SOME RESULTS ON A BSS ALGORITHM UNDER NON-STANDARD CONDITIONS

Saleem A. Kassam and Yinglu Zhang
Department of Electrical Engineering
University of Pennsylvania, Philadelphia, PA 19104
(kassam@ee.upenn.edu)
and
George V. Moustakides
Department of Computer Engineering and Informatics
University of Patras,
Patras, Greece

Abstract

We consider some aspects of performance, approximation, and robustness of the EASI *(equivariant adaptive separation by independence)* algorithms for adaptive blind source separation. This algorithm class is useful for separating unknown linear mixtures of unknown independent sources. We characterize the nature of the optimum solutions in this class of algorithms (that depend on some nonlinear g function.) The result is used to establish the nature of optimum quantizer nonlinearities, and also to introduce robustness against deviations from nominal source pdf assumptions.

1. Introduction

Separation of independent source sequences from a set of observed linear mixtures of the sequences is referred to as the *blind source separation* (bss) problem, because it is usually assumed that nothing is known about the mixing matrix, and about the sources themselves apart from their mutual independence. We will let $\mathbf{s}(k) = [s_1(k), s_2(k),\ldots,s_n(k)]^T$ denote the n-component vector time series of independent zero-mean sources that are to be separated from an observed mixture time series $\mathbf{x}(k) = \mathbf{A}\,\mathbf{s}(k)$. Here $\mathbf{A}$ is an instantaneous n×n matrix that is unknown. One has to identify $\mathbf{A}^{-1}$. This problem has attracted much interest in the last fifteen years [1,2,3]. We will start from a well-known approach and algorithm to identify the matrix $\mathbf{A}^{-1}$ and consider some particular aspects of the solution, including its optimization and robustness. We will consider the algorithm suggested by Cardoso and Laheld [2], called the EASI algorithm *(equivariant adaptive separation via independence)*. This algorithm is based on the idea of obtaining through an on-line or adaptive scheme a sequence of estimates $\mathbf{B}(k)$ of $\mathbf{A}^{-1}$. It combines a pre-whitening process with the optimization of a "contrast" function or criterion function of the statistics of the output $\mathbf{y}=\mathbf{Bx} =[y_1,\ldots,y_n]^T$ with respect to the matrix $\mathbf{B}$. The contrast function is chosen such that it achieves an extreme value when the $\mathbf{y}$ components are independent, under the constraint of orthogonality (pre-whitening). Specifically, if $\phi(\mathbf{B})=E\{\psi[\mathbf{Bx}]\}$ is such a contrast function, then it is shown in [2] that its minimization leads to use of the adaptive algorithm

$$\mathbf{B}(k+1) = \mathbf{B}(k) - \lambda\, \mathbf{H}[\mathbf{y}(k)]\mathbf{B}(k) \qquad (1)$$

where $\mathbf{H}[\mathbf{y}] = \psi'[\mathbf{y}]\,\mathbf{y}^T$ and ψ' is the gradient of ψ. The EASI algorithm is obtained by modifying $\mathbf{H}(\mathbf{y})$ in two ways. A whitening constraint is added, and $\mathbf{H}$ is made skew-symmetric, reflecting the fact that after a whitening transformation the linear transformation of $\mathbf{y}$ is reduced to an orthogonal rotation. Suppose ψ is a sum of component functions $\psi_i(y_i)$, then ψ' has components $\psi_i'(y_i)$. In this case we may generalize the algorithm by using any scalar function g in its definition of $\mathbf{H}[\mathbf{y}]$. These modifications and generalizations lead to the EASI algorithm based on

$$\mathbf{H}[\mathbf{y}] = \mathbf{y}\mathbf{y}^T - \mathbf{I} + \mathbf{g}(\mathbf{y})\mathbf{y}^T - \mathbf{y}\mathbf{g}(\mathbf{y})^T \qquad (2)$$

where $\mathbf{g}(\mathbf{y})=[g(y_1), \ldots, g(y_n)]^T$. The algorithm of (1) with $\mathbf{H}$ defined as in (2) was developed and evaluated for performance in [2].

The adaptive algorithm attempts to solve $E\{\mathbf{H}[\mathbf{y}]\}=0$. This holds when the components of $\mathbf{y}$ are independent with unit variance. We may also use the condition $E\{\mathbf{H}[\mathbf{y}]\}=0$ as the basis of a batch algorithm [1].

2. Performance of the Adaptive Algorithm

It is of interest to consider the choice of the scalar function g in the above algorithm. For this it is convenient to express the adaptation as an update for the matrix $\mathbf{C}(k)=\mathbf{B}(k)\mathbf{A}$ which should converge to $\mathbf{I}$. We find that

$$\mathbf{C}(k+1) = \mathbf{C}(k) - \lambda \mathbf{H}[\mathbf{C}(k)\mathbf{s}(k)]\,\mathbf{C}(k)$$

There are two aspects of performance that we are interested in. One is stability and convergence rate of the algorithm, the other is the variance of the solution in the steady state, both obtained in the *local case* of small step-size λ.

A stability analysis has been conducted in [2], with the result that a sufficient condition for stability of the solution $\mathbf{C}=\mathbf{I}$ depends on the sign of the parameters

$$\kappa_i = E\{g'(s_i)\} - E\{s_i\, g(s_i)\} \qquad (3)$$

The result $\mathbf{C}=\mathbf{I}$ is a stable solution if all the κ_i are *strictly positive*. The definition of the κ_i is in terms of *unit-variance* source signals s_i, because of the normalization inherent in the algorithm. If $g(s)=s^3$, then k_i is -1×kurtosis, and the algorithm is stable for sub-Gaussian sources with negative kurtoses. Note that for a unit-variance Gaussian source κ_i is always zero, and that for $g(s)=s$ the κ_i are always zero for any unit-variance source. The κ_i are stability indices and characterize the exponential convergence rate for the off-diagonal terms. The diagonal terms have a guaranteed stability.

The steady state covariance matrix of $\mathbf{C}(k) - \mathbf{I}$ can also be found [2,4] and from this the steady-state variance of the i,j-th term $C_{ij}(k)$ of the matrix $\mathbf{C}(k)$ may be obtained. For identically distributed sources for which $\kappa_i=\kappa$, we find that

$$\mathrm{Var}\{C_{ij}(k)\} = \frac{\lambda}{2}\left[\frac{1}{2}+\frac{\gamma}{\kappa}\right] \qquad (4)$$

where

$$\gamma = E\{g^2(s)\} - E^2\{s\, g(s)\} \qquad (5)$$

The constant additive term 1/2 in the variance expression comes from the pre-whitening part of the algorithm. Stability analysis also shows that the exponential convergence rate for the skew-symmetric g-dependent part of the algorithm is $\lambda\kappa$, whereas for the symmetric pre-whitening part it is simply λ, independent of g. We also find that the steady-state variance of $C_{ii}(k)$ is independent of g. A fair comparison of two algorithms should be based on both the exponential convergence rate as well as steady state error variance, and Moustakides [5] has introduced the notion of the local "efficacy" of an adaptive algorithm as a ratio of its exponential convergence rate and its steady state error variance. In our context the exponential convergence rate is $\lambda\, min\{1,\kappa\}$, so that the efficacy is

$$Q = \frac{min(1,\kappa)}{\frac{1}{4}+\frac{\gamma}{2\kappa}}$$

This is a maximum when $\kappa=1$, under which constraint the efficacy is found to be

$$Q = \frac{1}{\frac{1}{4}+\frac{\gamma}{2\kappa^2}}$$

This leads to a criterion of performance $P(g)$ as a function of g given as

$$P(g) = \frac{\kappa^2}{\gamma} \qquad (6)$$

We will call this the *"bss efficacy"*. This is *independent of amplitude scaling* of g. To enforce the condition $\kappa=1$, we may scale it in amplitude after finding the maximizing g.

Before we continue with a consideration of the bss efficacy, we should note that this same quantity occurs in the asymptotic variance expression for the $C_{ij}(k)$ in the batch-mode algorithm [1]. It also arises in the asymptotic variance result for a batch-mode algorithm minimizing a contrast function that is a generalization of the kurtosis, with the fourth-power function replaced by a function G with derivative g [6].

3. Optimization of BSS Efficacy and Mean-Squared Error of Fit

There have been several analyses of what we call the bss efficacy and its variants, and the results have given the optimum g function. We will give below a simple approach to establishing that the optimum function maximizing the bss efficacy is any function of the form

$$g_o(s) = a\, g_{LO}(s) + b\, s \qquad (7)$$

where

$$g_{LO}(s) = \frac{-f'(s)}{f(s)} \qquad (8)$$

the familiar optimum nonlinearity arising in locally optimum signal detection and in M-estimation; here f is the unit-variance common pdf of the independent sources. Related results have also been given in [7,1]

For this express the bss efficacy as

$$P(g) = \frac{(E\{[g(s)-cs][g_{LO}(s)-s]\})^2}{E\{[g(s)-sE\{sg(s)\}]^2\}} \qquad (9)$$

where c is *any* constant. This follows from the unit-variance condition on the identified sources. In particular, set $c = E\{sg(s)\}$. Now from the Schwarz inequality, it follows that $P(g)$ is upper bounded by $E\{[g_{LO}(s)-s]^2\} = I(f)-1$, where $I(f)$ is the Fisher information function. This maximum value of $P(g)$ is obtained with

$$g(s) - sE\{sg(s)\} = a\,[g_{LO}(s) - s]$$

where a is any constant, and this leads directly to the result of (7).

It is particularly significant that the optimum function is any linear-biased version of g_{LO}. In fact for *any* function $g(s)$, the bss efficacies obtained with $g(s)$ and $g(s)+bs$ are exactly the same. A stronger statement can be made: the adaptive algorithm is *invariant* to any linear bias of the function g.

Mean-Square Error of Fit to $g_{LO}(s)$

Let us now consider the discrepancy measure:

$$\Delta(g;\, a,\, b) = E\{[ag(s) + bs - g_{LO}(s)]^2\} \qquad (10)$$

under the pdf f of s. This is the mse between a scaled and linear-biased version of $g(s)$ and $g_{LO}(s)$. We expect that the best (minimum) value of Δ for a given function g used in the adaptive algorithm also indicates the efficacy of that nonlinearity, because we know that a best nonlinearity after scaling and linear biasing should be $g_{LO}(s)$. An equivalent discrepancy measure is

$$D(g;\, \alpha, \beta) = E\{[\alpha g(s) - \beta s - (g_{LO}(s) - s)]^2\} \qquad (11)$$

Minimizing D with respect to the parameters α and β, we find the optimum values satisfy

$$\alpha = \frac{E\{[g(s) - sE\{sg(s)\}][g_{LO}(s) - s]\}}{E\{[g(s)\} - sE\{sg(s)]^2\}}$$

and

$$\beta = \alpha\, E\{sg(s)\}$$

and the resulting best value $D^*(g)$ of the mean-square discrepancy of $g(s)$ from $g_{LO}(s)$ is easily found to be

$$D^*(g) = I(f) - 1 - P(g)$$

Thus minimizing $D(g;\alpha,\beta)$ wrt g and α, β is the same as maximizing the bss efficacy. This result provides an interesting connection between bss efficacy maximization and optimizing the fit of a scaled and linearly biased nonlinearity g to g_{LO}. It provides clearer insight as to the nature of the best g function that should be used if one is restricted to a specific class of functions to choose from.

4. Piecewise Constant (Quantizer) g Functions

In applications it may be very desirable to implement a simple g function from a restricted class of possibilities, in particular an M-interval quantizer, in the EASI algorithm or its variants. This might be useful to lessen computation or the data transmission burden, to obtain easily on-line-optimized g functions within the class, and to obtain algorithms more robust to mismatch between assumed and actual source distributions.

Let Q_M be the class of M-interval quantizers q, so that in the k-th interval $I_k = (t_{k-1}, t_k)$ the level is α_k. To optimize the choice of these intervals and associated levels for $g = q \in Q_M$, consider the discrepancy measure of (11) written as:

$$D(q;\beta) = \sum_{k=1}^{M} \int_{t_{k-1}}^{t_k} \{\alpha_k - \beta s - [g_{LO} - s]\}^2\, dF(s) \qquad (12)$$

Here F is the unit-variance distribution function with pdf f of each source. Writing the conditions for minimization with respect to choice of the t_k and the α_k, we get the expressions

$$\alpha_k = \frac{-\delta f_k}{\delta F_k} - \delta C_k[1-\beta] \qquad (13a)$$

$$\frac{\alpha_k + \alpha_{k+1}}{2} = g_{LO}(t_k) - t_k[1-\beta] \qquad (13b)$$

and

$$\beta = \sum_{k=1}^{M} \alpha_k \delta A_k \qquad (13c)$$

where $-\delta f_k = f(t_{k-1}) - f(t_k)$, $\delta F_k = F(t_k) - F(t_{k-1})$, the $\delta A_k = \int_{t_{k-1}}^{t_k} s\, dF(s)$ and $\delta C_k = \delta A_k / \delta F_k$

Note that the above conditions giving the optimum quantizer parameters for the bss algorithm parallel those for locally optimum quantization in a known-signal detection setting [8]. The key difference is that there is a linear bias term included in the $g_{LO}(s)$ function, which gives rise to the terms in β in the above equations.

Consider a unit-variance generalized Gaussian source pdf of the form $f(s)=K\exp(-k|s|^m)$, giving the result $g_{LO}(s)=km|s|^{m-1}sgn(s)$. For a sub-Gaussian source with $m>2$, the LO function increases quickly. To obtain for example a good 4-interval symmetric quantizer, the viewpoint of approximating some linear biased version of the LO function suggests biasing this function downwards and approximating the result with a quantizer with a negative level for an interval $[0,t_3)$ and a positive level for $[t_3,\infty)$. Indeed, Figure 1 shows the optimum 4-interval quantizer for the unit-variance generalized Gaussian pdf with $m=3$, and also the optimal linear-biased version of g_{LO} that this approximates. For this case the optimum parameters gave $\beta = -0.152$. This quantization result is counter-intuitive if a free linear bias is not taken into account.

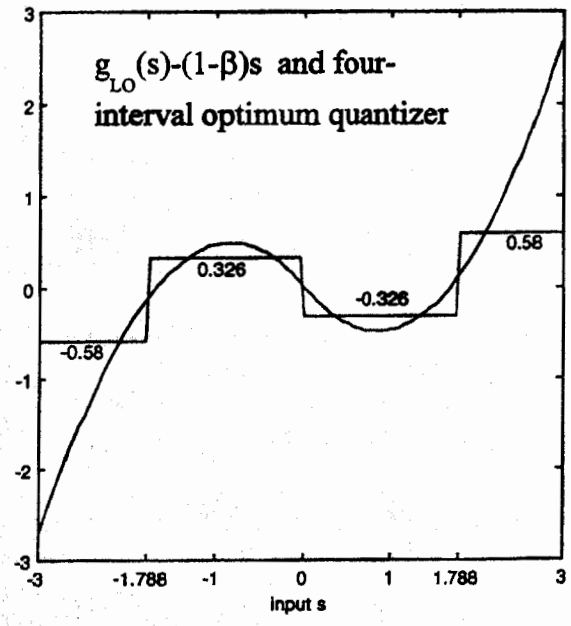
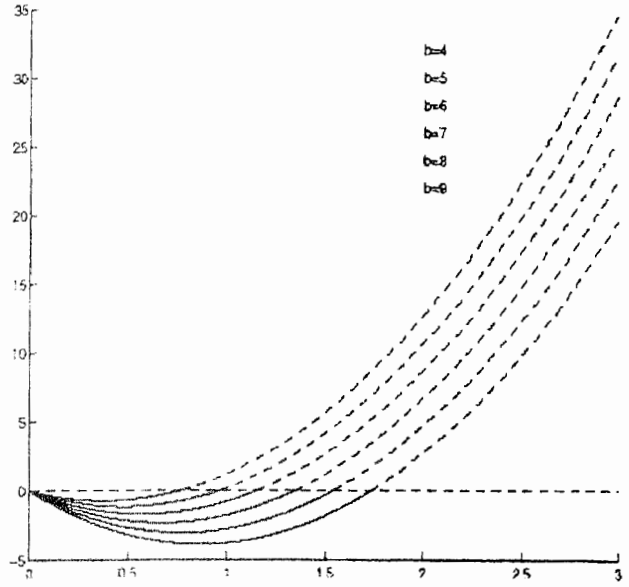


Figure 1. *4-Interval Optimum Quantization for m=3 Generalized Gaussian Sources*

We have to also consider the condition that κ defined in (3) should be set to value 1, by suitable amplitude scaling of the result. We find that for the quantizer function, the value of κ is $\kappa_q = -\sum_{k=1}^{M} \alpha_\kappa \delta f_k - \beta$ $= -\sum_{k=1}^{M} \alpha_k [\delta f_k + \delta A_k]$. For the above quantizer this is 0.0975, so that we should scale up the quantizer levels by 10.26. Note also that for $m=3$ the Fisher information $I(f)$ is 1.132, so that g_{LO} or its biased version must also be scaled by 1/0.132 or 7.58.

5. Robustness of Performance

One application of the idea of best approximation of and invariance to a linear-biased version of $g_{LO}(s)$ is in obtaining nonlinearities g that provide better robustness against deviations from the nominal assumption about the unit-variance f than is offered by $g_{LO}(s)$. In the case of a sub-Gaussian pdf of the type considered in the last section, the function $g_{LO}(s)$ may increase rapidly, as s^{m-1}, with $m>2$. The optimum nonlinearity maximizing the bss efficacy and yielding $\kappa=1$ is actually $g_o(s)=g_{LO}(s)/(I(f)-1)$. Suppose however that the actual pdf p is a mixture of the form $(1-\varepsilon)f_\sigma(s) + \varepsilon\eta(s)$, where f_σ is a variance-σ^2 version of f and η is a Gaussian pdf with variance >1 such that p has unit variance. This means that the actual pdf p has heavier Gaussian tails than the nominal. With $g_o(s)$ an odd power law nonlinearity of the form $s^{m-1}sgn(s)$, the algorithm performance may degrade considerably in the presence of the Gaussian "outlier" contamination. One approach to getting a more robust solution is to start by adding a negative linear bias, to obtain $g_o(s) - bs$. The resulting function descends from 0 to a negative value before eventually rising again towards large positive values. Suppose we now modify it by setting its value to 0 whenever $g_o(s)-bs$ becomes positive (for $s>0$, and symmetrically for $s<0$). By making b reasonably large, the mean-square discrepancy between the "limited" function and the original LO function (accordingly linearly-biased) is small, and the performance of the algorithm can be expected to remain good for the nominal case. In the presence of Gaussian contamination the limiter solution should provide better performance. Before using this function, the amplitude scale of the limiter should be adjusted to get a nominal value of 1 for the resulting κ value.



$g_o(s)$-bs *for* $m=3$ *Generalized Gaussian pdf*

Figure 2 *Limiter Functions, Generalized Gaussian (m=3) pdf*

This approach is equivalent to one in which the function $g_o(s)$ is modified beyond a point s_b to be *linear* with slope b; here $s_b b=g_o(s_b)$. The *linear increase* rather than the *power law increase* beyond s_b provides the robustness against more *Gaussian* rather than sub-Gaussian behavior in the tails. In Figure 2 we illustrate the nature of the robust nonlinear solutions that we are led to by the above considerations. The figure shows the limiter nonlinearities obtained for different values of b for the generalized Gaussian pdf with m=3. The dashed parts of the curves are the parts that are limited off to 0, the values of b increasing from left to right.

For b=8 and beyond, the bss efficacy using the limiter is more than 3/4 of the maximum for the nominal generalized Gaussian (m=3) pdf. Figures 3 and 4 show some typical results of running the algorithm with the unmodified (EASI) optimum g function and the limited version with b=8. These figures show the elements of the C matrix as a function of time. For Figure 3 the source pdf was the unit-variance nominal generalized Gaussian with m=3; the EASI algorithm used the optimum function $g_o(s)$ whereas the robust limiter used the amplitude scaled function of Figure 2 with b=8. In Figure 4 the source pdf was changed to a mixture pdf $p(s)=(1-\varepsilon)f_\sigma(s)+\varepsilon\eta_3(s)$

where $f_\sigma(s)$ is the m=3 generalized Gaussian pdf with variance $\sigma^2=7/9$ and the pdf η_3 is the Gaussian contaminating pdf with variance 3. This makes the variance of p remain 1, with ε=0.1

We are generally able to get rather better performance from the robust limiter. While we are able to obtain good robustness of performance with this approach, we have not yet succeeded in proving any specific minimax robustness property for useful classes of source pdf's. One complication is that the pdf's are constrained to be of unit-variance, so that results of the type established in earlier work on robust detection and estimation are not applicable.

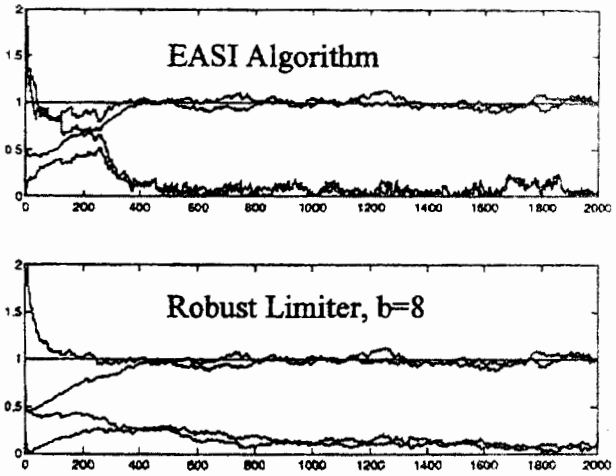


Figure 3 *Example Performance, Nominal Generalized Gaussian (m=3) pdf*

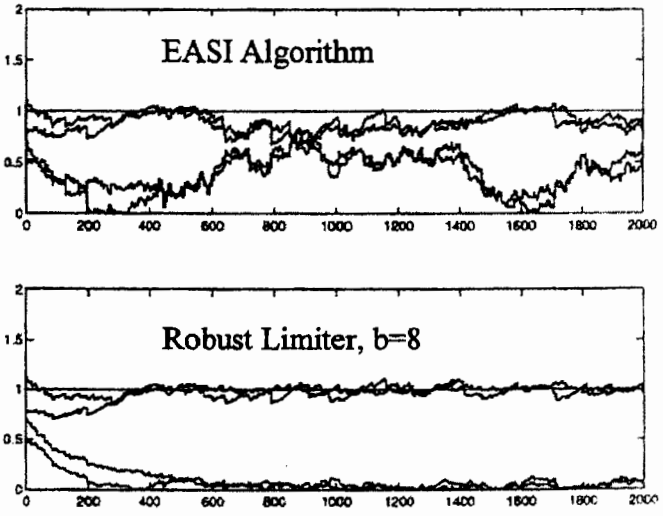


Figure 4 *Example Performance, Contaminated Generalized Gaussian (m=3) pdf*

One might consider also what happens in the case of super-Gaussian pdf's. For example, for the double-exponential source pdf (special case of generalized Gaussian with m=1) the function $g_o(s)$ is proportional to $sgn(s)$. The lack of robustness we are concerned with arises from the terms $\mathbf{g(y)y}^T$ in the algorithm of (2). (We will ignore the problem of robust pre-whitening here; one may implement a general version of this with appropriate nonlinear functions of $\mathbf{y}$ in place of $\mathbf{y}$ in $\mathbf{yy}^T$.) It is of interest to note that the algorithm of (2) is invariant not only to a linear-bias imposed on $g(s)$, but also to a $g(s)$-*bias* imposed on the linear term s. Thus, the algorithm is the same if we replace $g(y)$ by $[g(y) - by]$ and also y by $[y-ag(y)]$; we should have $ab \neq 1$ to prevent degeneracy. Just as we suggested truncating the function $[g_o(y)-by]$ above, we can also implement a modification of $[y-ag_o(y)]$ with a suitable value of a.

6. Conclusion

We have considered the choice of the nonlinearity $g(s)$ for the EASI algorithm, and by extension for a number of related approaches for blind source separation. Our results established how optimum quantized versions should be designed of nominally optimum g functions, and also suggest how robust performance can obtained by using suitably modified versions of the optimum functions. Exact minimax properties of the robust limiter-type functions remain to be proved.

Acknowledgement

This research was conducted under a Collaborative Research Grant from the NATO International Scientific Exchange Programme.

References

[1] J-F Cardoso, "Blind source separation: Statistical principles" *Proc. IEEE,* **86**, 2009-2025, October 1998

[2] J-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Sig. Process.*, **44**, 3017-3030, Dec. 1996

[3] P. Comon, "Independent component analysis: a new concept?" *Signal Process.*, **36**, 287-314, April 1994

[4] J.-F. Cardoso, "On the performance of orthogonal source separation algorithms," in *Proc. EUPISCO,* pp. 7776-779, Sept. 1994

[5] G. V. Moustakides, "Locally optimum adaptive signal processing algorithms," *IEEE Trans. Sig. Process.*, **46**, 3315-3325, Dec. 1998

[6] A. Hyvarinen, *Independent Component Analysis: A Neural Network Approach.* Ph.D. Thesis, Helsinki Univ. of Tech., Finland 1997

[7] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. Sig. Process.*, **45**, 1712-1725, July 1997

[8] S. A. Kassam, *Signal Detection in Non-Gaussian Noise.* Springer-Verlag, New York, 1988