi: 10.1090/S0002-9947-1950-0051437-7

[4] Z. Yang, A. T. Walden, and E. J. McCoy, "Correntropy: Implications of nongaussianity for the moment expansion and deconvolution," *Signal Processing*, vol. 91, pp. 864–876, 2011. doi: 10.1016/j.sigpro.2010.09.004

[5] W. Liu, P. P. Pokharel, and J. C. Príncipe, "Correntropy: A localized similarity measure," in *Int. Joint Conference on Neural Networks, Vancouver, Canada*, 2006, pp. 4919–4924.

[6] A. Singh and J. Príncipe, "Using correntropy as a cost function in linear adaptive filters," in *Proc. Of International Joint Conference on Neural Networks, Atlanta, USA*, 2009. doi: 10.1109/IJCNN.2009.5178823 pp. 2950–2955.

[7] A. Neves, C. Wada, R. Suyama, R. Attux, and J. M. T. Romano, *Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, ch. An Analysis of Unsupervised Signal Processing Methods in the Context of Correlated Sources, pp. 82–89.

[8] J. R. Treichler and B. G. Agee, "A new approach to multipath correction of constant modulus signals," *IEEE Transactions on Adaptive Speech and Signal Processing*, vol. 31, no. 4, pp. 349—-472, 1983. doi: 10.1109/TASSP.1983.1164062

[9] D. G. Fantinato, R. Attux, A. Neves, R. Suyama, and J. M. T. Romano, "Blind deconvolution of correlated sources based on second-order statistics," in *SBrT 2013*. Sociedade Brasileira de Telecomunicações, 2013. doi: 10.14209/sbrt.2013.103

[10] R. Li, W. Liu, and J. Príncipe, "A unifying criterion for instantaneous blind source separation based on correntropy," *Signal Processing*, vol. 87, pp. 1872–1881, 2007. doi: 10.1016/j.sigpro.2007.01.022

[11] A. Gunduz and J. Príncipe, "Correntropy as a novel measure for nonlinearity tests," *Signal Processing*, vol. 89, pp. 14–23, 2009. doi: 10.1016/j.sigpro.2008.07.005

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539

[13] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Prentice Hall, 2008.

[14] D. Erdogmus and J. Príncipe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Transactions on Signal Processing*, vol. 50, no. 7, pp. 1780 – 1786, 2002. doi: 10.1109/TSP.2002.1011217

[15] D. Erdogmus and J. C. Príncipe, "Generalized information potential criterion for adaptive system training," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1035–1044, 2002. doi: 10.1109/TNN.2002.1031936

[16] J. Zupanc, "Error-entropy minimization for dynamical systems modeling," in *Artificial Neural Networks-ICANN 2008*. Springer, 2008, pp. 417–425.

[17] R. J. Bessa, V. Miranda, and J. a. Gama, "Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting," *IEEE Transactions on Power Systems*, vol. 24, no. 4, pp. 1657–1666, 2009. doi: 10.1109/TPWRS.2009.2030291

[18] J. Príncipe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Verlag, 2010.

[19] L. M. Silva, J. M. de Sá, and L. A. Alexandre, "The mee principle in data classification: a perceptron-based analysis," *Neural computation*, vol. 22, no. 10, pp. 2698–728, oct 2010. doi: 10.1162/NECO_a_00013

[20] L. M. Silva, J. M. Santos, and J. M. de Sá, "Classification performance of multilayer perceptrons with different risk functionals," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 6, pp. 1–17, 2014. doi: 10.1142/S021800141450013X

[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2001.

[22] A. Singh and J. C. Principe, "A loss function for classification based on a robust similarity metric," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, jul 2010. doi: 10.1109/IJCNN.2010.5596485 pp. 1–6.

[23] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "A regularized correntropy framework for robust pattern recognition." *Neural computation*, vol. 23, pp. 2074–2100, 2011. doi: 10.1162/NECO_a_00155

[24] "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2011. doi: 10.1109/TPAMI.2010.220

[25] H.-J. Xing and X.-M. Wang, "Training extreme learning machine via regularized correntropy criterion," *Neural Computing and Applications*, pp. 1–10, 2012. doi: 10.1007/s00521-012-1184-y

[26] L. Chen, H. Qu, J. Zhao, B. Chen, and J. C. Principe, "Efficient and robust deep learning with correntropy-induced loss function," *Neural Computing and Applications*, 2015. doi: 10.1007/s00521-015-1916-x

[27] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice Hall, 2001.

[28] I. Santamaria, D. Erdogmus, and J. Príncipe, "Entropy minimization for supervised digital communications channel equalization," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1184–1192, 2002. doi: 10.1109/78.995074

[29] A. Singh and J. Príncipe, "A closed form recursive solution for maximum correntropy training," in *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 2070 – 2073.

[30] I. Santamaria, C. Vielva, and J. Príncipe, "Fast algorithm for adaptive blind equalization using order-α renyi's entropy," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 3. IEEE, 2002, pp. III–2657–III–2660.

[31] D. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Transactions on Communications*, vol. 28, no. 11, pp. 1867–1875, 1980. doi: 10.1109/TCOM.1980.1094608

[32] A. Benveniste, M. Goursat, and G. Ruget, "Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications," *IEEE Transactions on Automatic Control*, vol. AC-25, no. 3, pp. 385–399, 1980. doi: 10.1109/TAC.1980.1102343

[33] M. Lazaro, I. Santamaria, C. Pantaleon, D. Erdogmus, , and J. Principe, "Matched pdf-based blind equalization," in *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, 2003, pp. 297–300.

[34] M. Lazaro, I. Santamaria, C. Pantaleon, D. Erdogmus, K. E. H. II, and J. Príncipe, "Blind equalization by sampled PDF fitting," in *Proc. Fourth Int. Symp. Ind. Component Anal. Blind Equalization, Nara, Japan*, 2003, pp. 1041–1046.

[35] M. Lazaro, I. Santamaria, D. Erdogmus, K. Hild, C. Pantaleon, and J. Príncipe, "Stochastic blind equalization based on pdf fitting using parzen estimator," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 696–704, 2005. doi: 10.1109/TSP.2004.840767

[36] I. Santamaria, C. Pantaleon, L. Vielva, and J. Príncipe, "Adaptive blind equalization through quadratic pdf matching," in *Proc. Eur. Signal Processing Conf.*, 2002, pp. 289–292.

[37] D. Erdogmus, J. C. Príncipe, and L. Vielva, "Blind deconvolution with minimum renyi's entropy," in *Signal Processing Conference, 2002 11th European*, 2002, pp. 1–4.

[38] D. Erdogmus, K. Hild, II, and J. Príncipe, "Online entropy manipulation: Stochastic information gradient," *IEEE Signal Processing Letters*, vol. 10, no. 8, pp. 242–245, 2003. doi: 10.1109/LSP.2003.814400

[39] D. Erdogmus and J. Príncipe, "Adaptive blind deconvolution of linear channels using renyi's entropy with parzen window estimation," *IEEE Transactions on Signal Processing*, vol. 52, no. 6, pp. 1489–1498, 2004. doi: 10.1109/TSP.2004.827202

[40] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-Interscience, 2001.

[41] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994. doi: 10.1016/0165-1684(94)90029-9

[42] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995. doi: 10.1162/neco.1995.7.6.1129

[43] G.-S. Fu, R. Phlypo, M. Anderson, X.-L. Li, and T. Adali, "Blind source separation by entropy rate minimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4245–4255, 2014. doi: 10.1109/TSP.2014.2333563

[44] G.-S. Fu, W. Du, and T. Adali, "Entropy rate estimation for vector processes: Application to complex fmri analysis," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1867–1871.

[45] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation*. Oxford, UK: Academic Press, 2010.

[46] T. Adali, M. Anderson, and G.-s. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 18–33, 2014. doi: 10.1109/MSP.2014.2300511

[47] K. Hild, D. Erdogmus, J. Príncipe *et al.*, "Blind source separation using renyi's mutual information," *IEEE Signal Processing Letters*, vol. 8, no. 6, pp. 174–176, 2001. doi: 10.1109/97.923043

[48] K. Hild II, D. Erdogmus, and J. Príncipe, "An analysis of entropy estimators for blind source separation," *Signal Processing*, vol. 86, no. 1, pp. 182–194, 2006. doi: 10.1016/j.sigpro.2005.04.015

[49] F. Vrins, D. Pham, and M. Verleysen, "Is the general form of renyi's entropy a contrast for source separation?" in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 129–136.

[50] A. Taleb and C. Jutten, "Nonlinear source separation: The post-nonlinear mixtures," in *European Symposium on Artificial Neural Networks*, 1997, pp. 279–284.

[51] L. Duarte, R. Suyama, R. Attux, F. Von Zuben, and J. Romano, "Blind source separation of post-nonlinear mixtures using evolutionary computation and order statistics," *Independent Component Analysis and Blind Signal Separation*, vol. 3889, pp. 66–73, 2006. doi: 10.1007/11679363_9

[52] A. Yeredor, "ICA in boolean XOR mixtures," in *Proceedings of Independent Component Analysis and Signal Separation, ICA 2007*. Springer, 2007, pp. 827–835.

[53] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach," *Signal Processing*, vol. 45, no. 1, pp. 59–83, 1995. doi: 10.1016/0165-1684(95)00042-C

[54] H. Gutch, P. Gruber, and F. Theis, "ICA over finite fields," in *ICA 2010 - Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 645–652.

[55] A. Yeredor, "Independent component analysis over Galois fields of prime order," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5342–5359, 2011. doi: 10.1109/TIT.2011.2145090

[56] H. W. Gutch, P. Gruber, A. Yeredor, and F. J. Theis, "ICA over finite fields – separability and algorithms," *Signal Processing*, vol. 92, no. 8, pp. 1796 – 1808, 2012. doi: 10.1016/j.sigpro.2011.10.003
[57] D. G. Silva, R. Attux, E. Z. Nadalin, L. T. Duarte, and R. Suyama, "An immune-inspired information-theoretic approach to the problem of ICA over a Galois field," in *Information Theory Workshop (ITW), 2011 IEEE*, oct. 2011. doi: 10.1109/ITW.2011.6089571 pp. 618 –622.
[58] D. G. Silva, E. Z. Nadalin, G. P. Coelho, L. T. Duarte, R. Suyama, R. Attux, F. J. Von Zuben, and J. Montalvão, "A michigan-like immune-inspired framework for performing independent component analysis over galois fields of prime order," *Signal Processing*, vol. 96, pp. 153–163, Mar. 2014. doi: 10.1016/j.sigpro.2013.09.004
[59] T. Back, D. B. Fogel, and Z. Michalewicz, Eds., *Evolutionary Computation 1: Basic Algorithms and Operators*. Bristol, UK: Taylor & Francis, 2000.
[60] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: A survey," *IEEE Transactions on Neural Networks*, vol. 13, pp. 3–14, 2002. doi: 10.1109/72.977258
[61] J. Montalvão, B. Dorizzi, and J. C. M. Mota, "Channel estimation by symmetrical clustering," *IEEE Transactions on Signal Processing*, vol. 50, pp. 1459–1469, 2002. doi: 10.1109/TSP.2002.1003069
[62] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.
[63] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1866–1881, 2005. doi: 10.1109/TPAMI.2005.237

Daniel G. Silva was born in Botucatu, Brazil, in 1983. He received the B.S. degree in computer engineering and the M.S. and Ph.D. degrees in electrical engineering, all from the University of Campinas (UNICAMP), São Paulo, Brazil, in 2006, 2009, and 2013, respectively. Currently, he is a Professor at the Department of Electrical Engineering (ENE) of the University of Brasília (UnB). His main research interests are information theoretic learning, adaptive signal processing and computational intelligence.

Denis G. Fantinato was born in Americana, Brazil, in 1985. He received the B.S. and M.Sc. degrees in Electrical Engineering from the University of Campinas (UNICAMP) in 2011 and 2013, respectively. Currently, he is a Ph.D. student at the same institution. His main research interests are blind signal processing, adaptive filtering and information theoretic learning.

Jânio C. Canuto was born in Maceió, Brazil, in 1984. He received the B.S. degree in Electrical Engineering (2007) from the Federal University of Sergipe (UFS), M.S. degree in Electrical Engineering (2010) from the University of Campinas (UNICAMP) and Ph.D. degree in Computer Science (2014) from Télécom SudParis. He joined the Department of Computer Science of the Federal University of Sergipe (UFS) in 2016. His main research interests are pattern recognition and machine learning.

Leonardo T. Duarte received the B.S. and the M.Sc. degrees in electrical engineering from UNICAMP (Brazil) in 2004 and 2006, respectively, and the Ph.D. degree from the Grenoble Institute of Technology (Grenoble INP), France, in 2009. He is currently an assistant professor at the School of Applied Sciences at UNICAMP. His research interests are mainly associated with the theory of unsupervised signal processing and include signal separation, independent component analysis, Bayesian methods, and applications in chemical sensors and seismic signal processing. He is also working on unsupervised schemes for adjusting multiple-criteria decision analysis (MCDA) techniques. He is a Senior Member of the IEEE.

Aline O. Neves received the B.S. and M.S. degree in Electrical Engineering from the University of Campinas (UNICAMP), Brazil, in 1999 and 2001 respectively. She received her Ph.D. degree in 2005, also in Electrical Engineering, from the University René Descartes (Paris V), Paris, France. Recently, she is an associate professor at the Engineering, Modeling and Applied Social Science Center of the Federal University of ABC, Santo André, Brasil. Her research interests consist of equalization, channel estimation, source separation and information theoretic learning.

Ricardo Suyama was born in São Paulo, Brazil, in 1978. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the State University of Campinas (Unicamp), Campinas, Brazil, in 2001, 2003 and 2007, respectively. Currently he is an Assistant Professor at the Universidade Federal do ABC (UFABC), Sao Paulo, Brazil. His research interests include blind source separation, adaptive equalization, adaptive nonlinear filtering, and evolutionary algorithms.

Jugurta Montalvão was born in Aracaju, Brazil, in 1968. He received the title of Electrical Engineer (1992) from the University of Campina Grande (UFPB II), Master in Electrical Engineering (1995) from the University of Campinas (UNICAMP) and Doctor in "Automatique et traitement du signal" (2000) from the University Paris-Sud XI. He joined the Department of Electrical Engineering of the Federal University of Sergipe (UFS) in 2005. His main research interests are: pattern recognition and signal processing.

Romis Attux was born in Goiânia, Brazil, in 1978. He received the titles of Electrical Engineer (1999), Master in Electrical Engineering (2001) and Doctor in Electrical Engineering (2005) from the University of Campinas (UNICAMP), Brazil. Currently, he is an associate professor at the same institution. His main research interests are adaptive filtering, computational intelligence, dynamical systems / chaos and brain-computer interfaces.