# Simulated Annealing Based-GA Using Injective Contrast Functions for BSS

J.M. Górriz, C.G. Puntonet, J.D. Morales, and J.J. delaRosa

Facultad de Ciencias, Universidad de Granada,
Fuentenueva s/n, 18071 Granada, Spain
gorriz@ugr.es

**Abstract.** In this paper we present a novel GA-ICA method which converges to the optimum. The new method for blindly separating unobservable independent component signals from their linear mixtures (Blind Source Separation BSS), uses genetic algorithms (GA) to find the separation matrices which minimize a cumulant based contrast function. The paper also include a formal prove on the convergence of the proposed algorithm using guiding operators, a new concept in the genetic algorithms scenario. This approach is very useful in many fields such as biomedical applications i.e. EEG which usually use a high number of input signals. The Guiding GA (GGA) presented in this work converges to uniform populations containing just one individual, the optimum.

## 1 Introduction

The starting point in the Independent Component Analysis (ICA) research can be found in [1] where a principle of redundancy reduction as a coding strategy in neurons was suggested, i.e. each neural unit was supposed to encode statistically independent features over a set of inputs. But it was in the 90´s when Bell and Sejnowski applied this theoretical concept to the blindly separation of the mixed sources (BSS) using a well known stochastic gradient learning rule [2] and originating a productive period of research in this area [3]. In this way ICA algorithms have been applied successfully to several fields such as biomedicine, speech, sonar and radar, signal processing, etc. and more recently also to time series forecasting [4], i.e. using stock data.

In general, any abstract task to be accomplished can be viewed as a search through a space of potential solutions and whenever we work with large spaces, GAs are suitable artificial intelligence techniques for developing this optimization [4]. Such search requires balancing two goals: exploiting the best solutions and exploring the whole search space. In this work we prove how GA-ICA algorithms converge to the optimum. They work efficiently in the search of the separation matrix (i.e. EEG and scenarios with the BSS problem in higher dimension) proving the convergence to the optimum. We organize the essay as follows. In section 2 and 3 we give a brief overview of the basic ICA and GA theory. Then we introduce a set of genetic operators in sections 3 and 4 and prove convergence. Finally state some conclusions in section 6.

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3514, pp. 585–592, 2005.

## 2 ICA and Statistical Independence Criterion

We define ICA using a statistical latent variables model (Jutten & Herault, 1991). Assuming the number of sources $n$ is equal to the number of mixtures, the linear model can be expressed, using vector-matrix notation and defining a time series vector $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathbf{s}$, $\tilde{\mathbf{s}}$ and the matrix $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$ as:

$$\tilde{\mathbf{s}} = \mathbf{Bx} = \mathbf{BAs} = \mathbf{Gs} \tag{1}$$

where we define $\mathbf{G}$ as the overall transfer matrix. The estimated original sources will be, under some conditions included in Darmois-Skitovich theorem [5], a permuted and scaled version of the original ones. The statistical independence of a set of random variables can be described in terms of their joint and individual probability distribution. This is equivalent to [6]:

$$\mathit{\Pi} = \sum_{\{\lambda,\lambda^*\}} \beta_\lambda \beta^*_{\lambda^*} \mathbf{\Gamma}_{\lambda,\lambda^*} \qquad |\lambda| + |\lambda^*| < \tilde{\lambda} \tag{2}$$

where the expression defines a summation of cross cumulants [6] and is used as a fitness function in the GA. The latter function satisfies the definition of a contrast function $\mathit{\Psi}$ defined in [7] as can be seen in the following generalized proposition given in [8].

**Proposition 1.** *The criterion of statistical independence based on cumulants defines a contrast function $\mathit{\Psi}$ given by:*

$$\psi(\mathbf{G}) = \mathit{\Pi} - log|det(\mathbf{G})| - h(\mathbf{s}) \tag{3}$$

*where $h(\mathbf{s})$ is the entropy of the sources and $G$ is the overall transfer matrix.*

*Prove: To prove this proposition see Appendix A in [8] and apply the multilinear property of the cumulants.*

## 3 Genetic Algorithms: A Theoretical Background

Let $\mathcal{C}$ the set of all possible creatures in a given world and a function $f : \mathcal{C} \to R^+$, namely fitness function. Let $\mathit{\Xi} : \mathcal{C} \to \mathcal{V}_{\mathbf{C}}$ a bijection from the creature space onto the free vector space over $\mathcal{A}^\ell$, where $\mathcal{A} = \{\overline{a}(i), \quad 0 \leq i \leq a-1\}$ is the alphabet which can be identified by $\mathcal{V}_1$ the free vector space over $\mathcal{A}$. Then we can establish $\mathcal{V}_{\mathbf{C}} = \otimes_{\lambda=1}^{\ell} \mathcal{V}_1$ and define the free vector space over populations $\mathcal{V}_{\mathcal{P}} = \otimes_{\sigma=1}^{N} \mathcal{V}_{\mathbf{C}}$ with dimension $L = \ell \cdot N$ and $a^L$ elements. Finally let $S \subset \mathcal{V}_{\mathcal{P}}$ be the set of probability distributions over $\mathcal{P}_{\mathbf{N}}$, that is the state which identifies populations with their probability value.

**Definition 1.** *Let $S \subset \mathcal{V}_{\mathcal{P}}$, $n, k \in \mathcal{N}$ and $\{P_c, P_m\}$ a variation schedule. A Genetic Algorithm is a product of stochastic matrices (mutation, selection, crossover, etc..) act by matrix multiplication from the left:*

$$\mathbf{G^n} = \mathbf{F_n} \cdot \mathbf{C}^{\mathbf{k}}_{\mathbf{P^n_c}} \cdot \mathbf{M}_{\mathbf{P^n_m}} \tag{4}$$

*where* $\mathbf{F_n}$ *is the selection operator,* $\mathbf{C}^{\mathbf{k}}_{\mathbf{P}^{\mathbf{n}}_{\mathbf{c}}} = \mathbf{C}(K, P_c)$ *is the simple crossover operator and* $\mathbf{M}_{\mathbf{P}^{\mathbf{n}}_{\mathbf{m}}}$ *is the local mutation operator (see [6] and [10])*

In order to improve the convergence speed of the algorithm we could include another mechanisms such as elitist strategy (a further discussion about reduction operators, can be found in [11]). Another possibility is:

## 4 Guided GAs

In order to include statistical information into the algorithm we define an hybrid statistical genetic operator as follows. The value of the probability to go from individual $p_i$ to $q_i$ depends on contrast functions (i.e. based on cumulants) as: $P(\xi_{n+1} = p_i | \xi_n = q_i) = \frac{1}{\aleph(T_n)} \exp\left(-\frac{\Psi(p_i)+\Psi(q_i)}{T_n}\right)$; $p_i, q_i \in \mathbf{C}$ where $\aleph(T_n)$ is the normalization constant depending on iteration $n$; temperature follows a variation decreasing schedule, that is $T_{n+1} < T_n$ converging to zero, and $\Psi(q_i)$ is the value of the selected contrast function over the individual (an encoded separation matrix). This sampling (Simulated Annealing -SA- law) is applied to the population and offspring emerging from the canonical genetic procedure.

**Proposition 2.** *The guiding operator can be described using its associated transition probability function (t.p.f.) by column stochastic matrices* $\mathbf{M}^{\mathbf{n}}_{\mathbf{G}}$, $n \in \mathcal{N}$ *acting on populations.*

1. *The components are determined as follows: Let* $p$ *and* $q \in \wp_N$, *then we have*

$$\langle q, \mathbf{M}^{\mathbf{n}}_{\mathbf{G}} p \rangle = \frac{N!}{z_{0q}! z_{1q}! \ldots z_{a^L-1q}!} \prod_{i=0}^{a^L-1} \{P(i)\}^{z_i q}; \qquad p, q \in \mathcal{P}_{\mathbf{N}} \tag{5}$$

   *where* $z_{iq}$ *is the number of occurrences of individual* $i$ *on population* $q$ *and* $P(i)$ *is the probability of producing individual* $i$ *from population* $p$ *given above. The value of the guiding probability* $P(i) = P(i, \Psi)$ *depends on the fitness function used:*[1] $P(i) = \frac{z_{ip} \exp\left(-\frac{\Psi(p_i)+\Psi(q_i)}{T_n}\right)}{\sum_{i=0}^{a^L-1} z_{ip} \exp\left(-\frac{\Psi(p_i)+\Psi(q_i)}{T_n}\right)}$
2. *For every permutation* $\pi \in \Pi_N$, *we have* $\pi \mathbf{M}^{\mathbf{n}}_{\mathbf{G}} = \mathbf{M}^{\mathbf{n}}_{\mathbf{G}} = \mathbf{M}^{\mathbf{n}}_{\mathbf{G}} \pi$.
3. $\mathbf{M}^{\mathbf{n}}_{\mathbf{G}}$ *is an identity map on* $\mathbf{U}$ *in the optimum, that is* $\langle p, \mathbf{M}^{\mathbf{n}}_{\mathbf{G}} p \rangle = 1$*; and has strictly positive diagonals since* $\langle p, \mathbf{M}^{\mathbf{n}}_{\mathbf{G}} p \rangle > 0 \quad \forall p \in \mathcal{P}_{\mathbf{N}}$.
4. *All the coefficients of a GA consisting of the product of stochastic matrices: the simple crossover* $\mathbf{C}^{\mathbf{k}}_{\mathbf{P}_{\mathbf{c}}}$, *the local multiple mutation* $\mathbf{M}^{\mathbf{n}}_{\mathbf{Pm}}$ *and the guiding operator* $\mathbf{M}^{\mathbf{n}}_{\mathbf{G}}$ *for all* $n, k \in \mathcal{N}$ *are uniformly bounded away from 0.*

[1] The condition that must be satisfied the transition probability matrix $P(i, f)$ is that it must converge to a positive constant as $n \to \infty$ (since we can always define a suitable normalization constant). The fitness function or selection method of individuals used in it must be injective.

*Proof: (1) follows from the transition probability between states. (2) is obvious and (3) follows from [7] and checking how matrices act on populations. (4) follows from the fact that* $\mathbf{M}^{\mathbf{n}}_{\mathbf{P_m}}$ *is fully positive acting on any stochastic matrix* $\mathbf{S}$*.*

It can be viewed as a suitable fitness selection and as a certain Reduction Operator, since it preserves the best individuals into the next generation using a non heuristic rule, unlike the majority of GAs used.

The convergence and strong and weak ergodicity of the proposed algorithm can be proved using several ways. A MC modelling a CGA has been proved to be strongly ergodic (hence weak ergodic, see [10]). So we have to focus our attention on the transition probability matrix that emerges when we apply the guiding operator. We can write the overall process as:

$$\langle q, \mathbf{G}^{\mathbf{n}} p\rangle = \sum_{v\in\wp_N} \langle q, \mathbf{M}^{\mathbf{n}}_{\mathbf{G}} v\rangle\langle v, \mathbf{C}^n p\rangle \tag{6}$$

where $\mathbf{C}^n$ is the stochastic matrix associated to the CGA and $\mathbf{M}^{\mathbf{n}}_{\mathbf{G}}$ is given by equation 5.

**Proposition 3. Weak Ergodicity**
*A MC with transition probability function associated to guiding operators that converges to uniform populations (populations with the same individual) satisfies weak ergodicity.*

*Prove: If we define a GGA on CGAs, the ergodicity properties depends on the new defined operator since they satisfy them as we said before. To prove this proposition we just have to check the convergence of the t.p.f. of the guiding operator on uniform populations. If the following condition is satisfied:*

$$\langle u, \mathbf{G}^{\mathbf{n}} p\rangle \to 1 \quad u \in \mathbf{U} \tag{7}$$

*Then we can find a series of numbers which satisfies:*

$$\sum_{n=1}^{\infty} \min_{n,p}(\langle u, \mathbf{G}^{\mathbf{n}} p\rangle) = \infty \leq \sum_{n=1}^{\infty} \min_{q,p} \sum_{v\in\wp_N} \min\left(\langle v, \mathbf{M}^{\mathbf{n}}_{\mathbf{G}} p\rangle\langle v, \mathbf{C}^n q\rangle\right) \tag{8}$$

*which is equivalent to weak ergodicity [9].*

**Proposition 4. Strong Ergodicity**
*Let* $\mathbf{M}^{\mathbf{n}}_{\mathbf{P_m}}$ *describe multiple local mutation,* $\mathbf{C}^{\mathbf{k}}_{\mathbf{P_c^n}}$ *describe a model for crossover and* $\mathbf{F}^{\mathbf{n}}$ *describe the fitness selection. Let* $(P_m^n, P_c^n)_n \in \mathcal{N}$ *be a variation schedule and* $(\phi_n)_{n\in\mathcal{N}}$ *a fitness scaling sequence associated to* $\mathbf{M}^{\mathbf{n}}_{\mathbf{G}}$ *describing the guiding operator according to this scaling.* [2] *Let* $\mathbf{C}^{\mathbf{n}} = \mathbf{F}^{\mathbf{n}} \cdot \mathbf{M}^{\mathbf{n}}_{\mathbf{P_m}} \cdot \mathbf{C}^{\mathbf{k}}_{\mathbf{P_c^n}}$ *represent the first n steps of a CGA. In this situation,*

[2] A scaling sequence $\phi_n : (\mathcal{R}^+)^N \to (\mathcal{R}^+)^N$ is a sequence of functions connected with a injective fitness criterion $f$ as $f_n(p) = \phi_n(f(p)) \quad p \in \wp_N$ such that $\mathbf{M}^{\infty}_{\mathbf{G}} = \lim_{n\to\infty} \mathbf{M}^{\mathbf{n}}_{\mathbf{G}}$ exist.

$$v_\infty = \lim_{n\to\infty} \mathbf{G^n} v_0 = \lim_{n\to\infty} (\mathbf{M_G^\infty C^\infty})^n v_0 \tag{9}$$

*exists and is independent of the choice of* $v_0$*, the initial probability distribution. Furthermore, the coefficients* $\langle v_\infty, p\rangle$ *of the limit probability distribution are strictly positive for every population* $p \in \wp_N$*.*

*Prove: The demonstration of this proposition is rather obvious using the results of Theorem 16 in [10] and the point 4 in Proposition 2. In order to obtain the results of the latter theorem we only have to replace the canonical selection operator* $\mathbf{F_n}$ *with our guiding selection operator* $\mathbf{M_G^n}$ *which has the same essential properties.*

**Proposition 5. Convergence to the Optimum**
*Under the same conditions of propositions 3, 4 the GGA algorithm converges to the optimum.*

*Prove: To reach this result, one has to prove that the probability to go from any uniform population to the population containing only the optimum is equal to 1 when* $n \to \infty$*:*

$$\lim_{n\to\infty} \langle p^*, \mathbf{G^n} u\rangle = 1 \tag{10}$$

*since the GGA is an strongly ergodic MC hence any population tends to uniform in time. If we check this expression we finally have the equation 10. In addition we have to use point 3 in Proposition 2 to make sure the optimum is the convergence point. Thus any guiding operator following a simulated annealing law converges to the optimum uniform population in time.*

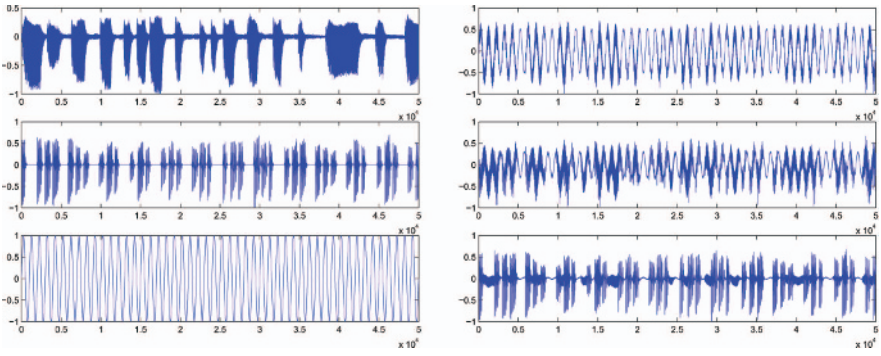
## 5 Simulations

At the first step, we compare the previous canonical method for apply GAs to ICA [12] with the GGA version for a reduced input space dimension ($n = 3$). The Computer used in these simulations was a PC 2 GHz, 256 MB RAN in the case of a low number of signals and the software used is an extension of ICATOOLBOX2.0 in MatLab code, protected by the Spanish law N° CA-235/04. We test these two algorithms for a set of independent signals plotted in figure 2(a) using 50 randomly chosen mixing matrices (50 runs); i.e. using the mixing matrix: $B = \{1.0000, -0.9500, 0.5700; -0.5800, 1.0000, 0.0900; 0.6300, -0.0100, 1.0000\}$, we get the signals shown in figure 2(b). We have chosen two super-gaussian signals and one bimodal signal for the first attempt ($n_i nps = 3$). The order of the statistics used is the same in both methods (cumulants of $4^{th}$order)[3] and the size

[3] Based on section 2, we can define the fitness function approach for BSS as:

$$f(p_o) = \sum_{i,j,\ldots} ||Cum(\overbrace{y_i, y_j, \ldots}^{stimes})|| \qquad \forall i, j, \ldots \in [1, \ldots, n] \tag{11}$$

where $p_o$ is the parameter vector (individual) containing the separation matrix and $||\ldots||$ denotes the absolute value.



(a) Original signals (b) Mixed signals

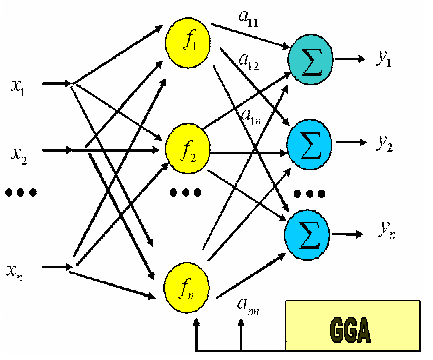
**Fig. 1.** Set of independent series used in the comparison GA-GGA and a mixed case

**Table 1.** Mean and deviation of the parameters in the separation over 50 runs for the cost function of $4^{th}$ order by the GA-method, GGA method and the FASTICA method. $1^{st}$ row GA-ICA, $2^{nd}$ row GGA-ICA and $3^{th}$ row FATICA

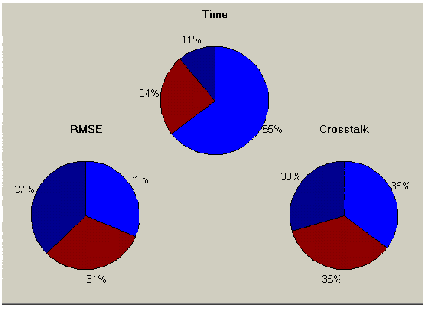
| param | $a_{11}$ | $a_{12}$ | $a_{13}$ | $a_{21}$ | $a_{22}$ | $a_{23}$ | $a_{31}$ | $a_{32}$ | $a_{33}$ |
|---|---|---|---|---|---|---|---|---|---|
| mean | -0.2562 | 0.1473 | -0.1657 | -0.0647 | -0.1393 | 0.2475 | -0.4910 | 0.0998 | -0.0350 |
| dev.(%) | $\leq 5$ | $\leq 5$ | $\leq 5$ | $\leq 5$ | $\leq 5$ | $\leq 5$ | $\leq 5$ | $\leq 5$ | $\leq 5$ |
| mean | -0.1481 | 0.1647 | -0.2564 | 0.1401 | -0.2464 | -0.0649 | -0.1003 | 0.0345 | -0.4914 |
| dev.(%) | $\leq 6.5$ | $\leq 6.5$ | $\leq 6.5$ | $\leq 6.5$ | $\leq 6.5$ | $\leq 6.5$ | $\leq 6.5$ | $\leq 6.5$ | $\leq 6.5$ |
| mean | 0.0756 | -0.1099 | 0.2271 | -0.0715 | 0.1648 | 0.0659 | 0.0512 | -0.0226 | 0.4435 |
| dev.(%) | $\leq 10$ | $\leq 10$ | $\leq 10$ | $\leq 10$ | $\leq 10$ | $\leq 10$ | $\leq 10$ | $\leq 10$ | $\leq 10$ |

of population was 100. In this way we can compare the search efficiency of both methods. Later we will focus our attention with a third statistical algorithm for ICA, the well-known FastICA [3]. This method uses the same level of information in its contrast function ($4^{th}$ order) thus the comparison is significant.

Results obtained from simulations are conclusive. We find out how the number of iterations (CPU time) needed to reach convergence is higher using the proposed method in [12]. This is due to blind search strategy used in the latter reference unlike the guided strategy proposed in this paper. We measure convergence by means of the well-known methods: Crosstalk (between original and recovered signals) and Normalized Round Mean Square Error (NRMSE). The set of recovering signals using GGA method can be found in figure 2(b). In the case of the three method comparison we observe how the efficiency of the FastICA in low dimension is better than the genetic approaches (see figure 2 somehow the standard deviation in time and error measure in higher than the genetic methods (see tables 1 and 2) since it suffers the local minima effect. As is shown in the latter tables the genetic procedures are slower but finally reach a better solution (the new proposed method is faster than the method in [12]).




(a) Schematic Representation of the Separation System in ICA-GA




(b) Comparison for number of inputs equal to 3 GGA (red) ,GA (light blue) and FastICA (blue). Observe how GA methods obtain the same level of recovery but the time efficiency is quite different

**Fig. 2.** Schematic representation and comparison of the 3 method

**Table 2.** Mean and deviation of the parameters in the separation over 50 runs for the cost function of $4^{th}$ order by the GA-method, GGA method and the FASTICA method (cont)

| Method | param. | Comp. Time(s) | NRMSE | Crosstalk(dB) |
|---|---|---|---|---|
| GA-ICA | mean | 10.21 | $1.5635^{-4}$ | -34.709 |
| | dev.(%) | $\leq 2$ | $\leq 1$ | $\leq 1$ |
| GGA-ICA | mean | 3.3 | $1.5408^{-4}$ | -37.7507 |
| | dev.(%) | $\leq 2$ | $\leq 1$ | $\leq 1$ |
| FastICA | mean | 1.64 | $1.6355^{-4}$ | -29.663 |
| | dev.(%) | $\leq 5$ | $\leq 2$ | $\leq 4$ |

Finally, we checked the performance of the proposed hybrid algorithm in a high dimensional scenario [6]. The results for the crosstalk were conclusive: FASTICA convergence rate decreases as dimension increases whereas GA approaches work efficiently. Of course we used the number of starting points equal to the number of individuals in the genetic generation.

## 6 Conclusions

A GGA-based BSS method has been developed to solve BSS problem from the linear mixtures of independent sources. The proposed method obtain a good performance overcoming the local minima problem over multidimensional domains.

Extensive simulation results prove the ability of the proposed method. This is particular useful in some medical applications where input space dimension increases and in real time applications where reaching fast convergence rates is the major objective. In this work we have focussed our attention to linear mixtures. The nonlinear problem can be interpreted as a piece-wise linear model and is expected that results improve even more since the higher parameters to encode the better results we obtain. GAs are the best strategies in high dimensional domains so it would be interesting how these algorithms (non CGAs) face the nonlinear ICA. The experimental work on this part is on the way. In the theoretical section we have prove the convergence of the proposed algorithm to the optimum unlike the ICA algorithms which usually suffer of local minima and non-convergent cases. Any injective contrast function can be used to build a guiding operator, as a elitist strategy i.e. the Simulated Annealing function defined in section 4. The convergence is shown under little restrictive conditions for the guiding operator: its effect must disappear in time like the simulated annealing.

## References

1. Barlow, H.B, Possible principles underlying transformation of Sensory messages. Sensory Communication, MIT Press, New York, (1961).
2. Bell,A.J. et al. An Information-Maximization Approach to BSS and Blind Deconvolution. Neural Comp., 7, 1129-1159 (1995).
3. Hyvärinen, A. et al. A fast fixed point algorithm for ICA Neural Comp., 9: 1483-1492
4. Górriz, J.M. et al. New Model for Time Series Forecasting using rbfs and Exogenous Data. Neural Comp. and Appl., 13/2 (2004)
5. Cao, X.R. et al. General Approach to BSS. IEEE Trans. on Signal Proc., 44/3, 562-571 (1996)
6. Górriz J.M. et al. Hybridizing GAs with ICA in Higher dimension LNCS 3195,414-421, (2004)
7. Comon, P., ICA, a new concept? Signal Proc. 36 (1994) 287-314
8. Cruces, S. et al. Robust BSS algorithms using cumulants. Neurocomp. 49 (2002) 87-118
9. Isaacson, D.L. et al. *MCs: Theory and Appli.*, Wiley, 1985.
10. Schmitt, L.M. et al. *Linear Analysis of GAs*, Theoretical Computer Science, 200, pp 101-134, 1998.
11. Rudolph, G., *Convergence Analysis of CGAs*, IEEE Trans. on NN, 5/1,(1994) 96-101.
12. Tan, Y. et al. Nonlinear BSS Using HOS and a GA. IEEE Trans. on Evol. Comp., 5/6 (2001)