Variable Transformations in Estimation of Distribution Algorithms

Davide Cucci, Luigi Malagò, and Matteo Matteucci

Department of Electronics and Information, Politecnico di Milano,
Via Ponzio, 34/5, 20133 Milano, Italy
{cucci,malago,matteucci}@elet.polimi.it

Abstract. In this paper we address model selection in Estimation of Distribution Algorithms (EDAs) based on variables trasformations. Instead of the classic approach based on the choice of a statistical model able to represent the interactions among the variables in the problem, we propose to learn a transformation of the variables before the estimation of the parameters of a fixed model in the transformed space. The choice of a proper transformation corresponds to the identification of a model for the selected sample able to implicitly capture higher-order correlations. We apply this paradigm to EDAs and present the novel Function Composition Algorithms (FCAs), based on composition of transformation functions, namely I-FCA and Chain-FCA, which make use of fixed low-dimensional models in the transformed space, yet being able to recover higher-order interactions.

Keywords: Function Composition Algorithm, Transformation of Variables, Minimization of Mutual Information, Chain Model.

1 Introduction

Estimation of Distribution Algorithms (EDAs) belong to the class of metaheuristics for optimization where the search is guided by a statistical model able to capture the interactions among the variables in the problem. When the model is not given a priori, model selection becomes crucial in order for the algorithm to be able to detect global optimal solutions. Indeed if the model chosen is not expressive enough, or if the wrong interactions are considered, model-based search strategies are prone to converge to local optima, cf. [16,10,6].

As a consequence, much of the literature in the EDAs community is focused on applying efficient algorithms for model selection, able to identify the correct interactions of the function from a sample of observations. Among the others we mention algorithms which reconstruct the topology of a Bayesian Network, as in BOA [12], clustering algorithms for the variables that appear to be correlated, eCGA [8], or model selection for Markov Random Fields (MRFs), as in DEUM [13]. When no prior information about the problem is available, EDAs need an efficient and scalable policy for model selection. In the general case learning an accurate model is exponential in the number of variables thus it is a

C.A. Coello Coello et al. (Eds.): PPSN 2012, Part I, LNCS 7491, pp. 428–437, 2012.

common practice to reduce the search space for the models, for example by limiting the interactions considered to the second order, when learning a MRF, by constraining the number of incoming edges in a BN or employing variable clustering techniques as in [14]. These restrictions can limit the performance of the algorithms in presence of certain structures of interactions among the variables.

In this paper we propose an approach to the problem of model selection based on the idea of applying a transformation of the variables and then employing a fixed low dimensional statistical model in the new transformed space. Obviously we moved much of the computational complexity from model selection to the choice of a good transformation; on the other side it becomes easier to select models able to capture higher-order interactions among the variables. Instead of limiting the search up to a given order of interactions, due to the family of transformations we introduce, we are able to identify non hierarchical models that can be efficiently employed in an EDAs.

To the best knowledge of the authors, the approach of transforming the original variables first appeared in [17], where the UMDA [11] is run over a set of new variables obtained applying ICA on the selected sample. More recently, Toussaint proposed to employ compression algorithms to the sample of promising solutions [15], Cho and Zhang cluster similar individuals in a group and explain the high order interactions with latent variables [3], Grosset et al. introduce physically meaningful auxiliary variables related to the application domain [7].

The paper is organized as follows. In Section 2 we review the MBS approach to optimization. In Section 3 we present the idea of employing variable transformations to perform model selection. In Section 4 we describe the Function Composition Algorithms (FCAs) family. First we review and discuss more in detail I-FCA, originally presented in [4], next we introduce a novel algorithm called Chain-FCA, which makes use of a fixed chain model. In Section 5, we discuss and compare the preliminary performance of the above-mentioned algorithms. In Section 6, we conclude by presenting some future directions of research.

2 Model Based Search and Stochastic Relaxation

We are interested in the minimization of a real-valued function f defined over a vector of binary variables. For mathematical convenience and without loss of generality we consider values in $\{\pm 1\}$, rather than classic 0/1 encoding. Let us introduce the notation that will be used in the following. Let $x = (x_1, \ldots, x_n) \in \Omega = \{\pm 1\}^n$ a vector of n binary variables, any $f : \Omega \mapsto \mathbb{R}$ can be represented uniquely as a square-free polynomial, i.e., the finite sum of monomials

$$f(x) = \sum_{\alpha \in F} c_\alpha x^\alpha, \quad c_\alpha \in \mathbb{R}^n, \tag{1}$$

where we employed the multi-index notation $\alpha = (\alpha_1, \ldots, \alpha_n) \in F \subset \{0,1\}^n$, and $x^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$. For instance, let $n = 3$ and $f = x_1x_2 + x_2x_3$, then $F = \{(1,1,0),(0,1,1)\}$. The monomials $\{x^\alpha\}$ with $\alpha \in \{0,1\}^n$ defines a basis for any function, while those identified by F correspond to the interactions present in f.

The paradigm of Model-Based Search (MBS) in stochastic optimization, consists in finding the minimum of f by solving the optimization problem of the stochastic relaxation of f, i.e., the minimization of the expected valued of f with respect to a density in a statistical model $\mathcal{M}$.

From now on, we consider models $\mathcal{M}$ that belong to the exponential family of probability distributions of the form

$$p(x;\theta) = \exp\left\{\sum_{i=1}^{k} \theta_i T_i(x) - \psi(\theta)\right\}, \quad \theta \in \mathbb{R}^k, \tag{2}$$

where $\theta = (\theta_1, \ldots, \theta_k)$ is the vector of natural parameters, $T_i(x) : \Omega \to \mathbb{R}$ are the sufficient statistics, and $\psi(\theta)$ is a normalizing factor. Such choice is not restrictive, indeed many models used in MBS belong to this family, such as log-linear models, the Gibbs distribution, and more in general MRFs.

Since the sum of the sufficient statistics is a function defined over Ω, with no prior information about f, it is convenient to choose the basis $\{x^\alpha\}$ itself as the set of sufficient statistics. However, the basis has $2^n - 1$ monomials, so is computationally intractable. For this reason, we consider lower-dimensional models identified by a subset of the sufficient statistics identified by a small subset of indices $M \subset \{0,1\}^n$, usually polynomial in n. Each monomial in M identifies one of the possible correlations between groups of variables.

The choice of the model is central in MBS. From a theoretical point of view, the best choice would be to chose $\mathcal{M}$ such that $M \triangleq F$, so that the stochastic relaxation admits no local minima [16]. Models with smaller number of monomials may admit local minima, so that algorithms are more prone to convergence to local minima for f. On the other side, larger models imply more computational costs for parameter estimation. Dealing with the exponential family, one possible approach for model-selection is to test all possible second-order interactions, and then in case move to higher-order correlations, as in e.g. [13]. The computational complexity of these techniques grows with the maximum order of f in Equation (1), c.f. [6]. Dealing with BNs, hBOA [12] solves this issue by introducing trees between variables, which allow to efficiently learn hierarchies between variables and thus higher-order correlations. Instead of employing standard statistical techniques able to learn high-dimensional models directly, we propose to employ variable transformations to perform implicit model selection.

3 Variable Transformations

In this section we describe an implicit approach to model selection in MBS, and in particular in EDAs, where the problem of identifying a model is replaced by a search for a transformation of the variables of f. By employing a fixed model for the transformed variables, we are implicitly choosing a different model in the original space, which depends on the transformation. We are interested in those transformations such that a model in the transformed space corresponds to a model in the original space which is able to capture the interactions of f.

Let us introduce a new vector of variables $y = (y_1, \ldots, y_n)$ in Ω and a one-to-one map h such that $y = h(x)$. We can express f as the composition of a function $g(y) : \Omega \to \mathbb{R}$ with h, i.e., $f = g \circ h$, and $g = f \circ h^{-1}$. Since h defines a permutation of the points in Ω, follows that $\min g = \min f$.

We can express h component-wise, i.e., $h = (h_1(x), \ldots, h_n(x)$. Since $h_i(x) : \Omega \to \{\pm 1\}$, each h_i admits an expansion as in Equation (1), i.e.,

$$h_i = \sum_{\alpha \in H_i} c_{i,\alpha} x^{\alpha}, \quad 1 \leq i \leq n. \tag{3}$$

Let $q(y;\xi) \in \mathcal{N}$ be a density for Y from the exponential family in (2), by expanding all the products obtained by substituting y_i with $h_i(x)$, we obtain a polynomial in x whose monomials are identified by a set of indices M, i.e.,

$$\begin{aligned} &\exp\left\{\sum_{\alpha \in N} \xi_\alpha y^\alpha - \phi(\xi)\right\} = \exp\left\{\sum_{\alpha \in N} \xi_\alpha \prod_{i=1}^{n} (h_i)^{\alpha_i} - \phi(\xi)\right\} = \\ &\exp\left\{\sum_{\alpha \in N} \xi_\alpha \prod_{i=1}^{n} \Big(\sum_{\gamma \in H_i} c_{i,\gamma} x^\gamma\Big)^{\alpha_i} - \phi(\xi)\right\} = \exp\left\{\sum_{\beta \in M} l_\beta(\xi) x^\beta - \psi(\xi)\right\}. \end{aligned} \tag{4}$$

By setting $\theta_\beta = l_\beta(\xi) \in \mathbb{R}$, we expressed $q(y, \xi)$ as the probability distribution $p(x;\theta)$ for the original variables, where l_β is a function which maps the parameters of the two exponential families. Notice that since h is one-to-one the dimension of the θ and the ξ parameter space are the same.

In other words, suppose we apply a transformation from x to y and consider an exponential family $\mathcal{N} = \{q(y;\xi), \xi \in \mathbb{R}^k\}$ over Y identified by the sufficient statistics in N. By Equation (4), $q(y;\xi) = p(x;\theta)$, with $y = h(x)$ and $\theta = l(\xi)$, thus $\mathcal{N}$ maps into the exponential family $\mathcal{M}$ for X, identified by a different set of sufficient statistics M. Such mapping is one-to-one, so that $\mathbb{E}_p[f] = \mathbb{E}_q[g]$, and $\min_{q \in \mathcal{N}} \mathbb{E}_q[g] = \min_{p \in \mathcal{M}} \mathbb{E}_p[f]$, so that the minimization of stochastic relaxation of f with respect to $\mathcal{M}$ is equivalent to those of g with respect to $\mathcal{N}$. Consider the following example. The function $f = x_1x_2 + x_2x_3$, $x \in \{\pm 1\}^3$ admits two global minima $x = (-1, 1, -1)$ and $(1, -1, 1)$. Let us apply following one-to-one map $y = h(x)$ and its inverse h^{-1}

$$h : \begin{cases} y_1 = x_1x_2 \\ y_2 = x_2x_3 \\ y_3 = x_3 \end{cases} \qquad h^{-1} : \begin{cases} x_1 = y_1y_2y_3 \\ x_2 = y_2y_3 \\ x_3 = y_3 \end{cases}$$

Let $\mathcal{N}$ be the exponential family defined over Y with $\{y_1, y_2, y_3\}$ as sufficient statistics. By expanding $q(y;\xi) \in \mathcal{N}$ we have that

$$\begin{aligned} \mathcal{N} \ni q(y, \xi) &= \exp\{\xi_1 y_1 + \xi_2 y_2 + \xi_3 y_3 - \phi(\xi)\} = \\ &= \exp\{\theta_1 x_1 x_2 + \theta_2 x_2 x_3 + \theta_3 x_3 - \psi(\theta)\} = p(x, \theta) \in \mathcal{M}, \end{aligned}$$

where $\theta = \xi$ and $\psi(\theta) = \phi(\xi)$. The sufficient statistics of $\mathcal{M}$ include the interactions on f, i.e., $F \subset M$. Follows that the stochastic relaxation of f with respect

to $\mathcal{M}$ does not admint local minima, for every $p \in \mathcal{M}$ the gradient of $\mathbb{E}_p[f]$ points into the direction of the global optimum of the relaxed problem, c.f., [10]. We can explicitly compute $g(y) = f \circ h^{-1}$. Since in the binary case $x_i^2 = 1$, $g = h_1^{-1}h_2^{-1} + h_2^{-1}h_3^{-1} = y_1 + y_2$, which is linear in y. The minimization of f can thus be performed considering the stochastic relaxation of g with respect to $\mathcal{N}$. This problem is simpler than the original one since we are minimizing the expected value of a linear function with respect to the independence model and the stochastic relaxation does not admit local minima, c.f., [9]. This example shows how the use of a properly chosen variable transformation can greatly affect the complexity of an optimization problem from the point of view of model-based search strategies.

4 Function Composition Algorithms

In this section we present the idea of learning a transformation of the variables before estimating the parameters of a fixed low-dimensional model in the transformed space. Such approach to model selection is applied to the EDAs paradigm, leading to a novel family of algorithm called Function Composition Algorithms (FCAs). Preliminary work appeared in [4].

Recall the basic iteration of an EDA,

$$\mathcal{P}^t \xrightarrow{\text{selection}} \mathcal{P}_s^t \xrightarrow{\text{estimation}} p(x;\theta^t) \in \mathcal{M} \xrightarrow{\text{sampling}} \mathcal{P}^{t+1}.$$

At each iteration t of an EDA, a subset $\mathcal{P}_s^t$ of the population $\mathcal{P}^t$ is selected according to a given selection policy. Then, a statistical model $\mathcal{M}$ is learned from the subsample, and the parameters of a distribution $p(x;\theta^t)$ are estimated. Finally, a new population $\mathcal{P}^{t+1}$ is generated by sampling. Some algorithms, such as PBIL [1] or UMDA [11], make the assumption of independent variables, others use low-dimensional models, such as the chain model, see MIMIC [5], while more powerful EDAs, e.g., hBOA [12] or DEUM [13,2] perform model selection in a larger class of models, able to capture higher-order correlations among variables.

In FCA, we implicitly learn a model by first choosing a variable transformation, and then using a fixed model for the new set of transformed variables. We introduce the following variation of the iteration of an EDA. Estimation and sampling are preceded and followed by two transformations. First a one-to-one map $y = h(x)$ is applied to each individual in the selected population obtaining $\tilde{\mathcal{P}}_s^t$, then after sampling from the estimated distribution, the population $\tilde{\mathcal{P}}^{t+1}$ is transformed back to the original space by means of h^{-1}, i.e.,

$$\mathcal{P}_s^t \xrightarrow{h} \tilde{\mathcal{P}}_s^t \xrightarrow{\text{estimation}} q(y;\xi^t) \in \mathcal{N} \xrightarrow{\text{sampling}} \tilde{\mathcal{P}}^{t+1} \xrightarrow{h^{-1}} \mathcal{P}^{t+1}.$$

From Equation (4), estimating a probability distribution $q(y;\xi) \in \mathcal{N}$ for the transformed sample $\tilde{\mathcal{P}}_s^t$ is equivalent to estimate a distribution $p(x;\theta) \in \mathcal{N}$ for $\mathcal{P}_s$. In general $\mathcal{N}$ and $\mathcal{M}$ are different, since the latter depends on the map h employed, so that the choice of h corresponds to choice of a model $\mathcal{M}$.

4.1 Independence-FCA

In the following we briefly review I-FCA, first introduced in [4]. I-FCA employs the independence model for the transformed variables y, and if the map h is properly chosen, the resulting low-dimensional model $\mathcal{M}$ can achieve a better approximation of the sample $\mathcal{P}_s$ with respect to the independence model for x.

The non-linear maps used in I-FCA are defined as follows. Consider the maps h indexed by j, $k \in \{1, \ldots, n\}$, with $j \neq k$, such that each h_i is defined as

$$h_i^{(j,k)} : \begin{cases} y_i = x_i x_k & \text{if } i = j \\ y_i = x_i & \text{otherwise.} \end{cases}$$

We have $n(n-1)$ different $h^{(j,k)}$ transformations. It is easy to see that they are one-to-one and that $h^{-1} = h$, since $x_i^2 = 1$. Next we extend the class of transformations we consider by allowing elements h to be the composition of a finite number $\overline{m}$ of maps of the form $h^{(j,k)}$:

$$h = h^{(j_1,k_1)} \circ \ldots \circ h^{(j_m,k_m)} \circ \ldots \circ h^{(j_{\overline{m}},k_{\overline{m}})}. \tag{5}$$

Since the inverse of each transformation in the sequence of compositions is the element itself, it is easy to see that h^{-1} is the composition of all the $h^{(j_m,k_m)}$ in the reversed order. Moreover, if the sufficient statistics of $\mathcal{N}$ are monomials in y, the sufficient statistics of the resulting model $\mathcal{M}$ for X are monomials in x.

In I-FCA we propose a strategy for the choice of map h based on the maximization of the likelihood of the transformed selected sample $\tilde{\mathcal{P}}_s$ with respect to the estimated distribution $q(y, \hat{\xi}) \in \mathcal{N}$, where $\mathcal{N}$ is the independence model for Y. This is equivalent to minimize the Kullback-Leibler divergence between the empirical distribution representing the selected population and its projection on the independence model (i.e. $KLD[\tilde{\mathcal{P}}_s \| q(y, \hat{\xi})] = -H[\tilde{\mathcal{P}}_s] - \mathcal{L}[\tilde{\mathcal{P}}_s \| q(y, \hat{\xi})]$), which gives a measure of the loss of information which occurs when $\tilde{\mathcal{P}}_s$ is approximated with $q(y, \xi)$. Note that since h is one-to-one $H[\tilde{\mathcal{P}}_s]$ does not depend oh h.

In order to make the search for h feasible, we choose a greedy approach. We initialize h to be the identity map $y = x$, then we iteratively examine all the $n(n-1)$ maps $h^{(j,k)}$ and compose the h map obtained at the previous step with the map $h^{(j,k)}$ which better improves the likelihood of $(h \circ h^{(j,k)})(\mathcal{P}_s)$ with respect to the independence model. The iteration stops when no improvement in the likelihood is achievable composing further maps of the form $h^{(j,k)}$ or when the maximum number $\overline{m}$ of transformations in h has been reached.

The representation of h as a composition of maps of the form $h^{(j,k)}$ is highly redundant, i.e., there exists more than one sequence of indices (j_m, k_m) which transforms the independence model $\mathcal{N}$ to the same exponential family $\mathcal{M}$. As a consequence, in order to reduce the complexity of the search strategy for h, we discard maps that produce models already examined in earlier stages of the search process. Each time a new map $h^{(j,k)}$ is considered, the sufficient statistics y_i are transformed and the corresponding monomials $x^\beta = y_i$ with $\beta \in M$ are computed. Next, maps for which the monomial x^β does not contain x_i, i.e., $\beta_i = 1$, or, for all i, the degree of x^β decrease when $h^{(j,k)}$ is applied, are discarded.

The worst case time complexity of the greedy search strategy for h is $\mathcal{O}(n^2\overline{m}N)$, where n is the number of variables in f and N is the population size. Note that it is possible to take advantage of the following log-likelihood decomposition:

$$\mathcal{L}[\tilde{\mathcal{P}}_s \| q(y, \hat{\xi})] = \frac{1}{N} \sum_{y \in \tilde{\mathcal{P}}_s} \log \Big(\prod_{i=1}^{n} q_i(y_i; \hat{\xi}_i) \Big) = \frac{1}{N} \sum_{i=1}^{n} \overbrace{\sum_{y \in \tilde{\mathcal{P}}_s} \log q_i(y_i; \hat{\xi}_i)}^{\mathcal{L}_i},$$

since q belongs to the independence model. When a new map $h^{(j_m,k_m)}$ is considered, since $y_i = x_i$, for $i \neq j$, we do not need to compute the terms $\mathcal{L}_i$ and the values already evaluated at the previous step $m-1$ can be used.

4.2 Chain-FCA

The variable transformation paradigm is general, and different models can be chosen for transformed variables. In the following we introduce Chain-FCA, a novel algorithm in FCAs family, where we fix a model with interactions, rather than the independence model, as in I-FCA. Consider the family of probability distributions for which the joint probability function factorizes as

$$p(y, \xi) = p(y_1) \prod_{j=2}^{n} p(y_j | y_{j-1}). \tag{6}$$

This is a chain model whose structure is fixed and each variable except the first depends only on the previous one. The parameter vector ξ has $2(n-1)+1$ components, one for the marginal probability of y_1, and two for each of the conditional probabilities, and can be easily estimated by means of max-likelihood estimation. The log-likelihood of a sample with respect to this model is given by

$$\mathcal{L}[\tilde{\mathcal{P}}_s \| q(y, \hat{\xi})] = \sum_{j=1}^{n-1} I(Y_j | Y_{j+1}) - \sum_{j=1}^{n} H(Y_j), \tag{7}$$

where $H(Y_j)$ is the marginal entropy and $I(Y_j|Y_k)$ is the mutual information.

Chain-FCA employs the chain model defined in (6) and a greedy search strategy to choose the sequence of maps $h^{(j,k)}$ which maximizes the likelihood of the transformed set of selected individuals $\tilde{\mathcal{P}}_s$. The order of the variables in the chain is fundamental. For this reason the class of the maps h is enriched by allowing the swap of couples of variables. This operation is equivalent to the composition of three maps $h^{(j,k)}$. Consider for example two variables x_1, x_2 and the map $y = h_1^{(1,2)} \circ h_2^{(2,1)} \circ h_3^{(1,2)}$. It turns out that $y_1 = x_2$ and $y_2 = x_1$, in fact

$$\{x_1, x_2\} \overset{h^{(1,2)}}{\Rightarrow} \{\overbrace{x_1 x_2}^{y_1}, \overbrace{x_2}^{y_2}\} \overset{h^{(2,1)}}{\Rightarrow} \Rightarrow \{\overbrace{x_1 x_2}^{y_1}, \overbrace{x_1}^{y_2}\} \overset{h^{(1,2)}}{\Rightarrow} \{\overbrace{x_2}^{y_1}, \overbrace{x_1}^{y_2}\}$$

The map h implies the swap of the variables x_1 and x_2. Notice that this was useless in I-FCA since the order of the variables is not relevant in the independence model, and such maps are discarded a priori. On the other hand, in Chain-FCA this allows to implicitly adapt the fixed structure of the interactions among

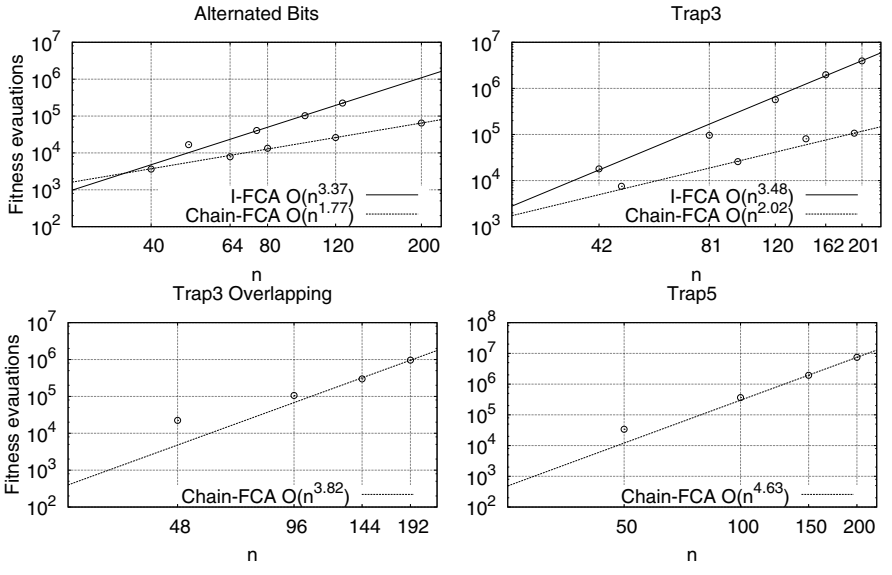

Fig. 1. Scalability of I-FCA and Chain-FCA

the variables in the chain model to the ones appearing in the set of candidate solutions. Notice that if the search for h is restricted to consider *only* variables swaps the result is a model selection algorithm very similar to MIMIC [5].

By means of a fixed structure chain model and a greedy search strategy for the choice of the transformation, Chain-FCA is able to implicitly learn a richer model compared to I-FCA, characterized by $2(n-1)+1$ parameters. The worst case time complexity of Chain-FCA is $\mathcal{O}(n^2\overline{m}N)$, since only $\frac{n(n-1)}{2}$ variables swaps have to be examined, along with the $n(n-1)$ maps of the form $h^{(j,k)}$.

5 Experimental Results

In this section we present the results of a preliminary scalability evaluation for I-FCA and for the novel Chain-FCA algorithm, over a set of well known benchmarks functions: Alternated Bits, Trap3, Trap3 overlapping, and Trap5. In Alternated Bits the variables interact in a chain structure and higher fitness is given to the instances for which the variables take opposite values with respect to their neighbors in the chain. Trap3 and Trap5 are deceptive functions and are composed of independent blocks of 3 and 5 variables, respectively. Each block has a global optimum and a deceptive local optimum. Trap3 overlapping is similar with respect to Trap3 but the blocks fully overlap.

In our algorithms we perform truncation selection and we choose the best S individuals. After a preliminary parameter tuning we fix $S = 10n$ for I-FCA and $S = 5n$ for Chain-FCA for all the problems considered, independently from the population size. Experiments show that in the case of I-FCA no improvement is

achievable when $\overline{m} > n$. This result is also supported by an empirical analysis on the set of models $\mathcal{M}$ which can be obtained mapping the independence model into the original space through h. In the case of Chain-FCA the class of the models that can be obtained by means of h is wider, so we set $\overline{m} = 4n$.

For each problem and for each dimension we determine the population size which ensure at most one failure out of 24 run. Then we estimated the average number of fitness evaluations performed when the global optimum for f first appears in the population. The results are presented in Figure 1, along with an asymptotic estimation obtained by means of a least square fit of the curve an^b. Notice that, with the only exception of Alternated Bits, these functions are *not* solved by other EDAs based on a fixed or low-dimensional model such as PBIL, UMDA, or MIMIC, and in general one has to move to more EDAs which perform model selection on a large class of complex models, such as BOA or DEUM. I-FCA solves Alternated Bits and Trap3 while Chain-FCA robustly solves all the benchmark functions considered. This proves the viability of the variable transformations approach. Both algorithms are part of the Evoptool toolkit. Source code and the detailed experimental settings are public availabe[1].

6 Conclusions and Future Works

Variables transformations can be employed as an alternative approach to model selection in MBS. In this paper we presented theoretical foundations of such approach, and proposed a novel algorithm in the FCA family, called Chain-FCA, which chooses a variable transformation maximizing maximum likelihood with respect to a fixed chain model. Both I-FCA and Chain-FCA choose the variable transformation to apply by iteratively composing basic modular maps. Besides the usual EDAs parameters, selection policy and population size, these algorithms have one more parameter which is the length of the composition sequence in the variable transformation, even though we argue that this parameter is problem independent and could be fixed a priori.

A preliminary experimental evaluation of the performances of Chain-FCA compared to I-FCA showed that these algorithms are able to solve functions characterized by higher-order interaction yet only employing fixed low dimensional models. This shows the viability of the variable transformation approach.

Some directions of future works include testing on different and more complex benchmark functions, experimenting more expressive classes of variables transformations and different models for the transformed variables, such as Chow-Liu trees. Performance enhancements could also come by the replacement of the greedy search strategy for h with more advanced policies.

References

1. Baluja, S., Caruana, R.: Removing the genetics from the standard genetic algorithm. In: Machine learning: proceedings of the Twelfth International Conference on Machine Learning, pp. 38–46. Morgan Kaufmann (1995)

[1] `http://airlab.elet.polimi.it/index.php/Evoptool`

2. Brownlee, A.E.I., McCall, J.A.W., Shakya, S.K., Zhang, Q.: Structure Learning and Optimisation in a Markov Network Based Estimation of Distribution Algorithm. In: Chen, Y.-p. (ed.) Exploitation of Linkage Learning. ALO, vol. 3, pp. 45–69. Springer, Heidelberg (2010)
3. Cho, D., Zhang, B.: Evolutionary optimization by distribution estimation with mixtures of factor analyzers. In: Proceedings of the 2002 Congress on Evolutionary Computation, CEC 2002, vol. 2, pp. 1396–1401 (2002)
4. Corsano, E., Cucci, D., Malagò, L., Matteucci, M.: Implicit model selection based on variable transformations in estimation of distribution. In: Learning and Intelligent OptimizatioN Conference LION 6. LNCS, vol. 7219. Springer (to apppear, 2012)
5. De Bonet, J., Isbell, C., Viola, P.: Mimic: Finding optima by estimating probability densities. In: Advances in Neural Information Processing Systems, p. 424. The MIT Press (1996)
6. Echegoyen, C., Zhang, Q., Mendiburu, A., Santana, R., Lozano, J.: On the limits of effectiveness in estimation of distribution algorithms. In: 2011 IEEE Congress on Evolutionary Computation (CEC), pp. 1573–1580 (June 2011)
7. Grosset, L., LeRiche, R., Haftka, R.: A double-distribution statistical algorithm for composite laminate optimization. Structural and Multidisciplinary Optimization 31, 49–59 (2006)
8. Harik, G.: Linkage learning via probabilistic modeling in the eCGA, 1999. Harik, G. R (1999); Linkage Learning via Probabilistic Modeling in the ECGA (IlliGAL Report No. 99010). University of Illinois at Urbana-Champaign
9. Hohfeld, M., Rudolph, G.: Towards a theory of population-based incremental learning. In: Proceedings of the 4th IEEE Conference on Evolutionary Computation, pp. 1–5. IEEE Press (1997)
10. Malagò, L., Matteucci, M., Pistone, G.: Towards the geometry of estimation of distribution algorithms based on the exponential family. In: Proceedings of the 11th Workshop on Foundations of Genetic Algorithms, FOGA 2011, pp. 230–242. ACM, New York (2011)
11. Mühlenbein, H., Mahnig, T.: Mathematical analysis of evolutionary algorithms. In: Essays and Surveys in Metaheuristics, Operations Research/Computer Science Interface Series, pp. 525–556. Kluwer Academic Publishers (2002)
12. Pelikan, M., Goldberg, D.: Hierarchical Bayesian Optimization Algorithm. In: Pelikan, M., Sastry, K., Cant Paz, E. (eds.) Scalable Optimization via Probabilistic Modeling. SCI, vol. 33, pp. 63–90. Springer, Heidelberg (2006)
13. Shakya, S., Brownlee, A., McCall, J., Fournier, F., Owusu, G.: A fully multivariate DEUM algorithm. In: IEEE Congress on Evolutionary Computation (2009)
14. Thierens, D.: The Linkage Tree Genetic Algorithm. In: Schaefer, R., Cotta, C., Kołodziej, J., Rudolph, G. (eds.) PPSN XI. LNCS, vol. 6238, pp. 264–273. Springer, Heidelberg (2010)
15. Toussaint, M.: Compact Genetic Codes as a Search Strategy of Evolutionary Processes. In: Wright, A.H., Vose, M.D., De Jong, K.A., Schmitt, L.M. (eds.) FOGA 2005. LNCS, vol. 3469, pp. 75–94. Springer, Heidelberg (2005)
16. Zhang, Q.: On stability of fixed points of limit models of univariate marginal distribution algorithm and factorized distribution algorithm. IEEE Transactions on Evolutionary Computation 8(1), 80–93 (2004)
17. Zhang, Q., Allinson, N., Yin, H.: Population optimization algorithm based on ica. In: 2000 IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks, pp. 33–36 (2000)