Online Entropy Manipulation: Stochastic Information Gradient

Deniz Erdogmus, *Member, IEEE*, Kenneth E. Hild, II, *Student Member, IEEE*, and Jose C. Principe, *Fellow, IEEE*

***Abstract*—Entropy has found significant applications in numerous signal processing problems including independent components analysis and blind deconvolution. In general, entropy estimators require $O(N^2)$ operations, N being the number of samples. For practical online entropy manipulation, it is desirable to determine a stochastic gradient for entropy, which has $O(N)$ complexity. In this letter, we propose a stochastic Shannon's entropy estimator. We determine the corresponding stochastic gradient and investigate its performance. The proposed stochastic gradient for Shannon's entropy can be used in online adaptation problems where the optimization of an entropy-based cost function is necessary.**

***Index Terms*—Shannon's entropy, stochastic gradient for entropy.**

I. Introduction

FOLLOWING Shannon's introduction of entropy [1], information-theoretic approaches experienced an increased interest among signal processing researchers [2]–[4]. This interest is mainly due to their intellectual appeal and the elegant mathematical theory. Alternative definitions such as Renyi's, which encompass Shannon's as a special case, have also emerged [5]. Independent components analysis (ICA) and blind deconvolution are two problems where entropy has been extensively adopted as the optimality criterion [2], [6], [7]. Recently, we have introduced a minimum-error-entropy (MEE) training method that outperformed mse in supervised training [8], [9]. In practice, for a real-time solution of these problems, a stochastic entropy gradient is required.

Information-theoretic approaches are becoming more widespread in the contemporary signal processing literature. However, in problems that require online adaptation, the stochastic gradient algorithms have so far been application-specific. For example, the entropy maximization procedure (InfoMax) proposed by Bell and Sejnowski [10] specifically applies to the adaptation of a layer of perceptrons coupled through their joint output entropies. Therefore, researchers who use this approach have to fit their problem into the structural requirements designated by the algorithm [11], [12]. Another approach in entropy estimation uses Jaynes' maximum-entropy principle [13]. Hyvarinen proposes a first-order approximation to Shannon's entropy based on probability density function (pdf) estimates based on Jaynes' principle [14]. However, this approach, in general, requires solving a set of nonlinear equations to determine the optimal pdf parameters, which impedes the derivation of a computationally efficient stochastic gradient. Viola [15] proposed an entropy estimator similar to the one we use. However, Viola [15] and Torkkola [16] (who uses the entropy estimator we proposed) take the approach of selecting small random subsets of the training data when utilizing the stochastic gradient approach. The computational complexity of these estimators are all $O(N^2)$, N being the number of samples. In this letter, we present a stochastic gradient for Shannon's entropy based on a nonparametric estimator that utilizes Parzen windowing and that exhibits $O(N)$ complexity.

Manuscript received April 11, 2002; revised October 9, 2002. This work was supported by the National Science Foundation under Grant ECS-9900394. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marcelo G. S. Bruno.

The authors are with the Computational NeuroEngineering Laboratory, University of Florida, Gainesville, FL 32611 USA (e-mail: deniz@cnel.ufl.edu; k.hild@ieee.org; principe@cnel.ufl.edu).

Digital Object Identifier 10.1109/LSP.2003.814400

II. Nonparametric Estimation of Shannon's Entropy

Shannon's entropy for a random variable Y with pdf $f_Y(y)$ is [1]

$$H_S(Y) = -\int_{-\infty}^{\infty} f_Y(y)\log f_Y(y)dy = E_Y\left[-\log f_Y(Y)\right]. \quad (1)$$

We have recently proposed and successfully applied to practical problems a nonparametric estimator for Renyi's entropy based on Parzen windowing given N samples $\{y_1,\ldots,y_N\}$ [9], [17], [18]. Parzen windowing approximates the unknown pdf underlying the samples by $\hat{f}_Y(y) = (1/N)\sum_{i=1}^{N}\kappa_\sigma(y-y_i)$, where $\kappa_\sigma(x)$ is the kernel function with size σ [19]. Returning to Shannon's entropy in (1), similar to [9], replacing the expectation in (1) with the sample mean and substituting the Parzen pdf estimate, one will obtain a nonparametric estimate of Shannon's entropy. It was shown that for even-symmetric, unimodal, and differentiable kernels the global minima of this entropy estimator and actual entropy coincide [9]. Although our principal objective is not to accurately estimate entropy, but to determine the optimal adaptive system weights under the entropy-based criterion, we refer the interested readers to [20] and [21] for a treatment of the problem of kernel size estimation in kernel pdf estimates.

III. Stochastic Entropy Estimator and Its Gradient

Having introduced the methodology to derive the estimator for Shannon's entropy, we can derive the stochastic entropy gradient. Widrow's stochastic estimator for mse in the derivation of LMS uses the instantaneous error-square value by dropping the expectation operator. In stochastically approximating (1), we will take Widrow's lead. Dropping the expectation and evaluating its argument at the most recent sample of the random variable Y, we obtain $H_S(Y) =$

$E_Y[-\log f_Y(Y)] \approx -\log f_Y(y_k)$, y_k denoting the most recent sample at time k. Since, in practice, the pdf of Y is unknown, a Parzen window estimate is utilized. In a nonstationary environment, the online pdf estimate can be obtained using a sliding window of samples. Assuming a window length of L samples, the stochastic pdf estimate of Y evaluated at y_k is

$$\hat{f}_Y(y_k) = \frac{1}{L}\sum_{i=k-L}^{k-1} \kappa_\sigma(y_k - y_i). \tag{2}$$

Thus, the stochastic entropy estimate at time k becomes

$$\hat{H}_{S,k}(Y) = -\log\left(\frac{1}{L}\sum_{i=k-L}^{k-1} \kappa_\sigma(y_k - y_i)\right). \tag{3}$$

Clearly, the expected value of (3) satisfies $E[\hat{H}_{S,k}(Y)] = E[-\log \hat{f}_Y(y_k)] = \hat{H}_S(Y)$, where $\hat{H}_S(Y)$ is Shannon's entropy estimated using the Parzen window method. Recall that the Parzen pdf estimator is biased [19]; therefore, the stochastic estimator in (3) becomes a biased estimator of the actual entropy given in (1). It is now trivial to see the stochastic gradient of entropy with respect to the weight vector of the adaptive system that generated the samples of Y according to $y_k = g(\mathbf{x}_k; \mathbf{w})$ is

$$\frac{\partial \hat{H}_{S,k}}{\partial \mathbf{w}} = -\frac{\displaystyle\sum_{i=k-L}^{k-1} \kappa'_\sigma(y_k - y_i)\left(\frac{\partial y_k}{\partial \mathbf{w}} - \frac{\partial y_i}{\partial \mathbf{w}}\right)}{\displaystyle\sum_{i=k-L}^{k-1} \kappa_\sigma(y_k - y_i)} \tag{4}$$

where $\kappa'_\sigma(x)$ is the derivative of the kernel function. We call this expression the stochastic information gradient (SIG). The selection of L is dictated by two factors: the length of the interval in which the signal of interest remains approximately stationary, and the computational load limitation on each update. Once the stochastic gradient in (4) is evaluated, the weights can be updated by $\mathbf{w} \leftarrow \mathbf{w} \pm \eta \partial \hat{H}_{S,k}/\partial \mathbf{w}$, where η is the learning rate. The sign of the update is determined by whether the task is minimization or maximization of the entropy of Y.

An important issue in adaptation is the convergence properties of the algorithm. Now, we will investigate the convergence of the proposed entropy adaptation algorithm, whose weight update is given in (4), for the special case of error entropy minimization in supervised training of a linear filter. In this case, the output of the linear filter is given by $z = \mathbf{w}^T\mathbf{x}$, and the error is defined as $e_k = d_k - z_k$. We also assume that the desired signal is generated by a linear system with weight vector w_* according to $d = \mathbf{w}_*^T\mathbf{x}$. Consider the stochastic error entropy minimization algorithm given below, for this setup.

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta\Delta\mathbf{w}_k = \mathbf{w}_k - \eta\frac{\displaystyle\sum_{i=k-L}^{k-1} \kappa'_\sigma(e_k - e_i)(\mathbf{x}_k - \mathbf{x}_i)}{\displaystyle\sum_{i=k-L}^{k-1} \kappa_\sigma(e_k - e_i)}. \tag{5}$$

Notice that at $\mathbf{w}_k = \mathbf{w}_*$, the expected value of the update becomes zero, i.e., $E[\Delta\mathbf{w}_k]|_{\mathbf{w}_*} = 0$, because all error samples become zero. Therefore, the true weight vector is a stationary point of the proposed algorithm. Now consider the propagation of the weight error vector $\varepsilon_k = \mathbf{w}_* - \mathbf{w}_k$ through the updates. Subtracting both sides of (5) from $\mathbf{w}_*$, we get $\varepsilon_{k+1} = \varepsilon_k + \eta\Delta\mathbf{w}_k$. Multiplying both sides of this equation with its transpose to get the weight error vector norm yields $\|\varepsilon_{k+1}\|^2 = \|\varepsilon_k\|^2 + 2\eta\varepsilon_k^T\Delta\mathbf{w}_k + \eta^2\|\Delta\mathbf{w}_k\|^2$. In order for the weights to converge to the true weights, we require $\|\varepsilon_{k+1}\|^2 < \|\varepsilon_k\|^2$, which is guaranteed when the step size satisfies the inequality $0 < \eta < -2\varepsilon_k^T\Delta\mathbf{w}_k/\|\Delta\mathbf{w}_k\|^2$. Using the identities $\varepsilon_k^T\mathbf{x}_k = e_k$, $\varepsilon_k^T\mathbf{x}_i = e_i$, and the definition of $\Delta\mathbf{w}_k$ given in (5), the upper bound on the positive step size becomes

$$0 < \eta < \frac{-2\displaystyle\sum_{i=k-L}^{k-1} \kappa'_\sigma(e_k - e_i)(e_k - e_i)}{\|\Delta\mathbf{w}_k\|^2 \displaystyle\sum_{i=k-L}^{k-1} \kappa_\sigma(e_k - e_i)}. \tag{6}$$

Notice that if the kernel function is selected to be a unimodal pdf with peak at the origin (e.g., symmetric), $\kappa'_\sigma(e_k - e_i)(e_k - e_i) < 0$. This is true: since the only zero-crossing of a unimodal kernel with peak at the origin is the origin, we have $\kappa'_\sigma(x < 0) > 0$ and $\kappa'_\sigma(x > 0) < 0$; therefore, $\text{sign}(\kappa'_\sigma(x)) = -\text{sign}(x)$. In addition, $\kappa_\sigma(e_k - e_i) > 0$ for any error value; therefore, the upper bound on the step size is positive and valid. Similar convergence analyses specific to the problem at hand can be conducted following the approach outlined above. The convergence analysis for a nonlinear adaptive system, however, would be more complicated, since the useful identity $\varepsilon_k^T\mathbf{x}_k = e_k$ would not be valid anymore.

IV. Relationship Between SIG and LMS

In the extreme case, one might choose to utilize only a single sample in (4), corresponding to $L = 1$. Then, especially for the supervised training of ADALINE using the MEE criterion [9], SIG becomes quite simple. In ADALINE, the output is a linear combination of the inputs, i.e., $z_k = \mathbf{w}^T\mathbf{x}_k$, and the error is $e_k = d_k - z_k$, where d_k is the desired output corresponding to the input vector $\mathbf{x}_k$. Selecting $L = 1$ and $y = e$ in (4), SIG reduces to

$$\begin{aligned}\frac{\partial \hat{\mathbf{H}}_{\mathbf{S},\mathbf{k}}}{\partial \mathbf{w}} &= \frac{\kappa'_\sigma(e_k - e_{k-1})}{\kappa_\sigma(e_k - e_{k-1})} \cdot (\mathbf{x}_k - \mathbf{x}_{k-1}) \\ &\triangleq f(e_k - e_{k-1}) \cdot (\mathbf{x}_k - \mathbf{x}_{k-1}).\end{aligned} \tag{7}$$

Notice that in this problem, the aim is to minimize the entropy of the error signal, which is obtained using the output of the adaptive system. In (7), $f(x) = \kappa'_\sigma(x)/\kappa_\sigma(x)$. When using Gaussian kernels with variance σ^2, (7) reduces to

$$\frac{\partial \hat{\mathbf{H}}_{\mathbf{S},\mathbf{k}}}{\partial \mathbf{w}} = -\frac{1}{\sigma^2}(e_k - e_{k-1}) \cdot (\mathbf{x}_k - \mathbf{x}_{k-1}) = -\frac{1}{\sigma^2}\Delta e_k \Delta\mathbf{x}_k \tag{8}$$

due to the identity $G'_\sigma(x) = -x \cdot G_\sigma(x)/\sigma^2$ for the Gaussian kernel $G_\sigma(x)$. Recall that LMS updates use the instantaneous values of the error and the input vector (i.e., the LMS update is $e_k\mathbf{x}_k$) and attempts to minimize the correlation between e_k and $\mathbf{x}_k$[3], [4]. The SIG presented in (8) is similar in structure to the

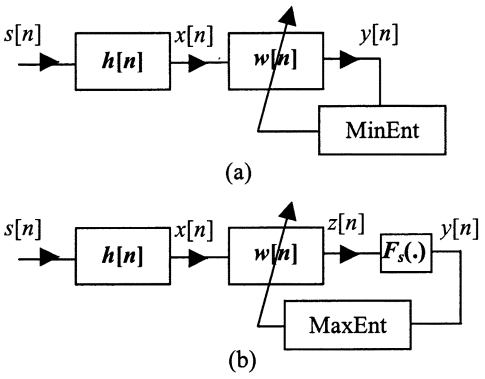

Fig. 1. Schematic diagram of blind deconvolution. (a) Minimum entropy. (b) Maximum entropy.

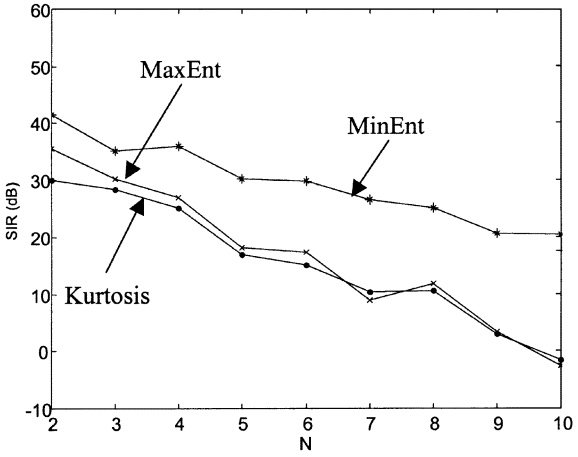

Fig. 2. Performance of the stochastic entropy and kurtosis algorithms in blind deconvolution versus the equalizer length. (*) Minimum entropy. (×) Maximum entropy. (.) Kurtosis.

LMS updates, except for the fact that it acts on the instantaneous increments Δe_k and $\Delta \mathbf{x}_k$ trying to decorrelate them when error entropy is minimized. This is an interesting observation: minimizing (maximizing) entropy is achieved by minimizing (maximizing) the correlation between sample separations of the input and the error (output) signals.

V. Simulations

SIG performance in entropy manipulation will be demonstrated in two case studies involving sample-by-sample training of adaptive systems. First, we investigate the performance of SIG in online blind deconvolution. A schematic diagram of the blind deconvolution problem is depicted in Fig. 1, where two alternative approaches are presented. In this problem, it is assumed that the linear channel impulse response $h[n]$ and the specific source signal $s[n]$ are unknown. The only knowledge regarding the source is that it is a sequence of independent and identically distributed samples (when second- order criteria are employed, the independence condition is relaxed to uncorrelatedness). The minimum-entropy approach is based on the fact that for constant variance, Gaussian density exhibits maximum entropy, and convolution forces signal pdfs toward Gaussian due to central limit theorem. Therefore, minimizing the output entropy achieves deconvolution. The maximum-entropy approach uses the facts that for fixed range, 1) uniform density exhibits maximum entropy, 2) the pdf of the equalizer output (before the nonlinearity) is equal to the derivative of the nonlinearity, which must be selected to be the source cdf, and 3) the input and the output of a filter have the same non-Gaussian pdf if and only if the overall impulse response is δ. In this scheme, the knowledge of the source cdf $F_s(\cdot)$ is required.

Based on the minimum-entropy deconvolution principle [6], we minimize the following modified entropy criterion for blind deconvolution using the presented stochastic entropy gradient approach: $J(Y) = H_S(Y) - \log[\text{Std}(Y)]$, where Std stands for the standard deviation. The modification of the entropy criterion with the standard deviation of the output results in a scale-invariant criterion, which is essential for blind deconvolution. For the maximum-entropy deconvolution, we simply maximize the entropy of the output of the nonlinearity [10]. Monte Carlo simulations are performed using randomly selected all-poles infinite impulse response channel filters (this allows us to determine the length of the ideal equalizer easily for simulation purposes). For each of the window lengths $L = 10, 25, 50, 75,$ and 100, we have performed 90 Monte Carlo simulations using 10 000-sample sets of Laplacian distributed source signals (ten simulations for each of the equalizer lengths from two to ten). The average signal-to-interference ratio (SIR) (defined as the ratio of the power of the source signal to that of the convolutive interference from different lags of the source in decibels) turned out to be 28.7, 28.6, 28.3, 29.9, and 29.0 dB, using the minimum-entropy approach, versus 10.7, 14.5, 13.2, 17.0, and 15.7 dB, using the maximum-entropy approach, respectively. The lower performance of the maximum-entropy approach can be attributed to the fact that Gaussian kernels are used in Parzen windowing, which does not provide a sufficiently accurate representation of the desired uniform density at the output of the nonlinearity when the entropy is maximized. For a comparison, we have performed the same experiments using the normalized kurtosis criterion [6] with a sliding window of samples to compute its stochastic gradient. The average SIRs came out to be 9.15, 16.9, 11.6, 15.2, and 14.4 dB, respectively, for each of the window lengths given above. The performance of the kurtosis-based algorithm was low, because kurtosis requires a large number of samples for an accurate estimate [2]. As expected, the performance of the stochastic gradient algorithms (entropy and kurtosis) degraded as the length of the weight vector to be optimized increased. The average SIRs for all three algorithms as a function of filter length is presented in Fig. 2. Notice that, in all cases, the SIG-based minimum-entropy blind deconvolution algorithm outperforms kurtosis.

The second case study is the blind separation of two audio sources, instantaneously mixed with a 2×2 matrix whose entries are selected randomly from the interval $[-1, 1]$. In the blind source separation (BSS) problem, a common assumption that is exploited about the sources is their independence. In this experiment, one of the source signals is a female voice, and the other is a male voice sampled at 8 kHz. To separate these sources from a vector of mixtures of the two, the Mermaid-SIG algorithm will be used [23]. This algorithm adapts an orthonormal matrix by minimizing the sum of the marginal output entropies, i.e.,

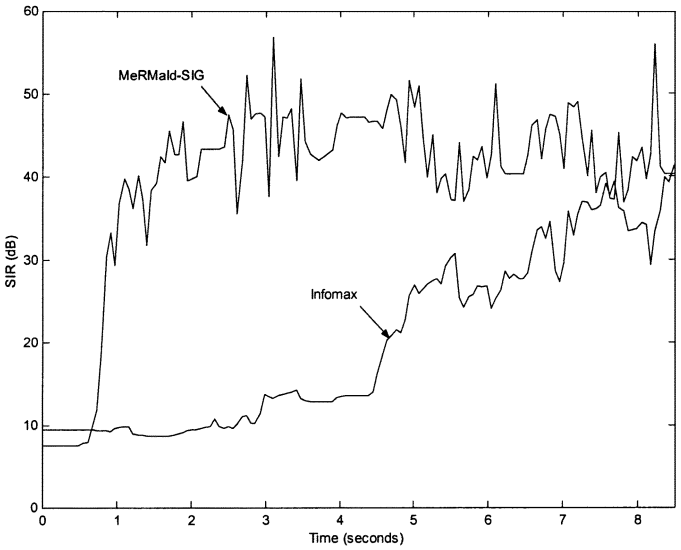

Fig. 3. Average SIR (in decibels) versus time (seconds) for Mermaid-SIG and Infomax algorithms in online two-source two-mixture BSS using 20 Monte Carlo results with random mixtures.

$\sum_{o=1}^{n} H_S(Y^o)$, where Y^o is the o^{th} separated output signal. A whitening procedure precedes the orthonormal matrix step, and any fast, online principal component analysis algorithm can be used to perform the whitening task. We employ the SIPEX-G algorithm in the simulations [22]. In the following simulations, we will use $L = 10$ and Gaussian kernels in SIG. Details regarding this topology can be found in [18]. As a comparison, we also applied the well-known Bell–Sejnowski algorithm (Infomax) to the same data [10]. For a fair comparison, the same whitening procedure is used to enhance the convergence speed of Infomax. Average of the SIR, defined as in [18], versus time is shown in Fig. 3. The averaging is performed over 20 Monte Carlo runs with different mixing matrices in order to obtain a better and fair comparison of the two algorithms. Step sizes of both algorithms are adjusted to yield the same magnitude of oscillations after convergence, yet Mermaid-SIG converges in about 0.5 s (after speech starts) whereas Infomax needs 7 s to achieve the same level of performance.

VI. Conclusion

In this letter, a stochastic entropy estimator based on Parzen window estimates of the underlying pdf is presented. We have derived the corresponding stochastic gradient and established that a special case of this gradient is closely related to LMS, differing in aim by acting on the instantaneous increments of the input and output signals instead of instantaneous values. For the special case of supervised training of a linear filter using the minimum-error entropy criterion, we have derived the upper bound on the step size to guarantee absolute convergence to the optimal solution. The performance of the proposed *stochastic information gradient* in online entropy manipulation problems is verified by blind source separation and blind deconvolution case studies. The kernel function is an important parameter that might affect the performance of the proposed entropy manipulation algorithm greatly. Although it might be possible to optimize the kernel function for different signals and pdfs, this issue is not addressed here. In addition, the Parzen window pdf estimator assumes that the samples are independent and also drawn from the same distribution. In the case of ICA and MEE applications, no problems associated with this concern were encountered. However, in some situations, the dependencies between (consecutive) samples used in the algorithm might hinder the performance. This issue must also be investigated in future work.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.
[3] S. Haykin, Ed., *Unsupervised Adaptive Filtering*. New York: Wiley, 2000, vol. 1, Blind Source Separation.
[4] S. Haykin, Ed., *Unsupervised Adaptive Filtering*. New York: Wiley, 2000, vol. 2, Blind Deconvolution.
[5] A. Rényi, *Probability Theory*. Amsterdam, The Netherlands: North-Holland, 1970.
[6] D. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II*. New York: Academic, 1981, pp. 565–609.
[7] A. Benveniste, M. Goursat, and G. Ruget, "Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications," *IEEE Trans. Automat. Contr.*, vol. 25, pp. 385–399, June 1980.
[8] D. Erdogmus and J. C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Processing*, vol. 50, pp. 1780–1786, July 2002.
[9] ——, "Generalized information potential criterion for adaptive system training," *IEEE Trans. Neural Networks*, vol. 13, pp. 1035–1044, Sept. 2002.
[10] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.
[11] G. Miller and D. Horn, "Probability density estimation using entropy maximization," *Neural Comput.*, vol. 10, no. 7, pp. 1925–1938, 1998.
[12] A. Touzni, I. Fijalkow, M. G. Larimore, and J. R. Treichler, "A globally convergent approach for blind MIMO adaptive deconvolution," *IEEE Trans. Signal Processing*, vol. 49, pp. 1166–1178, June 2001.
[13] E. T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, R. D. Rosenkrantz, Ed. Dordrecht, The Netherlands: D. Reidel, 1983.
[14] A. Hyvarinen, "New approximations of differential entropy for independent component analysis and projection pursuit," Helsinki Univ. Technol., Helsinki, Finland, Rep. A47, 1997.
[15] P. Viola, N. N. Schraudolph, and T. J. Sejnowski, "Empirical entropy manipulation for real-world problems," *Advances in Neural Information Processing Systems*, vol. 8, pp. 851–857, 1995.
[16] K. Torkkola, "On feature extraction by mutual information maximization," in *Proc. ICASSP*, Orlando, FL, 2002, pp. 821–824.
[17] J. C. Principe, J. W. Fisher, and D. Xu, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, 2000, pp. 265–319.
[18] D. Erdogmus, K. E. Hild II, and J. C. Principe, "Blind source separation using Renyi's α–marginal entropies," *Neurocomput.*, vol. 49, no. 1, pp. 25–38, Dec. 2002.
[19] E. Parzen, "On estimation of a probability density function and mode," in *Time Series Analysis Papers*. San Francisco, CA: Holden-Day, 1967.
[20] S. R. Sain and D. W. Scott, "On locally adaptive density estimation," *J. Amer. Stat. Assoc.*, vol. 91, pp. 1525–1534, 1996.
[21] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer, 2001.
[22] D. Erdogmus, Y. N. Rao, J. C. Principe, J. Zhao, and K. E. Hild II, "Simultaneous extraction of principal components using Givens rotations and output variances," in *Proc. ICASSP*, Orlando, FL, 2002, pp. 1069–1072.
[23] K. E. Hild II, D. Erdogmus, and J. C. Principe, "Blind source separation of time-varying, instantaneous mixtures using an on-line algorithm," in *Proc. ICASSP*, Orlando, FL, 2002, pp. 993–996.