BLIND DECONVOLUTION WITH MINIMUM RENYI'S ENTROPY

Deniz Erdogmus[1], Jose C. Principe[1], Luis Vielva[2]
[1] Computational NeuroEngineering Laboratory, NEB 454, University of Florida, Gainesville, Florida 32611, USA
[2] Grupo de Tratamiento Avanzado de Señal, Universidad de Cantabria, Santander, Spain
e-mail: [deniz , principe]@cnel.ufl.edu, luis@dicom.unican.es

ABSTRACT

Blind techniques attract the attention of many researchers due to their numerous promising applications in different fields of signal processing, from communications to control systems. Blind deconvolution is a problem that has been investigated in detail over the last two decades. Many approaches adopting various optimization criteria have been proposed to determine the inverse of the unknown channel filter. Minimum entropy deconvolution, as summarized by Donoho, provides an effective tool for determining the deconvolving filter using only the observed data. Recently, we have proposed an estimator for Renyi's entropy based on Parzen windowing, and demonstrated its superior performance over other entropy estimators in blind source separation and other problems. In this paper, we present a blind deconvolution algorithm based on the minimization of this entropy estimator and investigate its performance through Monte Carlo simulations.

1 INTRODUCTION

Blind techniques have become significantly important in the last two decades, and they have found many applications in signal processing, communications, control, and other fields [1-4]. Blind deconvolution is one such technique where the aim is to determine the inverse filter for an unknown linear filter (*channel*) using only the observations from the output of the channel.

Typically, blind deconvolution is represented by a block diagram as shown in Fig. 1, where both the channel impulse response ***h*** and the input signal is unknown [2]. Donoho summarized the well-known approach of entropy minimization to solve this problem [3]. Minimum entropy deconvolution assumes that the source signal is a non-Gaussian distributed wide-sense-stationary (WSS), white process. Since at the time, effective entropy estimators were not available, the methods summarized by Donoho usually adopted higher order moments, which mimic the properties of entropy, of the signals under investigation, e.g. the kurtosis. In some cases, the designer may have knowledge about some certain statistics of the input signal and this information may be used to obtain better deconvolution results. For example, if the source probability density function (pdf) is known and if the source signal samples are assumed to be *iid* (independent and identically distributed), then the maximum entropy approach may be used [5].

A structural concern in the blind deconvolution problem is the choice between the causal/non-causal equalizer filter structures. Recall that if the unknown channel filter is minimum phase, it will have a stable, causal inverse, whereas, if it is non-minimum phase in general, its stable inverse will be non-causal [4]. In such situations, a sufficiently long delay line may be put before the deconvolving filter, thus allowing us to implement a non-causal filter structure for deconvolution. However, this structural concern need not affect our cost function design and the algorithms used, for once a *suitable* delay line length is chosen, all blind deconvolution problems can be treated the same way.

In the following, we will undertake the minimum entropy deconvolution approach. We will show that Renyi's entropy can be used as a Gaussianity measure as required in [3]. A nonparametric

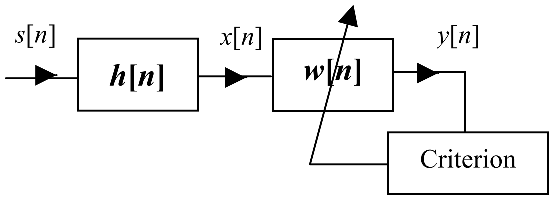

Figure 1. Schematic diagram of blind deconvolution

estimator for Renyi's entropy will also be presented. Motivated by the successful comparisons of this recently proposed entropy estimator in blind source separation and other problems [6-8], we investigate its performance in blind deconvolution.

2 RENYI'S ENTROPY AND MINIMUM ENTROPY BLIND DECONVOLUTION

In this section, we will provide the motivation for using Renyi's entropy as a measure for blind deconvolution. As an initial step, consider the following theorem, which gives the relationship between the entropies of linearly combined random variables.

Theorem 1. Let S_1 and S_2 be two independent random variables with pdfs $p_{S_1}(.)$ and $p_{S_2}(.)$, respectively. Let $H_\alpha(.)$ denote the order-α Renyi's entropy for a continuous random variable. If a_1 and a_2 are two real coefficients in $Y=a_1S_1+a_2S_2$, then

$$H_\alpha(Y) \ge H_\alpha(S_i) + \log|a_i|, \quad i=1,2.$$

Proof: In the appendix.

An immediate extension of this theorem is obtained by increasing the number of random variables in the linear combination to n.

Corollary 1. If $Y=a_1S_1+\ldots+a_nS_n$, with *iid* $S_i \sim p_S(.)$, then

$$H_\alpha(Y) \ge H_\alpha(S) + \frac{1}{n}\log|a_1 \ldots a_n|$$

where equality the two entropies occur if and only if $a_i = \delta_{ij}$, where δ denotes the Kronecker-delta function.

Proof: In the appendix.

Notice that the blind deconvolution problem is structurally very similar to the situation presented in the above corollary. In that context, the coefficients a_i of the linear combination are replaced by the impulse response coefficients of the overall filter ***h*w***. In addition, the random variables S and Y are replaced by the source signal and the deconvolving filter output signal, respectively. Especially, when close to the ideal solution, i.e. when ***h*w*** is close to an impulse, the second term in Cor. 1 will approach rapidly to zero and the two entropy values will converge as the two signals Y and S converge to each other.

3 THE COST FUNCTION

The entropy of a random variable is not scale invariant. In order to be able solve the blind deconvolution problem using unconstrained optimization techniques and without having to normalize the weights at the end of every iteration, we need a scale invariant cost function to minimize. For this purpose, consider the following modified cost function.

Fact 1. The modified cost function

$$J(Y) = H_\alpha(Y) - \frac{1}{2}\log[Var(Y)] \tag{1}$$

is scale invariant. That is $J(aY) = J(Y)$, $\forall a \in \Re$.

Proof: In the appendix.

In practice, one needs to estimate the underlying pdf from the data samples, since in general an analytical expression for it is not available. An effective nonparametric estimator that proved useful in other applications exists for Renyi's entropy. Renyi's entropy for a random variable Y with pdf $p_Y(.)$ is defined as [9]

$$H_\alpha(Y) = \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} p_Y^\alpha(y)dy \tag{2}$$

Alternatively, one can express this with an expectation operator.

$$H_\alpha(Y) = \frac{1}{1-\alpha}\log E_Y[p_Y^{\alpha-1}(Y)] \tag{3}$$

In order to estimate the pdf of Y, we employ Parzen windowing on its N samples, which is a kernel based consistent method [10].

$$\hat{p}_Y(y) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(y-y_i) \tag{4}$$

Here, $\kappa_\sigma(.)$ is the kernel function (a valid pdf) with size σ. Typically, Gaussian kernels are utilized, and in that case, the kernel size becomes the standard deviation of the kernel function.

Approximating the expectation operator with the sample mean and substituting the Parzen estimate in (3), we obtain the nonparametric estimator for Renyi's entropy of order-α [8].

$$\hat{H}_\alpha(Y) = \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(y_j-y_i)\right)^{\alpha-1} \tag{5}$$

In [8], it was established that for smooth, symmetric, unimodal kernel functions, the global minima of the entropy estimator in (5) and the actual entropy in (2) coincide and furthermore, this global minimum of the estimator is smooth, i.e. has zero gradient and a positive semi-definite Hessian (semi-definite, because a zero eigenvalue exists due to the fact that entropy is invariant to the mean of the random variable). This property of the estimator allows gradient and Hessian based optimization procedures to safely converge to the desired optimum point in adaptation.

Lemma 1. Assume the data distribution is not $\delta_Y(.)$. The entropy estimate in (5) as N goes to infinity (or roughly on the average), or equivalently the entropy of the estimated pdf in (4), provides an upper bound to the actual entropy of Y, i.e. $\hat{H}_\alpha(Y) \geq H_\alpha(Y)$. Equality is possible if and only if the kernel size is set to zero.

Proof: Outlined roughly in the appendix.

This lemma allows us to minimize the estimated entropy of the of the data in place of the actual entropy in blind deconvolution, since we will be minimizing an upper bound for a quantity that we wish to minimize.

Practically, the deconvolving filter $\boldsymbol{w}$ is a causal FIR filter, after the addition of the sufficiently long delay line as mentioned before. Assuming the input signal to filter $\boldsymbol{w}$ is x_k at time k, one can express its output as a linear combination of these input samples at consecutive time steps as

$$y_k = w^T X_k \tag{6}$$

where the weight vector $w = [w_0 \ldots w_L]^T$ consists of the FIR impulse response coefficients and $X_k = [x_k \ldots x_{k-L}]^T$ consists of the most recent values of the input signal to the filter.

As for the variance term in (1), under the assumption that the source signal is zero-mean WSS and the unknown channel is time-invariant, we can write

$$Var(Y) = Var(X)\cdot\sum_{i=0}^{L} w_i^2 \tag{7}$$

Substituting (5) and (7) in (1), we get the nonparametric estimate of the cost function as

$$J(w) = \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(y_j-y_i)\right)^{\alpha-1} - \frac{1}{2}\log\sum_{i=0}^{L} w_i^2 \tag{8}$$

where *Var(X)* dropped out because it does not depend on the weights of the adaptive filter. Now, using (6), the gradient of the cost function in (8) with respect to the weight vector is obtained as

$$\frac{\partial J}{\partial w} = -\frac{\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(y_j-y_i)\right)^{\alpha-2}\left(\sum_{i=1}^{N}\kappa'_\sigma(y_j-y_i)(X_j^T-X_i^T)\right)}{\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(y_j-y_i)\right)^{\alpha-1}} - \frac{w^T}{w^T w} \tag{9}$$

where $\kappa'_\sigma(.)$ is the derivative of the kernel function with respect to its argument. Given N samples of X_k, the adaptive filter may be trained to converge to the inverse of the channel. The gradient in (9) may be used in both off-line and on-line training to minimize (8). Choosing a small enough (depending on the computational requirements) window length of N, which may be sliding or non-overlapping, it is possible to estimate the source signal on-line.

As for the optimization techniques that can be applied to obtain the optimal solution, simple gradient descent, conjugate-gradient, Levenberg-Marquardt, or other approaches may be taken [11]. If the kernel size is chosen sufficiently large (usually, a kernel width that covers about 10 samples on the average yields good results), then the performance surface is reasonably simple to search and based on numerous simulations, we conjecture that there does not exist local minima. In fact, in all previous problems, it was observed that as long as the kernel size is chosen in a moderate value range (as prescribed above), its precise value is not crucial to the final performance of the adaptive system. An important property of the proposed estimator that deserves mentioning at this point is its close relationship with the global optimization method of convolution smoothing. The smoothing effect of increasing the kernel size is mentioned and investigated in more detail in [8], where hints of a possible link to convolution smoothing [12] are pointed out in a conjecture.

4 SIMULATIONS

In order to test the performance of the proposed blind deconvolution algorithm, we have performed a series of Monte Carlo runs using different entropy orders and batch-sizes. However, to demonstrate the behavior of the performance surface, we will first consider a single example where the channel is a single pole (at 0.5) IIR filter, thus a 2-tap deconvolving filter is sufficient to perfectly invert the channel. Recall that in the blind deconvolution problem, a scaling indeterminacy exists; hence, one can only determine the inverse filter up to a gain and sign uncertainty. The optimal solution for this specific problem, therefore, lies on the $w_0 = -2w_1$ line. In this example, the cost function is evaluated using N = 1000 samples. In order to demonstrate the effect of the kernel size, we have evaluated the performance surface for two different values, $\sigma = 0.1$ and $\sigma = 1$ for a Gaussian kernel. Fig. 2 depicts the performance surface and

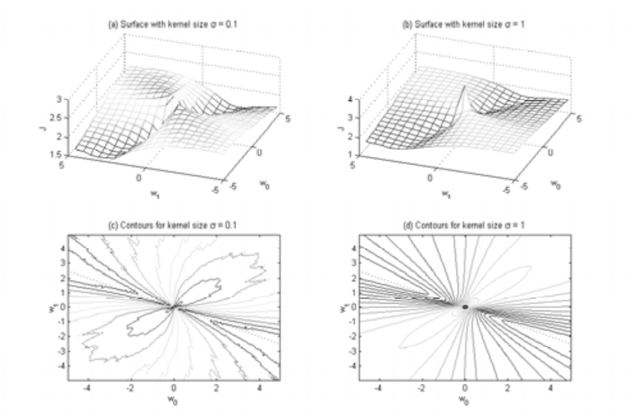

Figure 2. Illustration of the performance surface and its contour lines for a 2-tap deconvolving filter case using two different kernel size values in the entropy estimator.

	$\alpha = 1.01$	$\alpha = 2$	$\alpha = 3$
$N = 50$	21.25 ± 1.55	20.49 ± 1.15	20.45 ± 1.09
$N = 75$	21.45 ± 1.45	21.18 ± 1.08	21.09 ± 1.15
$N = 100$	22.14 ± 1.33	21.99 ± 1.04	21.93 ± 1.05
$N = 200$	22.81 ± 1.72	22.49 ± 1.11	22.30 ± 1.17
$N = 300$	22.70 ± 1.81	22.68 ± 1.20	22.65 ± 1.11
$N = 400$	23.04 ± 1.84	22.45 ± 1.48	22.53 ± 1.58
$N = 500$	23.27 ± 2.50	22.71 ± 1.65	22.87 ± 1.75

Table 1. E[SIR] ± std[SIR] in dB over 100 Monte Carlo runs for each combination after convergence.

	$\alpha = 1.01$	$\alpha = 2$	$\alpha = 3$
$N = 50$	72 ± 22	62 ± 30	62 ± 29
$N = 75$	66 ± 25	64 ± 27	62 ± 29
$N = 100$	65 ± 24	67 ± 28	67 ± 30
$N = 200$	68 ± 25	61 ± 27	62 ± 28
$N = 300$	68 ± 24	64 ± 29	64 ± 30
$N = 400$	70 ± 23	62 ± 29	60 ± 28
$N = 500$	67 ± 25	63 ± 26	64 ± 28

Table 2. E[T_c] ± std[T_c] in iterations over 100 Monte Carlo runs for each combination.

their contour lines for these two kernel size values. Notice that, in both cases, the actual solution denoted by a dotted line lies (almost perfectly) along the global minimum of the cost surface, however, as the kernel size gets larger, the surface is smoothened and local minima are eliminated as expected.

In the second case study, we perform Monte Carlo runs to evaluate the performance of the proposed adaptation algorithm. In the Monte-Carlo runs, a 15-tap FIR filter is chosen for the unknown channel impulse response, and the length of the deconvolving filter is set to that of the ideal inverse filter. For various values of N and α, 100 random-choice (both for Cauchy distributed data samples and deconvolver initial weights) simulations are run for each combination of (N,α). The results of these Monte Carlo simulations are summarized in Table 1 and Table 2, where the average and standard deviations of both signal-to-interference-ratio (SIR) and convergence time (T_c) are given.

The SIR of a single run is defined as the average of the SIR values of the last 100 iterations after convergence of that simulation (since due to the constant step size, the performance rattles slightly after convergence). The SIR value at a given iteration is computed as the ratio of the power of the maximum component of the over all filter to the power of the other components, i.e. if we let $a=h*w$ be the overall filter where w is the current estimate of the deconvolving filter, we evaluate

$$SIR = 10\log_{10}\frac{[\max(a_i)]^2}{\sum_i a_i^2 - [\max(a_i)]^2} \tag{10}$$

Note that under the assumption of WSS source signals, the power of the observed signal is time-invariant, therefore, the overall filter weights can equivalently be utilized to determine the signal-to-interference ratio.

The convergence time is defined as the largest iteration index smaller than the maximum number of iterations minus 100, such that the SIR value is less than or equal to the minimum SIR value attained in the last 100 iterations.

6 CONCLUSIONS

Blind deconvolution is a crucial signal processing technique that has numerous important applications ranging from communications to geophysics. Blind deconvolution has been studied in detail over the last two decades and many algorithms to obtain the deconvolving filter have been proposed. These algorithms adopted information theoretic or higher order statistical measures, measures that are also commonly utilized in the blind source separation context. In correspondence, a nonparametric estimator for Renyi's entropy has been proposed and studied in detail by the authors and successful results were obtained in many problems including supervised adaptive system training, chaotic time-series prediction, etc. Especially in the blind source separation context, it was shown that the proposed entropy estimator outperforms its alternatives in terms of data efficiency and overall solution performance.

Motivated by these recent developments, the same nonparametric Renyi's entropy estimator has been applied to the blind deconvolution problem in this paper. The *minimum entropy deconvolution* approach is followed, and it is shown that the optimal solution of the blind deconvolution problem lies at the minimum of Renyi's entropy (and the proposed estimator). In order to bring the optimization problem into an unconstrained form (in terms of the adaptable weights), a modification is introduced to the cost function to make it scale invariant.

The smoothness of the performance surface and the performance of the adaptation algorithm were validated through the use of Monte Carlo runs. The high average signal-to-interference ratio and its small variance verified that the cost function does not exhibit local minima. Monte Carlo runs also demonstrated that the proposed algorithm converges to the optimal solution in a very small number of iterations, and using a small number of data samples efficiently.

The effect of the entropy order on the convergence properties was also investigated. For the Cauchy-distributed source signal used in the Monte Carlo runs, no value of entropy order significantly outperformed the other values, thus we cannot conclude in favor of a specific entropy order from current results.

The proposed algorithm, nevertheless, still used a batch of data samples in the iterations. Future work will be conducted to introduce a robust and fast-converging stochastic gradient technique that provides weight updates with less computational requirements.

Acknowledgments: This work is partially supported by NSF grant ECS-9900394.

APPENDIX

Proof of Theorem 1 : Since S_1 and S_2 are independent, the pdf of Y is given by

$$p_Y(y) = \frac{1}{|a_1|} p_{S_1}(y/a_1) * \frac{1}{|a_2|} p_{S_2}(y/a_2) \qquad \text{(A.1)}$$

Recall the definition of Renyi's entropy for Y given in (1). Notice that we can write

$$\begin{aligned} e^{(1-\alpha)H_\alpha(Y)} &= \int_{-\infty}^{\infty} p_Y^\alpha(y)dy \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} \frac{1}{|a_1 a_2|} p_{S_1}(\frac{\tau}{a_1}) p_{S_2}(\frac{y-\tau}{a_2}) d\tau\right] dy \end{aligned} \qquad \text{(A.2)}$$

Using Jensen's inequality for convex and concave cases, we get

$$\begin{aligned} e^{(1-\alpha)H_\alpha(Y)} &\underset{\alpha<1}{\overset{\alpha>1}{\lesseqgtr}} \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} \frac{1}{|a_1|} p_{S_1}(\frac{\tau}{a_1}) \left[\frac{1}{|a_2|} p_{S_2}(\frac{y-\tau}{a_2})\right]^\alpha d\tau\right] dy \\ &= \int_{-\infty}^{\infty} \frac{1}{|a_1|} p_{S_1}(\frac{\tau}{a_1}) \left[\int_{-\infty}^{\infty}\left[\frac{1}{|a_2|} p_{S_2}(\frac{y-\tau}{a_2})\right]^\alpha dy\right] d\tau \\ &= \int_{-\infty}^{\infty} \frac{1}{|a_1|} p_{S_1}(\frac{\tau}{a_1}) V_\alpha(a_2 S_2) d\tau \\ &= V_\alpha(a_2 S_2) \cdot \int_{-\infty}^{\infty} \frac{1}{|a_1|} p_{S_1}(\frac{\tau}{a_1}) d\tau = V_\alpha(a_2 S_2) \end{aligned} \qquad \text{(A.3)}$$

where $V_\alpha(X)$ is called the order-α information potential for a random variable X with pdf $p_X(.)$, and is given by [7,8]

$$V_\alpha(X) = \int_{-\infty}^{\infty} p_X^\alpha(x)dx \qquad \text{(A.4)}$$

Notice that the information potential is the argument of the *log* in the entropy definition, and is named after its resemblance to potentials of physical particles [13].

Reorganizing the terms in the last inequality and using the relationship between entropy and information potential, regardless of the value of α and the direction of the inequality, we arrive at the conclusion $H_\alpha(Y) \geq H_\alpha(S_i) + \log|a_i|$, $i=1,2$. □

Proof of Corollary 1 : It is trivial to generalize the result in Theorem 1 to n random variables using mathematical induction. Thus, for the case where all n random variables are identically distributed we get n inequalities.

$$\begin{aligned} H_\alpha(Y) &\geq H_\alpha(S) + \log|a_1| \\ H_\alpha(Y) &\geq H_\alpha(S) + \log|a_2| \\ &\vdots \\ H_\alpha(Y) &\geq H_\alpha(S) + \log|a_n| \end{aligned} \qquad \text{(A.5)}$$

Adding these inequalities, we get the desired result. The necessary and sufficient condition for the equality of entropies is obvious from the formulation. If $a_i = \delta_{ij}$, then Y=S, therefore the entropies are equal. If $a_i \neq \delta_{ij}$, then due to Thm. 1, entropy of Y is greater than the entropy o S (assuming normalized coefficients). □

Proof of Fact 1 : It is trivial to show by a simple change of variables in the integral that for Renyi's entropy (as for Shannon's entropy), we have the following identity between the entropies of two scaled random variables.

$$H_\alpha(aY) = H_\alpha(Y) + \log|a| \qquad \text{(A.6)}$$

where we can replace $\log|a|$ with $(1/2)\log a^2$. We also know that for variance

$$Var(aY) = a^2 Var(Y) \qquad \text{(A.7)}$$

Combining these two identities, the terms with a cancel out and we get the desired result. □

Proof of Lemma 1 : Remember that the expected value of the Parzen window pdf estimate given in (4) is the convolution of the actual pdf underlying the samples and the kernel function (as N goes to infinity, the pdf estimate converges to this value since Parzen windowing is consistent). We can consider the average pdf as the pdf of a random variable, which is the sum of two independent random variables: one with the same pdf as the data samples, the second with pdf equal to the kernel function. If we define S=Y+K, where Y and K correspond to the independent random variables mentioned in the previous sentence, we can apply Thm. 1 to conclude that the entropy of S is larger than either of the entropies of Y and K. □

REFERENCES

[1] S. Haykin (ed), *Unsupervised Adaptive Filtering, Vol .2: Blind Deconvolution*, Wiley, New York, 2000.

[2] S. Haykin (ed), *Blind Deconvolution*, Prentice-Hall Inc., New Jersey, 1994.

[3] D. Donoho, "On Minimum Entropy Deconvolution," in *Applied Time Series AnalysisII*, Academic Press, 1981, pp. 565-609.

[4] A. Benveniste, M. Goursat, G. Ruget, "Robust Identification of a Nonminimum Phase System: Blind Adjustment of a Linear Equalizer in Data Communications," IEEE Trans. Automatic Control, vol. 25, no. 3, pp. 385-399, June 1980.

[5] A. Bell, T. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," Neural Computation vol. 7, pp. 1129-1159, 1995.

[6] K.E. Hild II, D. Erdogmus, J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information", IEEE Signal Processing Letters, vol. 8, no. 6, pp. 174-176, 2001.

[7] D. Erdogmus, K.E. Hild II, J.C. Principe, "Blind Source Separation Using Renyi's α–Marginal Entropies," to appear in Special Issue on BSS and ICA, Neurocomputation , 2001.

[8] D. Erdogmus, J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," to appear in IEEE Transactions on Neural Networks, 2001.

[9] A. Renyi, *Probability Theory*, American Elsevier Publishing Company Inc., New York, 1970.

[10] E. Parzen, "On Estimation of a Probability Density Function and Mode", in *Time Series Analysis Papers*, Holden-Day, Inc., CA, 1967.

[11] D.G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley Pub. Co., Massachussetts, 1973.

[12] R.Y. Rubinstein, *Simulation and the Monte Carlo Method*, John Wiley & Sons, NY, 1981.

[13] J. C. Principe, D. Xu, J.W. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering, Vol. 1: Blind Source Separation*, S. Haykin Editor, pp.265-319, John Wiley & Son, New York, 2000.