On the Relationships between Blind Equalization and Blind Source Separation – Part I: Foundations

Romis Attux, Aline Neves, Leonardo T. Duarte, Ricardo Suyama, Cynthia Junqueira, Leandro Rangel, Tiago M. Dias, João M. T. Romano

***Abstract*— The objective of this two-part work is to present and discuss the relationships between the problems of blind equalization and blind source separation. This first part, which is essentially a tutorial, begins with a systematic exposition of the basic concepts that form the core of equalization theory, starting from the fundamental idea that characterizes the zero-forcing solution and reaching, after an explanation of the supervised Wiener paradigm, an analysis of the unsupervised or blind techniques. Afterwards, the problem of blind source separation and the main approaches to solving it are studied; important concepts are discussed, such as those of principal component analysis (PCA), independent component analysis (ICA) and strategies founded on bases as diverse as the use of mutual information as a measure of independence, the idea of nongaussianity and the employment of the classical process of estimation via the method of maximum-likelihood.**

***Index Terms*— Adaptive Filtering, Blind Equalization, Blind Source Separation, Independent Component Analysis.**

I. INTRODUCTION

A recurrent necessity in signal processing is that of extracting or restoring information from a corrupted version thereof. This fundamental requirement is embodied in the problems of *blind equalization* and *blind source separation* (BSS), on which it can be said that the theory of unsupervised adaptive filtering is based. Interestingly enough, the development of the theory underlying these most interrelated problems took place along different lines: while most techniques for blind equalization were conceived in the context of a classical SISO (single-input / single-output) model, BSS evolved basically under the aegis of formulations of a purely spatial character. Two decades of efforts led these fields of research into a significant degree of maturity; nevertheless, the potential synergy between these branches, which could be decisive to enrich and facilitate their development, still remains to be fully exploited.

Influenced by this spirit, we investigate, in this work, the various relationships between the problems of blind equalization and blind source separation. The work is divided into two parts, the first devoted to the exposition of the main theoretical results concerning the problems of interest, and the second, to a broad discussion of several aspects of the relation between them. This division serves an important purpose: to allow a non-expert reader to understand both problems and, afterwards, to have access to the points of contact between them.

In this first part, our starting point is the classical problem of SISO equalization. Afterwards, the problem of multichannel equalization is briefly discussed and serves as a conceptual link to the next subject: blind source separation. The problem of BSS and the most important approaches to solve it are then presented. It is important to remark that the entire exposition is based on the aim of preserving an illustrative sequence of the manner whereby the concepts are treated in the literature.

II. BLIND EQUALIZATION

The objective of a communication system is to allow information to be adequately interchanged between a transmitter and a receiver that, by hypothesis, are interconnected by a *channel*. As a rule, this channel is responsible for the introduction of a certain level of distortion in the transmitted message that, if not properly dealt with, may decisively compromise the quality of the reconstructed signal of interest.

A typical strategy to overcome this practical difficulty is to use a filter whose structure and parameters are carefully chosen in order to counterbalance the channel influence: the *equalizer*. In Fig. 1, we depict a simple scheme of a baseband communication system endowed with a device of this sort. As

Romis Attux, is with the Laboratory of Signal Processing for Communications (DSPCom) – FEEC – University of Campinas (Unicamp) – CP 6101 – CEP 13083-970 – Campinas – SP – Brazil (phone: +55-1935213857; fax: +55-19-35213857; e-mail: attux@dca.fee.unicamp.br).

Ricardo Suyama, Leandro Elias Paiva Rangel, Tiago Macedo Dias and João Marcos Travassos Romano are with the Laboratory of Signal Processing for Communications (DSPCom) – FEEC – University of Campinas (Unicamp) – CP 6101 – CEP 13083-970 – Campinas – SP – Brazil (e-mails: {rsuyama, romano}@dmo.fee.unicamp.br, {leandroe, tdias}@decom.fee.unicamp.br)

Aline Neves is now with the Engineering, Modeling and Applied Social Sciences Center, Federal University of ABC, Brazil (e-mail: aline.neves@ufabc.edu.br)

Leonardo Tomazeli Duarte is with GIPSA Lab, Domaine Universitaire BP 46 38402 Saint Martin d'Hères Cedex (e-mail: leonardo.duarte@gipsa-lab.inpg.fr)

Cynthia Cristina Martins Junqueira is with the Institute of Aeronautics and Space (IAE) – Aerospace Technical Center (CTA) (e-mail: cynthia@iae.cta.br)

ISSN 1980-6604/$20.00 © 2007 SBrT/IEEE

shown in Fig. 1, the transmitted signal[1], $s(n)$, is sent through a channel and received as a distorted version, $x(n)$. This received signal is then processed by the equalizer and yields an output signal $y(n)$. At this point, it would be quite natural to consider what kind of output signal we should expect to produce at the equalizer output. This is a crucial issue, and, therefore, it is convenient to analyze its multiple aspects in separate.

A. The Zero-Forcing Equalizer

A first approach to the equalization problem is to consider in a very direct manner the idea outlined in the preliminary discussion above: the notion of an equalizer *as the inverse model of a channel*. Let us assume a SISO linear channel model, i.e.:

$$x(n) = \sum_{k=-\infty}^{\infty} h(k)s(n-k) = h(n) * s(n) \tag{1}$$

where $h(n)$ is the channel impulse response and '$*$' corresponds to the discrete-time convolution operator. In addition to that, let us suppose that the equalizer is also a linear device, that is:

$$y(n) = \sum_{k=-\infty}^{\infty} w(k)x(n-k) = w(n) * x(n) \tag{2}$$

where $w(n)$ is the equalizer impulse response. Substituting (2) into (1), we get to:

$$\begin{aligned} y(n) &= w(n) * x(n) = w(n) * \left[h(n) * s(n) \right] = \\ &= \left[w(n) * h(n) \right] * s(n) = c(n) * s(n) \end{aligned} \tag{3}$$

where $c(n)=w(n)*h(n)$ denotes the *combined response (channel+equalizer)*. Now, if we want the equalizer to be the inverse of the communication channel, it is necessary to ensure that

$$h(n) * w(n) = \alpha\delta(n-\tau) \tag{4}$$

Under these circumstances, it holds that

$$y(n) = \alpha s(n-\tau) \tag{5}$$

which, as expected, means that the transmitted signal is perfectly recovered up to a gain α and a delay τ, the so-called *equalization delay*. Since, as shown in (5), the combined response $c(n)$ is zero in all instants except for $n=\tau$, the condition defined by equations (4) and (5) is referred to as *zero-forcing (ZF) condition* [1].

The ZF condition is perhaps the most direct expression of the intuition behind the equalization problem. When a ZF solution is attained, the characteristics of both channel and equalizer are precisely inverse, and, as a consequence, the desired information is perfectly recovered. Nevertheless, the employment of this condition in the process of establishing a design criterion suffers from some difficulties. In order that the optimal equalizer impulse response $w(n)$ be obtained, it is necessary that the channel impulse response be perfectly known, which is, in practice, a stringent requirement. Moreover, the deterministic nature of the ZF solution excludes the effects caused by additive noise, present in most real-world applications.

Finally, the ZF approach is founded on a tacit assumption: that the chosen equalizer is capable, from a structural standpoint, of inverting the channel model. This is a relevant issue, because, in practice, the most popular channel and equalizer models are linear FIR (Finite Impulse Response) filters. In such a case, it is possible to show that a ZF solution is attainable only in two cases [2]: a) in the trivial case of a channel that is a simply gain and b) in the impractical case of an equalizer with an infinite number of coefficients in both the causal and anticausal parts of its impulse response.

B. The Wiener Equalizer

In the ZF approach, the focus is on the ability of the equalizer to compensate for the structural features of the channel. Now, we will turn our attention to a different strategy: to use the statistical information contained in the received and transmitted signals in order to obtain an efficient filtering device. This is the essence of the Wiener approach. Following this approach, noise can be easily taken into account, differently from the ZF method, generally leading to a superior performance.

Let us return to equation (5). This expression summarizes the final objective of any communication system: to reconstruct, at the receiver, a reliable version of the transmitted message. Suppose that it is not possible, for some reason (e.g. structural limitations and/or the presence of noise), to operate in a condition *exactly* like the one described by (5); it would then be natural to choose the parameters of the equalizer accordingly with the goal of reaching a condition *as close as possible* to (5). Therefore, in the Wiener approach, we define a *desired signal*:

$$d(n) = s(n-\tau) \tag{6}$$

and adjust the equalization device in order that its output be as close as possible, in a statistical sense, to (6). This idea of statistical closeness can be translated into mathematical terms by taking the expected value of the square of the error signal, defined as the difference between the desired output and the actual one (the *mean square error*, or *MSE*):

$$J_{Wiener} = E\left[e^2(n) \right] = E\left[\left(d(n) - y(n) \right)^2 \right] \tag{7}$$

where $E[.]$ denotes the *ensemble average operator* and $e(n)$ the error signal.

[1] In this work, we will consider only discrete-time zero-mean wide-sense stationary signals, which will be assumed to depend on a time index *n*.

The next step is to look for the best solution, i.e., the equalizer that minimizes the cost function (7). Considering that the equalizer is a linear device, (2) can be rewritten as

$$y(n) = \mathbf{w}^T(n)\mathbf{x}(n) \tag{8}$$

where $\mathbf{w}(n)=[w(0)\ w(1)\ \ldots\ w(N\text{-}1)]^T$ is the *equalizer parameter vector*, $\mathbf{x}(n) = [x(n)\ x(n\text{-}1)\ \ldots\ x(n\text{-}N\text{+}1)]^T$ is the *input vector*, N is the number of free coefficients, and $(.)^T$ denotes the transpose. Under these hypotheses, it is possible to show that the above cost function becomes [1]:

$$J_{Wiener} = \mathbf{w}^T\mathbf{R}\mathbf{w} - \mathbf{w}^T\mathbf{p} - \mathbf{p}^T\mathbf{w} + \sigma_d^2 \tag{9}$$

where

$$\mathbf{R} = E\left[\mathbf{x}(n)\mathbf{x}^T(n)\right] \tag{10}$$

is the *correlation matrix*,

$$\mathbf{p} = E\left[d(n)\mathbf{x}(n)\right] \tag{11}$$

is the *cross-correlation vector* and σ_d^2 is the variance of the desired signal. For a given $d(n)$, this function is a paraboloid and possesses a single minimum, called *Wiener solution*. This solution is defined by

$$\mathbf{w}_{opt} = \mathbf{R}^{-1}\mathbf{p} \tag{12}$$

This analysis reveals that the optimal equalizer in the MSE sense can be obtained from the second-order statistics of the received signal (contained in **R**) and the cross-correlation between the input signal and the desired signal (given by **p**). This is the typical path to solve the *supervised* equalization problem, i.e., that of finding the optimal equalizer with the aid of a desired signal.

Although the fundamentals of the Wiener approach are given above, there remains an important aspect to be discussed: how should the desired signal be chosen in a typical equalization problem? The question, which is deceptively simple, must be considered in some detail. As we saw in (6), to recover the transmitted signal up to a delay is acceptable in the equalization framework, as a simple delay does not corrupt the information contained in the signal of interest. Therefore, all choices of the kind $d(n) = s(n\text{-}\tau)$ are equally acceptable *a priori*. Nonetheless, a more detailed analysis shows that different delays give rise, as a rule, to different Wiener solutions, and, moreover, to solutions associated with different residual MSEs. In other words, the choice of the desired equalization delay may either facilitate or compromise the task of finding the optimal Wiener equalizer. Furthermore, the existence of multiple possible delays must be taken into account when comparisons are established between the Wiener approach and the unsupervised strategies that will be discussed later [3][4].

The "Wiener recipe" can be considered, in a certain sense, the cornerstone of the entire optimal filtering theory. Notwithstanding, it is a theoretical construction founded on a somewhat complicated hypothesis: that it is possible to have access to samples of the transmitted signal at the receiver. In some systems, this is not necessarily a problem, because information of this nature may be required for other purposes (e.g. synchronization). However, the idea of transmitting periodic training sequences known at the receiver may be impractical or undesirable from a practical standpoint. This fact is the main reason for interest in *unsupervised* or *blind* approaches.

C. Blind Equalization and Linear Prediction

A *blind* or *unsupervised* equalization criterion can be defined as one that makes exclusive use of statistical information about the transmitted signal to guide the choice of the free equalizer parameters, i.e., that does not rely on knowledge of the *transmitted samples*. In digital communications, the general characteristics of the transmitted messages are deeply related to intrinsic features of the communication system itself [2] (e.g. the employed modulation and the existence of interleaving), which indicates that it should not be a problem for the receiver to have perfect knowledge of their *statistical properties*. Consequently, the concept of using information of this sort to build an effective method for mitigating the distortion introduced by the channel is, in principle, a sound one.

Since the Wiener approach, which is based on second-order statistics, constitutes a noticeable compromise between effectiveness and mathematical simplicity, it would be desirable to keep our efforts to build a blind equalization criterion within its bounds. In order to do so, it is important that we turn our attention to another classical filtering problem: that of *prediction*. The problem of *prediction* can be stated as follows: given the present and past samples of a given time-series, build a filter, called *predictor*, to process them and produce an estimate of its future values. Mathematically, the input-output relationship of a predictor is:

$$y(n) = F_{pred}\left[x(n), x(n-1), \ldots, x(n-N+1)\right] \tag{13}$$

where $F_{pred}[.]$ represents a generic input-output mapping of order *N-1*. If the predictor is a linear filter, we may resort once more to equation (8).

In the Wiener framework, the problem of finding the parameters of the predictor corresponds to the task of finding the Wiener solution for $d(n) = x(n+1)$. From (12), we obtain

$$\mathbf{w}_{pred} = \mathbf{R}^{-1}\mathbf{r} \tag{14}$$

where $\mathbf{r} = E\left[x(n+1)\mathbf{x}(n)\right]$. Thus, a filter with parameters equal to $\mathbf{w}_{pred}$ produces the best linear estimate (in the minimum MSE sense) of a given signal from its past samples. It is also important to define the concept of *prediction-error filter*, a device whose output is the *prediction error* $e(n) = x(n+1) - y(n)$, illustrated in Fig. 2.

It is demonstrable that the prediction-error filter produces an error *as white as possible*, i.e., that the samples of $e(n)$ will tend to be temporally uncorrelated if the order of the employed predictor is sufficiently large [1]. The importance of this fact cannot be overlooked, for, in most theoretical analysis, it is assumed that the transmitted signal is composed of i.i.d. (Independent and Identically Distributed) samples, i.e., that $s(n\text{-}a)$ and $s(n\text{-}b)$ are two independent random variables when $a \neq b$. Since independence implies uncorrelatedness, it is tempting to ponder that, by whitening the received signal, a prediction-error filter may work as an efficient equalizer. This argument, however, is not as sound and complete as it may seem at first sight.

In the frequency domain, a white signal is conceived as one that has a constant power spectral density. Accordingly, the notion of "whitening a signal" can be understood as a filtering process that moulds its power spectrum into a pattern as close as possible to that of a white signal. Let us assume that an i.i.d. signal $s(n)$ is transmitted through a channel whose frequency response is $H(f)$, and received as a distorted version $x(n)$. In this case, the power spectral densities of the two signals are related by the formula [5]:

$$S_x(f) = \left|H(f)\right|^2 S_s(f) \tag{15}$$

where $S_x(f)$ and $S_s(f)$ are, respectively, the power spectral densities of $x(n)$ and $s(n)$. It is important, at this point, to recall that the power spectral density of a random process is the Fourier transform of its autocorrelation function, i.e., the Wiener-Khintchine theorem.

The expression shows that the power spectral density (and, consequently, the autocorrelation) *do not take into account the phase information of the channel frequency response*, which means that the whitening process is able to reduce or even eliminate the *amplitude distortion* introduced by a given channel, but, on the other hand, is not influenced by its *phase distortion*. Consequently, an effective prediction-error filter will counterbalance very well the channel amplitude response, but will not be reliable insofar as phase distortion is concerned. As a matter of fact, there is only one class of channels that can be adequately equalized by prediction-error channels: that of *minimum-phase* channels. This is entirely justifiable, since the property of minimum-phase implies that the amplitude and phase responses of these channels are uniquely related [6].

The prediction approach allowed us to build a blind equalizer based exclusively on second-order statistics. However, it has a very serious limitation: it is effective only for minimum-phase channels. In the most general case, *whitening does not mean equalizing*, because there will remain a phase distortion to be dealt with. This line of reasoning leads us to an inevitable conclusion: *second-order statistics are not, in general, sufficient to perform blind equalization*. It is necessary to consider other sources of information, sources that propitiate an adequate treatment of the phase distortions introduced by the communication channel.

D. Blind Equalization Theorems

Since second-order statistics are not sufficient to perform blind equalization in a general context, it is possible to conceive a direct improvement: to use statistical features of order higher than two, which allow, in principle, that the channel response be completely characterized. Based on this knowledge, Benveniste, Goursat and Ruget (BGR) [8] and, afterwards, Shalvi and Weinstein (SW) [9], defined the necessary conditions to achieve ideal equalization in an unsupervised context.

Firstly, let us consider the following conditions: 1) the samples of the transmitted signal are i.i.d.; 2) the channel and the equalizer are linear filters and, moreover, there is no additive noise; 3) it is possible to invert the channel perfectly, i.e., it is possible to attain the ZF solution. After establishing the necessary assumptions, it becomes necessary to address a pair of particularly relevant questions: how to define an appropriate criterion to find the channel inverse without resorting to supervised training? On what characteristics of the involved signals should this criterion be based?

The first answer to these topics was given by Benveniste, Goursat and Ruget [8]. Their goal was to use a comparison between the probability density function (pdf) of the equalizer output and that of the transmitted signal as the basis of a criterion devised to verify if an ideal (zero-forcing) condition had been attained. The use of a pdf as the source of the required statistical information is both intuitive and clever: implicitly, the pdf contains all the statistics of the random variable it describes. At this point, we must emphasize the fact that the validity of the above comparison is subject to an important assumption: the pdf of the transmitted signal, $p_s(s(n))$ must be non-Gaussian. Since Gaussian processes filtered by a non-trivial system remain Gaussian [5], the comparison would be reduced to a power adjustment [8].

The Benveniste-Goursat-Ruget (BGR) theorem is simply a translation of these arguments into mathematical terms:

Theorem 1 (BGR Theorem) *Under conditions 1, 2 and 3, if the probability density functions of the transmitted signal (supposed to be non-Gaussian) and of the equalizer output are equal, then the ZF solution is necessarily attained, i.e.,* $c(n) = \pm\delta(n\text{-}\tau)$.

The simplicity of this theorem must not overshadow its importance. We have just exposed a criterion capable of describing a condition of perfect equalization *without mentioning either a training sequence or the channel model*. This is the reason why the BGR theorem is considered to be the first important theoretical result of blind equalization theory.

Ten years later, Shalvi and Weinstein proposed a theorem [9] that can be conceived as a refinement of the ideas of Benveniste, Goursat and Ruget. Under the same conditions, the authors arrive at a blind expression of the ZF condition with the help of a less restrictive amount of information concerning the involved signals. The key to understanding how this is possible is the statistical concept of *kurtosis*.

Before presenting it, however, we need to define what is a *cumulant.*

Cumulants are statistical measures derived from the characteristic function [7]. We denote an order *p* cumulant of a real-valued random variable ξ as C_p^{ξ}, which is equivalent to writing *cum*(ξ : *p*). Until third order, the cumulants are equal to the moments of a zero-mean random variable. Thus, C_2^{ξ}, e.g., for a zero mean signal, is equal to its variance. The fourth-order cumulant, C_4^{ξ}, is called *kurtosis*. Its definition based on moments is given by:

$$K(\xi) = E\left[\xi^4\right] - 3E^2\left[\xi^2\right] \tag{16}$$

Theorem 2 (SW Theorem) *Under conditions 1, 2 and 3, and considering* $K\left[s(n)\right]$ *to be nonzero, if* $E\left[y^2(n)\right] = E\left[s^2(n)\right]$ *and* $\left|K\left[y(n)\right]\right| = \left|K\left[s(n)\right]\right|$*, then, necessarily, c(n)=* $\pm\delta(n-\tau)$.

Even though the theorem above is stated considering real random variables, it is also valid for complex signals, as long as conditions 1, 2 and 3 are satisfied.

The BGR theorem associated the "blind expression" of the ZF condition to a sort of "pdf matching", that is, to an implicit comparison between *all moments* of the equalized and transmitted signals. The beauty of the SW theorem lies in that it fulfills the same task using only two statistics: the variance and the kurtosis. Whereas the former statistic is simply responsible for a sort of "power adjustment", the latter statistic, which is of higher-order, bears the information that gives support to the process of channel inversion.

The theorems we have just presented are the pillars of the blind equalization theory. However, although they are of paramount importance as means of expressing the fundamental objective of the equalization task in suitable terms, it cannot be said that they explicitly indicate how the parameters of a given filter can be chosen in a real-world situation. Since this is the main concern of any communication engineer, it is not surprising that the research field of blind equalization also evolved through a different path: that of finding *blind criteria* which operate in a manner analogous to that which was previously referred to as "Wiener recipe". We will analyze two classes of these criteria: that of *Bussgang techniques*, which, although conceptually related to the theorems presented above, is formed by techniques derived, as a rule, from *ad hoc* motivations and conceptions; and that of the Shalvi-Weinstein methods, which, as the name suggests, are based on the SW theorem.

E. Bussgang Techniques

A blind equalization criterion must necessarily provide the means to choose the parameters of an equalizer by resorting exclusively to statistical characteristics of the input sequence. In the so-called *Bussgang methods*, this crucial information comes from the use of a zero-memory nonlinear function whose role is to produce an estimate of the unavailable transmitted signal [10]. Typically, one considers the use of this estimate, to which we shall refer to as *g(y(n))*, in the context of an LMS-like adaptation scheme. However, it is not unusual to find the idea also applied in a sort of Wiener-like cost function as

$$\begin{aligned} J_{Buss} &= E\left[e^2(n)\right] = E\left[\left(\hat{s}(n) - y(n)\right)^2\right] \\ &= E\left[\left(\varphi\left(y(n)\right) - y(n)\right)^2\right] \end{aligned} \tag{17}$$

where $\hat{s}(n)$ is the estimated transmitted symbol and $\varphi\left(y(n)\right)$ is a nonlinear function. Note that J_{Buss} is a nonconvex cost function, which means that it may have local minima in addition to global minima[2].

Historically, the motivations for choosing the nonlinear function were related to specific applications and theoretical insights. Nonetheless, before we consider these particular instances, it is interesting to discuss how $\varphi\left(y(n)\right)$ could be chosen in a more systematic way. In this sense, we start by defining convolutional noise. Considering *w*(*n*) as the impulse response of the correct inverse filter and $\hat{w}(n)$ the impulse response of the approximate inverse filter, obtained, for example, through the optimization of (17), we can write the equalizer output as

$$y(n) = s(n) + \eta(n) \tag{18}$$

where $\eta(n)$ is the convolutional noise, which represents the residual intersymbol interference and is given by

$$\eta(n) = \left(\hat{w}(n) - w(n)\right) * h(n) * s(n) = c_e(n) * s(n), \tag{19}$$

that is, the convolution of the input data with the residual inverse filter error. If the impulse response $c_e(n)$, which corresponds to the convolution of the channel impulse response and the residual inverse filter error, is long enough, then, inspired by the central limit theorem [5], we may model $\eta(n)$ as a white Gaussian noise. It is clear from (18) and (19) that $\eta(n)$ is correlated with *s*(*n*), at least in the beginning of an adaptive procedure. However, it is possible to show that this correlation is negligible compared to the variance of $\eta(n)$, and so $\eta(n)$ can be considered orthogonal to the data sequence *s*(*n*) [10][11]. We also suppose the data symbols *s*(*n*) to be uniformly distributed with zero mean and unit variance.

Returning to the non-zero memory nonlinear function, in the context defined above, the requirement is to derive an estimate of *s*(*n*) that is optimal in some statistical sense. A sensible and robust choice for *g*(*y*(*n*)) is the conditional mean [1][11]

$$\hat{s}(n) = \varphi\left(y(n)\right) = E\left[s(n) \mid y(n)\right] \tag{20}$$

[2] The function also contains a maximum and, in general, saddle points [2].

which minimizes the root mean-square error between the actual transmission $s(n)$ and the estimation $\hat{s}(n)$. It is important to remark that, although the estimate (20) is optimal in the mean-square error sense, it is, in practice, suboptimal due to the simplifying hypotheses, specially those regarding $\eta(n)$.

From *Bayes'* rule, we have

$$f_s\left(s(n)\,|\,y(n)\right)=\frac{f_y\left(y(n)\,|\,s(n)\right)f_s(s(n))}{f_y\left(y(n)\right)} \tag{21}$$

where $f_s\left(s(n)\right)$ and $f_y\left(y(n)\right)$ are the pdf of $s(n)$ and $y(n)$ respectively and $f_y\left(y(n)\,|\,s(n)\right)$ is the conditional pdf of $y(n)$ given $s(n)$. We may then rewrite (20) as

$$\hat{s}(n)=\frac{1}{f_y\left(y(n)\right)}\int_{-\infty}^{\infty}sf_y\left(y(n)\,|\,s\right)f_s\left(s\right)ds \tag{22}$$

Once $f_y\left(y(n)\right)$, the denominator of (22), has been evaluated, another useful expression for the conditional mean is [11]

$$\varphi\left(y(n)\right)=E\left[s(n)\,|\,y(n)\right]=y(n)-\sigma^2 b\left(y(n)\right), \tag{23}$$

where $\sigma^2=E\left[\eta^2(n)\right]$ is the convolutional noise variance and $b\left(y(n)\right)=\dfrac{-\left.\dfrac{df_y(y)}{dy}\right|_{y=y(n)}}{f_y\left(y(n)\right)}$.

Although this methodology is a systematic and rational approach to the problem, we must not forget that at its core a number of hypotheses were assumed that may render any practical application significantly difficult. This remark indicates why the statistically sound approach we have just described has not reduced the interest for the more idiosyncratic and intuitive choice of $\varphi\left(y(n)\right)$ associated with the most celebrated members of the Bussgang class.

After choosing the nonlinear function according to some methodology, it is straightforward to obtain an adaptive algorithm to optimize (17) with the help of the usual stochastic approximation of the gradient vector:

$$\mathbf{w}(n+1)=\mathbf{w}(n)-\mu\mathbf{x}(n)\left(y(n)-g(y(n))\right) \tag{24}$$

where μ is the step size and $g(.)$ is the nonlinearity resulting from the stochastic approximation of (17). We can readily see that (24) differs from the classical LMS (Least Mean Square) algorithm [1] only in the nonlinear function $g(y(n))$, which plays the role of a "blind reference signal". The use of a nonlinear mapping as the desired signal is what characterizes a Bussgang algorithm: expression (24) can thereby be considered to define the general form of a technique of this class.

As we have anticipated, a number of important approaches to the blind equalization problem fit the mould of a Bussgang technique, of which are specially relevant for us the decision-directed (DD) algorithm [1], the Sato algorithm [12] and the Godard / constant modulus (CM) algorithm [13]. In Table I, we list these algorithms together with their correspondent nonlinearities $g(.)$.

F. Shalvi-Weinstein Methods

The Bussgang algorithms are undoubtedly a prominent part of the *corpus* of unsupervised equalization techniques. However, any discussion on the subject of blind deconvolution would be incomplete without another branch of techniques: the class of the Shalvi-Weinstein methods.

Writing the equalizer output as

$$y(n)=\sum_{k=-\infty}^{\infty}c(k)s(n-k) \tag{25}$$

where $c(n)$ is the combined channel-equalizer response, the cumulant of $y(n)$ can be obtained as (considering $p>2$):

$$\left|C_p^{y(n)}\right|=\left|C_p^{s(n)}\right|\left|\sum_{k=-\infty}^{\infty}c^p(k)\right|\le\left|C_p^{s(n)}\right|\left|\sum_{k=-\infty}^{\infty}c^2(k)\right| \tag{26}$$

Following Theorem 2, the first condition for obtaining a ZF solution is $E\left[y(n)^2\right]=E\left[s(n)^2\right]$. Using this condition in (26), we find that $\sum_{k=-\infty}^{\infty}c^2(k)=1$, which means that $\left|c(k)\right|\le 1$. Therefore, the last inequality of (26) will become an equality only if the vector of the combined channel-equalizer response has a single nonzero element with magnitude equal to one (i.e., a ZF solution).

Based on (26) and on the discussion above, the criterion proposed by Shalvi and Weinstein is to maximize $\left|C_p^{y(n)}\right|$, with $p>2$ (higher-order cumulant), subject to the restriction $E\left[y(n)^2\right]=E\left[s(n)^2\right]$ [9][14]. The adopted values for p are $p=3$ (third-order cumulant) and $p=4$, which results in the kurtosis. The latter is used more often because the third-order cumulants of symmetric distributions are zero.

The power restriction may also be substituted by a normalization what gives rise to the following criterion: maximization of $\left\{\left|C_{2p}^{y(n)}\right|\Big/\left(C_2^{y(n)}\right)^p\right\}$. This criterion is used for the derivation of the super-exponential algorithm, also proposed by Shalvi and Weinstein [15]:

$$\begin{aligned}\tilde{\mathbf{w}}&=\mathbf{R}^{-1}\mathbf{d}\\ \mathbf{w}&=\frac{\tilde{\mathbf{w}}}{\tilde{\mathbf{w}}^T\mathbf{R}\tilde{\mathbf{w}}}\end{aligned} \tag{27}$$

where **R** is a matrix whose elements are given by

$$\mathbf{R}_{ij} = \frac{E\left[x(n-j)x(n-i)\right]}{E\left[s(n)^2\right]} \tag{28}$$

and **d** is a vector whose elements are given by

$$\mathbf{d}_i = \frac{cum(y(n):2p-1, x(n-i))}{C_{2p}^{s(n)}} \tag{29}$$

The term $\mathbf{R}^{-1}$ corresponds to a whitening operation. Due to this term, when compared to Bussgang algorithms, the super-exponential algorithm converges faster, at the cost of being computationally more complex. In [15], it is shown that this algorithm has an optimum step size with respect to convergence speed. In addition, it shows that the method can also be viewed as a gradient optimization scheme of a criterion given by the maximization of the ratio $\frac{\left(\sum_k c(k)^{2p}\right)}{\left(\sum_k c(k)^2\right)^p}$, a criterion known as *Donoho's criterion* [17]. This ratio can also be obtained by the manipulation of the normalized SW criterion, writing it as a function of the cumulants of $s(n)$ [14], highlighting the close relation between these two criteria.

G. SIMO/MISO/MIMO channel models

The development of the already presented criteria was carried out under the auspices of a SISO linear channel model. In some cases, though, it may be interesting to consider models with more inputs and/or outputs, which can be viewed as direct extensions of the SISO case. For example, in a wireless communication application, one can distinguish three different scenarios [2]:

1) A single antenna is used by the transmitter and a single antenna is present in the receiver – a SISO channel, as in the previous discussion;

2) A single antenna is employed by the transmitter, whereas two or more antennas are used in the receiver – a single input / multiple output (SIMO) channel – or, conversely, multiple antennas are used by the transmitter and a single antenna is part of the receiver – a multiple input / single output (MISO) channel;

3) Two sets of antennas are constituents of both the transmitter and the receiver – a multiple input / multiple output channel model.

These possibilities are illustrated in Fig. 3.

Although a detailed analysis of the equalization problem in these scenarios is beyond the scope of this work, they are, nonetheless, important for a particular reason: the perspective they open of devising a situation in which it is important to extract simultaneously *multiple information signals* using some kind of diversity (e.g. that originated by the use of several sensors). Thus, the reader should keep in mind that the study of the multichannel problem can be a solid bridge between the world of SISO equalization we have been considering and that of blind source separation, to which we shall, without further ado, turn our attention.

III. Blind Source Separation

Imagine there are three people talking in a restaurant, and their voices were recorded by three microphones placed at different locations in the room. After recording the voices, you obtain three signals, $x_1(n)$, $x_2(n)$ and $x_3(n)$, each one representing a mixture of the voices of the three people. In this particular case, the problem is to estimate the original voice signals, $s_1(n)$, $s_2(n)$ and $s_3(n)$, based on the mixtures $x_1(n)$, $x_2(n)$ and $x_3(n)$ captured by the microphones. Neither the mixing process nor the samples of the original signals are known *a priori*, which means that the problem must necessarily be solved in a blind fashion. The question is: how can it be done?

This example is known in the literature as the *cocktail party problem*, and illustrates the idea behind the blind source separation problem, also depicted in Fig. 4. Let us consider a set of N signals $s_1(n)$, …, $s_N(n)$, referred to as *sources*, and a set of M received signals $x_1(n)$, …, $x_M(n)$, the *observations.* The main goal of BSS is to provide accurate estimates of the sources based solely on the observed signals.

An important point in the development of methods to perform the signal separation is related to the mixing process model. It would be very interesting to deal with a general model in order to obtain a flexible algorithm, capable of performing BSS in wide range of scenarios. Due to the complexity of the problem, we shall consider the linear instantaneous model, i.e.

$$x_i(n) = \sum_{j=1}^{N} a_{ij} s_j(n) \tag{30}$$

or, in matrix form:

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) \tag{31}$$

where **A** is an NxM matrix, and

$$\begin{aligned} \mathbf{s}(n) &= \left[s_1(n) \quad s_2(n) \quad \cdots \quad s_N(n)\right]^T \\ \mathbf{x}(n) &= \left[x_1(n) \quad x_2(n) \quad \cdots \quad x_M(n)\right]^T \end{aligned} \tag{32}$$

denote the vectors containing the sources and the observations.

Under these assumptions, it seems reasonable to consider a separating system consisting of a matrix **W** such that

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n) \tag{33}$$

is a vector with accurate estimates of the sources. In the absence of noise, **W** should have a configuration that is the

inverse of the mixing system, i.e., $\mathbf{W} = \mathbf{A}^{-1}$, an idea similar to that behind the ZF equalization. Nevertheless, it would be equally satisfactory if one could recover a scaled version of the source vector, or even a permutation of it. These conditions can be expressed in mathematical terms as

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n) = \mathbf{PDs}(n) \tag{34}$$

where **P** and **D** denote a permutation and an invertible diagonal matrix, respectively. Thus, any valid BSS criterion should lead to a solution in accordance to (34).

In order to recover the sources in a blind fashion, likewise in the equalization problem, the BSS method must rely on statistical information about the sources. A hypothesis that proved itself valid in a large number of applications is that the source signals are mutually statistically independent. In fact, this simple assumption is the key to the solution of the BSS problem, as it will be discussed in the following.

A. PCA – Principal Component Analysis

A first attempt to solve the problem of determining the separating matrix **W** would be to explore the second-order statistics of the signals, an idea similar to the one described in section II.C. Since the sources are assumed to be independent, the source vector components are also uncorrelated, i.e.:

$$E\left[\mathbf{s}(n)\,\mathbf{s}^T(n)\right] = \mathbf{\Lambda} \tag{35}$$

where $\mathbf{\Lambda}$ is a diagonal matrix. Thus, one could be tempted to propose a criterion for adapting **W** in accordance with the idea of forcing **y**(*n*) to have uncorrelated components. This approach brought to the BSS community a well-known statistical tool – the *principal component analysis (PCA)*. In its original form, PCA can be understood as a strategy devised to find a set of variables smaller than and yet representative of a larger amount of multivariate measurements – a method for redundancy removal in data analysis.

The first step to implement a PCA technique is to subtract the mean from **x**(n):

$$\mathbf{x}(n) = \mathbf{x}(n) - E\{\mathbf{x}(n)\} \tag{36}$$

After a zero-mean vector is obtained, it is necessary to find a linear transformation that, when applied to it, produces a vector **y**(*n*) whose *M* elements are uncorrelated**.** In geometrical terms, the transformation is responsible for generating a rotated orthogonal coordinate system in which the elements of **x**(*n*) are uncorrelated. In simple terms, we obtain projections of **x**(*n*) on the new coordinate axes in a way that the projection on the first axis corresponds to the maximal variance, the projection on the second axis corresponds to the maximal variance orthogonal to the first axis, and so on.

An alternative interpretation of the process arises when it is assumed that the main goal of PCA is to find a projection such as the average error between **x**(*n*) and the projection of **x**(*n*) on the previously found subspace is minimal [18]. This idea is behind the criterion of mean square error minimization (MSEM):

$$\min_{\mathbf{w}_1} E\left\{\left\|\mathbf{x}(n) - \sum_{i=1}^{m}\left(\mathbf{w}_i^T\mathbf{x}(n)\right)\mathbf{w}_i\right\|^2\right\} \tag{37}$$

where $\boldsymbol{w}_i$ is an orthonormal base of the subspace defined by the vector **y**(*n*). A careful analysis of (37) reveals that the idea is to find a set of orthogonal directions that be the most suitable to represent the available data.

The very structure of the PCA process reveals that the technique is effectively built to find a set of "special directions" determined by the characteristics of the data of interest, directions on which its *principal components* are projected. Being a method based on second-order statistics, PCA seeks these directions under a strict restriction of orthogonality i.e. the *principal components are always orthogonal*. This is exactly the reason why PCA is commonly used in source separation algorithms as a tool *to whiten the components of a vector*.

After having defined the concept of PCA, we are ready to address a crucial problem in blind source separation: that of finding the *independent components* of a random vector.

This problem, as a matter of fact, is the essence of the so-called *Independent Component Analysis* (ICA), the objective of which, as the name already indicates, is to allow that a random vector be decomposed into a set of *independent components*. This motivation establishes a clear contrast with PCA, founded on the weaker assumption of orthogonality between the elements of interest, and, moreover, indicates that the problem to be solved is of a subtler and more complex nature.

B. ICA – Independent Component Analysis

In light of these fundamental hypotheses, it is straightforward to state the ICA problem: find the vector **s**(n), or, equivalently, the matrix **A** that produces the observed signals x_j(n). If the idea is to find the sources, typically a separating matrix **W** is calculated that, when applied to **x**(n), produces a vector containing the sources; this matrix is, in a certain sense, the inverse of **A** up to scaling factors and line permutations. For simplicity, in this section we will consider the absence of noise.

Independent component analysis is, *per se*, simply a method for decomposing a data signal into its independent components (in analogy with the original motivation of PCA). However, it is also, as we have shown with our discussion about the cocktail-party problem, a method for recovering a number of independent signals from mixed versions of them. The latter interpretation of ICA explains why it became the "standard formulation" in blind source separation theory.

We have presented the general ICA problem, but nothing has been said about the methods to solve it. Therefore, in the next sections, the reader will be introduced to a number of different proposals that attempt to explore every possible

mathematical expression of the notion of statistical independence to establish effective separation criteria.

1) Criteria based on Independence

According to the previous discussion, a possible way to obtain the separating system based only on the information contained in the observed samples is to find **W** such that the components of the estimated vector are mutually independent. Thus, it is necessary to find ways to quantify the degree of independence between random variables.

A natural "measure" of dependence is the information theory concept of mutual information. In order to properly define it, let us define the Kulback-Leibler divergence (KLD) [5] between two pdfs, p(x) and q(x), as:

$$D\left(p(x)\|q(x)\right)\stackrel{\Delta}{=}\int p(\xi)\log\frac{p(\xi)}{q(\xi)}d\xi \tag{38}$$

Let us consider a vector of random variables $\boldsymbol{\psi}$, with associated joint pdf, $p_{\boldsymbol{\Psi}}(\boldsymbol{\psi})$, and its marginal pdfs, $p_{\Psi i}(\psi_i)$. The mutual information between the elements of $\boldsymbol{\psi}$ is defined as:

$$I(\boldsymbol{\psi})\stackrel{\Delta}{=}D\left(p_{\boldsymbol{\psi}}(\boldsymbol{\psi})\|\prod_i p_{\psi_i}(\psi_i)\right)=\int p_{\boldsymbol{\psi}}(\boldsymbol{\psi})\log\frac{p_{\boldsymbol{\psi}}(\boldsymbol{\psi})}{\prod_i p_{\psi_i}(\psi_i)}d\boldsymbol{\psi} \tag{39}$$

Bearing in mind the notion of distance brought by the KLD, we can conclude that the mutual information is, in a certain sense, a measure of "distance" between the joint pdf of $\boldsymbol{\psi}$ and the pdf of a related vector with statistically independent components, the pdf of which is the product of the marginal distributions ψ_i. The mutual information can also be defined as

$$I(\boldsymbol{\psi})=\sum_{i=1}^{N}H(\psi_i)-H(\boldsymbol{\psi}) \tag{40}$$

where H(.) denotes Shannon's entropy [20], defined as:

$$H(\psi_i)=-\int p_{\psi_i}(\psi_i)\log p_{\psi_i}(\psi_i)d\psi_i$$

and

$$H(\boldsymbol{\psi})=-\int p_{\boldsymbol{\psi}}(\boldsymbol{\psi})\log p_{\boldsymbol{\psi}}(\boldsymbol{\psi})d\boldsymbol{\psi} \tag{41}$$

A crucial property of the mutual information is that it is always non-negative, and is zero if an only if $p_{\boldsymbol{\psi}}(\boldsymbol{\psi})=\prod_i p_{\psi_i}(\psi_i)$, i.e., if $\boldsymbol{\psi}$ *is composed by statistically independent components*.

Thus, the minimization of the mutual information between the estimates $\boldsymbol{\psi}=\mathbf{y}(n)$ represents a valid criterion to guide the optimization of the separating matrix **W**. As long as the mixing matrix **A** is invertible, it will be possible to recover the sources up to scale and permutation ambiguities.

2) Criteria based on Nongaussianity

The theoretical foundations of the linear BSS problem were laid by Comon [19], who demonstrated that source separation can be performed, under the conditions stated in the beginning of section III, by driving the outputs of a linear system, called separating system, to be statistically independent. On one hand, this idea constitutes the very essence of the ICA-based paradigm in BSS, but, on the other hand, its direct implementation may be extremely inefficient. Algorithms directly based on mutual information (MU), which we have described in section III.B.1, demand a high computational effort. This is the reason why the sue of MU based algorithms requires approximations. This is what has been done with the MU Edgeworth expansion yielding a cumulant-based optimization criterion introduced in [19]. Numerical algorithms based on cumulants are not computationally heavy, contrary to what is sometimes understood (they have a polynomial complexity).

However, there are other means to carry out the separation task. For instance, in [21], a criterion based on nongaussianity was considered. In a very intuitive way, it is possible to understand the principle of the nongaussianity approach in BSS in connection with the central limit theorem (CLT). Since, in summary, the CLT states that a sum of independent random variables tends toward a Gaussian distribution, it is expected that each mixture signal, which is the result of a linear combination of sources, be "more Gaussian" than the sources themselves. Taking this observation into account, a straightforward strategy to BSS is to adjust the separating system *in order to maximize the nongaussianity of its outputs*.

Despite the simplicity of the above justification, the nongaussianity approach is solidly and closely related to the idea of minimizing the mutual information, as shown in [19]. This fact becomes clear after some algebraic manipulation of (40), which leads to:

$$\begin{aligned}I\left(y_1(n),y_2(n),\ldots,y_N(n)\right)&=\\&=\sum_{i=1}^{N}H\left(y_i(n)\right)-H\left(\mathbf{x}(n)\right)-\log\left|\det\mathbf{W}\right|\end{aligned}. \tag{42}$$

From this expression, one can notice that $H(\mathbf{x}(n))$ does not depend on the parameters of the separating system, and, therefore, can be ignored in the optimization task. Furthermore, when the matrix W is restricted to be orthogonal and the variance of $y(n)$ is forced to be constant, the last term of (42) is also constant during the optimization procedure [19], which permits us to conclude that the minimization of the mutual information, in this case, is equivalent to the minimization of the marginal entropies of $y(n)$. Besides, from information theory, it is well-known [20] that the Gaussian distribution is the one with maximum entropy over all distributions with the same variance. Therefore, the maximization of nongaussianity is equivalent to the minimization of the marginal entropies, hence, to the minimization of mutual information – tacitly, the

nongaussianity approach also follows the guideline proposed by Comon: to recover the property of statistical independence.

Naturally, as in the independence-based criteria, we must have a quantitative measure of Gaussianity in the nongaussianity-based approach. In view of what was discussed in section II.F, a natural choice could be the kurtosis [18] of a random variable, since this cumulant assumes a non-zero value for the great majority of random variables and is equal to zero almost exclusively for the Gaussian distribution. Differently from the mutual information approach, the evaluation of kurtosis does not demand any sort of distribution estimation, being thus a more efficient solution from the computational standpoint.

Another classical measure of nongaussianity is the negentropy of a random variable χ, a quantity defined as follows:

$$J(\chi) = H(\chi_{gauss}) - H(\chi) \tag{43}$$

where χ_{gauss} is a Gaussian random variable with the same variance of χ. This measure is always non-negative, being zero only for a Gaussian distribution. In comparison with kurtosis, the sample estimation of the negentropy is more robust to outliers, which explains the preference for negentropy-based criteria in BSS.

At first glance, we notice that the evaluation of negentropy demands probability density estimation, as it requires the evaluation of marginal entropies as in the mutual information approach. Fortunately, it is possible to resort to the following approximation [18]:

$$J(\chi) \propto \left[E\{G(\chi)\} - E\{G(\chi_{gauss})\} \right]^2 \tag{44}$$

where $G(\cdot)$ is a nonquadratic function. Note that if we define $G(\chi) = E\{\chi^4\}$, we obtain a measure of Gaussianity very similar to the kurtosis, which indicates that this sort of approximation is in accordance with the idea of nongaussianity.

3) Maximum Likelihood Estimation in BSS

The work of Bell and Sejnowski [22] was important to popularize the problem of BSS in the signal processing community, which can be chiefly attributed to the fact that their technique was the first to propitiate the recovery of a great number of sources. Later, Cardoso [23] demonstrated that the principle presented in [22], the so-called infomax criterion, is equivalent to the maximum likelihood (ML) estimator for the ICA model, which established the paradigm ML/Infomax as one of the most important techniques in BSS. We devote this subsection to an exposition of the basics of the ML/Infomax approach. As it will be seen in the sequel, the idea behind the ML approach in BSS can be understood in a very illustrative manner using the divergence of Kullback-Leibler.

In the ML approach [23], the estimation of the parameters $\boldsymbol{\theta} = [\theta_1\ \theta_2\ \ldots\ \theta_N]$ from the vector samples $\mathbf{Q} = [\mathbf{q}(1)\ \mathbf{q}(2)\ \ldots\ \mathbf{q}(\tau)]$, is performed through the maximization of the likelihood function $p_Q(\mathbf{Q} \mid \boldsymbol{\theta})$, which corresponds to the joint pdf of the samples $\mathbf{Q}$ conditioned upon the parameters $\boldsymbol{\theta}$. Generally, it is assumed that the samples $\mathbf{q}(n)$ are mutually independent, which permits us to rewrite the likelihood function as follows:

$$L(\boldsymbol{\theta}) = p_Q(\mathbf{Q} \mid \boldsymbol{\theta}) = \prod_{n=1}^{\tau} p_q(\mathbf{q}(n) \mid \boldsymbol{\theta}), \tag{45}$$

where $p_q(\mathbf{q}(n) \mid \boldsymbol{\theta})$ corresponds to the pdf of the sample conditioned to $\boldsymbol{\theta}$.

When ML estimation in applied to the BSS problem, the parameters to be estimated and the samples available are, respectively, the elements of the mixture matrix $\mathbf{A}$ (or equivalently its inverse, denoted by $\mathbf{W}$) and the mixture signals $\mathbf{x}(n)$. Having this in mind, and recalling that $p_{\mathbf{x}}(\mathbf{x}(n)|\mathbf{A}) = p_{\mathbf{s}}(\mathbf{A}^{-1}\mathbf{x}(n))|\det(\mathbf{A}^{-1})|$ [5], it is possible, after some manipulations, to obtain the likelihood function associated with the BSS problem, which is given by:

$$L(\mathbf{W}) = \prod_{n=1}^{\tau} p_s(\mathbf{W}\mathbf{x}(n)) |\det(\mathbf{W})| \tag{46}$$

It is very usual to consider, instead of the likelihood function, its logarithm, which is often called the log-likelihood function and, in our case, is given by:

$$\log L(\mathbf{W}) = \sum_{n=1}^{\beta} \log p_s(\mathbf{W}\mathbf{x}(n)) + \beta \log |\det(\mathbf{W})| \tag{47}$$

Since the logarithm function is monotonic, the maximization of (47) over $\mathbf{W}$ also results in the ML solution of the BSS problem.

From expressions (46) and (47), it is evident that the ML approach in BSS requires *a priori* knowledge of the pdf of the sources, which, in a certain sense, contradicts the essence of ICA, since, as we saw in section III.B, the basic formulation of this problem demands no assumption aside from that of independence between sources. At first sight, this requirement could render unfeasible the ML approach in BSS; however, it is still possible to separate the sources by only considering approximations of the sources pdf. In fact, when an approximate $p_{\tilde{\mathbf{s}}}(\cdot)$ of the sources pdf is considered in expression (47), one readily obtains the cost function associated with the Infomax paradigm. This equivalence was proved by Cardoso [23], who also answered one relevant question resulting from this equivalence [24]: how large can the mismatch between $p_{\tilde{\mathbf{s}}}(\cdot)$ and $p_{\mathbf{s}}(\cdot)$ be in order to guarantee the validity of such an approximation? It was shown that there is a considerable tolerance in this mismatch.

A very illustrative interpretation of the ML approach arises when expression (47) is rewritten in terms of the divergence

of Kullback-Leibler. When β tends to infinity, the log-likelihood function becomes [23]:

$$\log L(W) = -D\left(p_{\mathbf{Wx}}\left(\mathbf{Wx}(n)\right) \| \, p_{\tilde{s}}\left(\tilde{s}(n)\right)\right) + \text{constant} \qquad (48)$$

From this expression, we notice that the maximization of the log-likelihood function is equivalent to the minimization of the divergence of KL between the hypothesized source and the separating mixture output distributions. Given that the divergence expresses an idea of distance between distributions, it is clear that the ML approach in BSS is associated with a distribution matching strategy, i.e., the aim in this paradigm is to adjust the separating matrix **W** in order to obtain an output distribution as close as possible to the hypothesized one.

IV. Conclusions and Perspectives

The objective of this first part was to present the basic concepts and main lines of development that form the core of the theories of blind equalization and blind source separation. The work was structured in order to provide the reader with a representative exposition of the conceptual sequence to which the researchers of both areas are accustomed. We have revisited the most important results in the literature and shown the existing criteria and algorithms in each case. At a first glance, the fact that the equalization problem was predominantly discussed in a SISO context, while the BSS problem is inherently of a MIMO nature could give the impression that they have very little, or even nothing in common. The basic concepts discussed here, however, will be essential in the second part of this paper to show that this assumption is far from being true. It will be seen that it is actually possible to establish a number of connection between both problems and the proposed solutions to them.

Acknowledgment

The authors thank FAPESP and CNPq for the financial support. Our warmest thanks go to Dr. Pierre Comon, from I3S/CNRS France, as well as to the anonymous reviewers for their important comments and suggestions.

.

References

[1] S. Haykin, Adaptive Filter Theory, 3rd edition, Prentice Hall, 1996.
[2] Z. Ding, Y. Li, Blind Equalization and Identification, CRC Press, 2001.
[3] C. R. Johnson, P. Schniter, T. J. Endres, J. D. Behm, D. R. Brown, R. A. Casas, "Blind equalization using the constant modulus criterion: a review", Proceedings of the IEEE, Vol. 86, No. 10, pp. 1927 – 1950, 1998.
[4] R. Suyama, R. R. F. Attux, J. M. T. Romano, M. Bellanger, "Relations entre les critères du module constant et de Wiener", Proceedings of the 19e Colloque GRETSI, Paris, France, 2003.
[5] A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill, 3rd edition, 1991
[6] A. V. Oppenheim, A. S. Willsky, H. Nawab, Signals and Systems, 2nd edition, Prentice Hall, 1996.
[7] C. Nikias, A. Petropulu, Higher Order Spectra Analysis - A Nonlinear Signal Processing Framework, Prentice Hall, 1993.
[8] A. Benveniste M. Goursat, 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.
[9] O. Shalvi, E. Weinstein, "New criteria for blind deconvolution of non-minimum phase systems (channels)", IEEE Transactions on Information Theory, Vol. 36, No. 2, pp. 312-321, 1990.
[10] R. Godfrey, F. Rocca, "Zero memory non-linear deconvolution", Geophysical Prospecting, Vol. 29, pp. 189-228, 1981.
[11] S. Bellini, "Bussgang techniques for blind deconvolution and equalization", in S. Haykin (ed.), Blind Deconvolution, Prentice Hall, 1994.
[12] Y. Sato, "A method for self recovering equalization", IEEE Transactions on Communications, Vol. COM-23, No. 6, pp. 679–682, 1975.
[13] 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.
[14] O. Shalvi, E. Weinstein, "Universal methods for blind deconvolution", in S. Haykin (ed.), Blind Deconvolution, Prentice Hall, 1994.
[15] O. Shalvi, E. Weinstein, "Super-exponential methods for blind deconvolution", IEEE Transactions on Information Theory, Vol. 39, No. 2, pp. 504–519, 1993.
[16] M. Mboup, P. Regalia, "A gradient search interpretation of the super-exponential algorithm", IEEE Transactions on Information Theory, Vol. 46, No 7, pp. 2731-2734, 2000.
[17] D. Donoho, "On minimum entropy deconvolution", in Applied Time Series Analysis II, D. F. Findley, Ed. New York: Academic, 1981.
[18] A. Hyvärinen, J, Karhunen, E. Oja, Independent Component Analysis, John Wiley and Sons, 1 rd edition, 2001.
[19] P. Comon, "Independent component analysis: a new concept?," Signal Processing, Vol. 36, No. 3, pp.287-314, 1994.
[20] T. Cover, J. Thomas, Elements of Information Theory, John Wiley and Sons, 1991.
[21] N. Delfosse, P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach," Signal Processing, Vol. 45, No. 1, pp.59-83, 2005.
[22] A. Bell, T. Sejnowski, "Blind separation and blind deconvolution: an information- theoretic approach", Proceeding of the ICASSP, Detroit, EUA, 1995.
[23] J. -F. Cardoso, "Infomax and maximum likelihood for blind source separation", IEEE Signal Processing Letters, Vol. 4, No. 4, pp. 112 – 114, 1997.
[24] J. -F. Cardoso, B. H. Laheld, "Equivariant adaptive source separation", IEEE Transactions on Signal Processing, Vol. 44, No. 12, pp. 3017 – 3030, 1996.

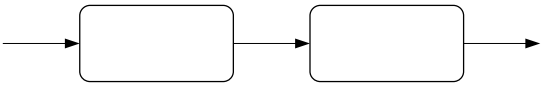

Fig. 1. Simplified Baseband Communication System

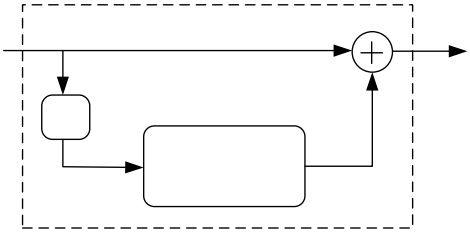

Fig. 2. Predictor and Prediction-error Filter

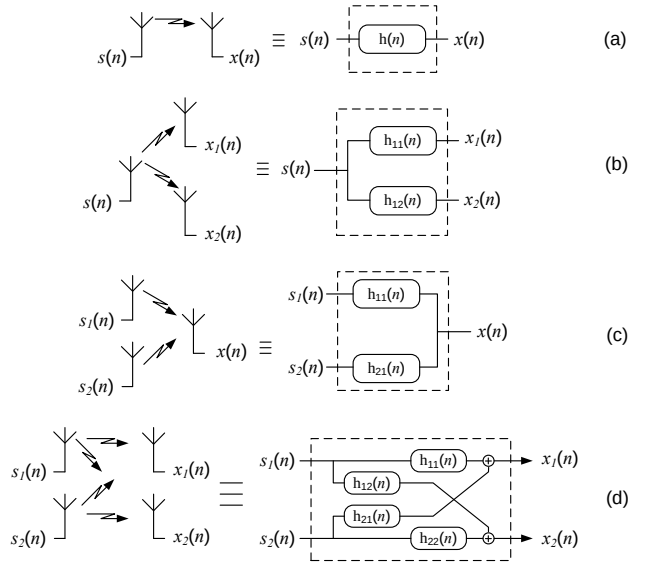

Fig. 3. (a) SISO system, (b) SIMO system, (c) MISO system, (d) MIMO system

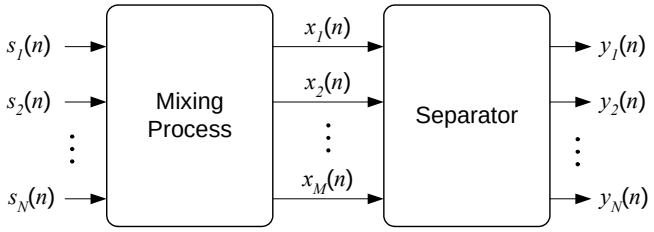

Fig. 4. The problem of blind source separation.

Algorithm	$g(y(n))$	Definitions
DD	$sgn(y(n))$	
Sato	$\gamma sgn(y(n))$	$\gamma = \frac{E\left[s^2(n)\right]}{E\left[\lvert s(n)\rvert\right]}$
CMA	$\frac{y(n)}{\lvert y(n)\rvert}\left(\lvert y(n)\rvert + R_2\lvert y(n)\rvert - \lvert y(n)\rvert^3\right)$	$R_2 = \frac{E\left[\lvert s(n)\rvert^4\right]}{E\left[\lvert s(n)\rvert^2\right]}$

Table I: Comparison between Bussgang Algorithms