Review and Study of Genotypic Diversity Measures for Real-Coded Representations

Guillaume Corriveau, Raynald Guilbault, Antoine Tahan, and Robert Sabourin

***Abstract*—The exploration/exploitation balance is a major concern in the control of evolutionary algorithms (EAs) performance. Exploration is associated with the distribution of individuals on a landscape, and can be estimated by a genotypic diversity measure (GDM). In contrast, exploitation is related to individual responses, which can be described with a phenotypic diversity measure. Many diversity measures have been proposed in the literature without a comprehensive study of their differences. This paper looks at surveys of GDMs published over the years for real-coded representations, and compares them based on a new benchmark, one that allows a better description of their behavior. The results demonstrate that none of the available GDMs is able to reflect the true diversity of all search processes. Nonetheless, the normalized pairwise diversity measurement (D_{PW}^{N}) proves to be the best genotypic diversity measurement for standard EAs, as it shows nondominated behavior with respect to the desired GDM requirements.**

***Index Terms*—Diversity measures, evolutionary algorithms, exploration/exploitation balance, premature convergence.**

I. Introduction

ONE OF THE major problems with evolutionary algorithms (EAs) is premature convergence [1]–[4]. However, no method exists that offers adequate control of this phenomenon. The origin of premature convergence is the exploration/exploitation balance (EEB) [5]. Too much exploration leads to random searching and a waste of computational resources, while too much exploitation leads to local searching and premature convergence. This balance could be controlled by setting the EA parameters [6]. Here, we consider parameter-setting in the broad sense of the term. For example, the population number, the type of evolution model, and restart strategies are all possible options for controlling the EEB. It is worth noting that the EEB dilemma is not unique to EAs, as it is essentially a resource allocation problem that any adaptive system must face [7]–[9].

Manuscript received July 19, 2010; revised November 16, 2010, May 5, 2011 and September 5, 2011; accepted September 14, 2011. Date of publication February 10, 2012; date of current version September 27, 2012. This work was supported in part by the Fonds Quebecois de Recherche sur la Nature et les Technologies.

G. Corriveau, R. Guilbault, and A. Tahan are with the Department of Mechanical Engineering, Ecole de Technologie Superieure, Montreal, QC H3C 1K3, Canada (e-mail: guillaume.corriveau@etsmtl.ca; raynald.guilbault@etsmtl.ca; antoine.tahan@etsmtl.ca).

R. Sabourin is with the Department of Automated Manufacturing Engineering, Ecole de Technologie Superieure, Montreal, QC H3C 1K3, Canada (e-mail: robert.sabourin@etsmtl.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TEVC.2011.2170075

The EEB can be viewed conceptually following one of two paradigms. In one of these, exploration and exploitation can be regarded as opposing forces, which means that increasing one reduces the other, while in the other, they can be regarded as orthogonal forces [7]. This second paradigm offers the possibility of increasing both exploration and exploitation simultaneously.

In fact, the opposing forces paradigm is a special case of the orthogonal forces paradigm, in that, for a unimodal landscape, reducing exploration increases exploitation proportionally. However, the situation is different for a multimodal landscape, where exploration and exploitation can be intensified simultaneously. For instance, to locate each peak of a landscape having uniformly distributed peaks of the same amplitude, and a population size equal to the number of peaks, exploration and exploitation have to be maximized concurrently. With very rugged landscapes, exploration could be in a maximal state, but with poor exploitation. In contrast, if the population converges over a very rugged, restricted region, exploration and exploitation would be in a minimal state at the same time. Finally, a converged population over a small plateau would be characterized by poor exploration and extensive exploitation. From this we can conclude that the orthogonal EEB concept is more suitable than the opposing forces concept to represent any landscape type. It also demonstrates that it could be useful to consider both genotypic and phenotypic diversity to characterize a given landscape effectively.

Exploration is adequately monitored by genotypic diversity evaluation (diversity of solutions), whereas exploitation is better described by phenotypic diversity (diversity of solution responses). These two diversity measurements also refer to the quantity and quality of the population diversity, respectively, [10]. In fact, genotypic diversity is built from the spread of the individuals over the search space, and phenotypic diversity is defined by the fitness distribution of the population. This means that phenotypic diversity is synonymous with fitness diversity. With normalized evaluation, unitary genotypic and phenotypic diversity values relate to maximum exploration and exploitation, respectively. It is important to note that, unlike genotypic diversity, phenotypic diversity defines maximum exploitation when it is in a state of convergence. However, evaluating genotypic and phenotypic diversity properly is difficult. Multiple diversity measures have been proposed in the literature without a clearly elucidated study of their differences being performed. This paper focuses on a comprehensive study of genotypic diversity measures (GDMs).

Depending on the problem and the representation used, the number of diversity measures could be infinite [11]. It is important, therefore, to clearly define the scope of applicability of this paper. First, the diversity measures considered here are those that can be related to the search space location of the individuals in the population. The diversity measures related to the tree structure representation, used in the genetic programming field, are not covered [11]–[13]. Also, the analysis is restricted to real-coded representations.

Diversity assessment is critical to monitoring and/or controlling the EEB. The aim of this paper is to provide an initial stepping stone toward EEB management, and it does so by studying the similarities and differences among GDMs. Of course, a good GDM should be capable of representing the real genotypic diversity of a population. However, it should also:

1) demonstrate repeatability with respect to a similarly scattered population;
2) be robust with respect to the simulation parameters, like population size and landscape dimensionality;
3) adequately describe the presence of outliers inside the population.

To perform this comparison task, our complete analysis makes use of a new, simple benchmark that allows clear definition of the tested indicator behavior. Furthermore, we restricted this study to the available GDM formulations provided in the literature. This paper is organized as follows. The next section, Section II, describes the various GDMs studied, Section III presents the published comparative studies, Section IV discusses the proposed benchmark, Section V presents the results, Section VI validates the use of the proposed benchmark, and Section VII provides our discussion, and Section VIII draws our conclusion.

II. Genotypic Diversity Measure

Even though no consensus has emerged on the definition of diversity [14], [15], the concept can be defined as the degree of heterogeneity or homogeneity between individuals in a studied population [16].

A. General Concept

Genotypic diversity can be evaluated using one of two approaches. The first is based on a measurement of the distance between individuals. This distance may be evaluated from the mean spatial position of the population [17]–[19], from the position of the fittest individual [10], or the position of each of the individuals, which in this case would range from the pairwise measure [20], [21] to the maximum distance between two individuals [20]. The Euclidian distance is more common for distance estimation with real-coded genes, since the landscape is defined in a Euclidian space R^n, where n represents the landscape dimensionality.

The second approach scans gene frequency (GF). This concept is generalized from binary representations, where the probability of the alleles at each locus is calculated within the complete population [22]. In a real-coded framework, all genes are continuous. Consequently, gene scanning requires gene partitioning. The predefined intervals are then considered as possible alleles. Ichikawa and Ishii [23] applied this procedure to integer representations, and the technique was later generalized to any symbolic alphabet by Wineberg and Oppacher [22]. Nevertheless, the number of intervals involved in the discretization constitutes a severe limitation; they directly influence diversity estimation, which could make it difficult to achieve meaningful usage for a small population size or high dimensionality. Moreover, the GF combination among all landscape variables must be defined. For example, Gouvêa, Jr., and Araújo [16] proposed using a representative gene to characterize the population diversity. In other words, the diversity measure is reduced to the consideration of only one gene or landscape variable characterizing the individuals. As they mentioned, the selected gene has to be a significant one. Therefore, to avoid a misleading diversity estimation, an average evaluation obtained from the diversity measure of each gene may be preferred [22]. Collins and Jefferson [24] also used the average GF to determine the population diversity. However, this paper was limited to binary representations.

Finally, Table I provides a summary of the symbols used throughout this paper.

TABLE I
Symbols Defined for This Study

Symbol	Definition
i, j	Individual number $\in \{1, 2, \ldots, N\}$
k	Gene locus $\in \{1, 2, \ldots, n\}$
m	Interval number
M	Total number of intervals
n	Landscape dimensionality
N	Population size
$x_{i,k}$	Gene k of individual i
$\overline{x}_k$	Average gene value of the population
$p_{m,k}$	Fraction of N that belongs to interval m on gene k
LD	Length of the landscape diagonal
NMDF	Normalization with maximum diversity so far

B. Normalization

Normalization of the various GDMs is preferable for comparison purposes, as the descriptors can then be evaluated on the same basis.

When defined, the maximum value can be used as a normalization factor. In the case of distance measurement, the landscape diagonal (LD), i.e., the maximum distance between opposite corners of the landscape, can also be used for normalization. Otherwise, the following simple normalization approach is proposed: the maximum value obtained so far during the evolution process of a given problem could serve as a normalization factor. The first iteration then becomes the reference, until a more diverse population is found. Since the initial EA population is generally created from a random uniform distribution, it is supposed to be the most diverse population. However, as information continues to arrive during the process, the indicator is updated if required. This normalization method is referred to here as normalized with maximum diversity so far (NMDF). NMDF is similar to the normalization used in [25].

C. Genotypic Diversity Measures

The GDMs based on distance measurements (D) and GF considered in this paper are listed in the following table. They are presented in their normalized form. The asterisk in the reference column indicates that the corresponding measure uses a normalization method not defined in its original form.

The first GDM in this table corresponds to the diameter of the population (D_{DP}^{N}), which is a pairwise measure considering only the distance between the two most widely separated individuals in the population.

The second GDM represents the radius of the population (D_{RP}^{N}), and determines the distance between the individual farthest away and the mean position of the population. It is possible to generalize D_{RP} to account for only a certain fraction (f) of the individuals around the mean position. This leads to the third GDM in Table II, $D_{\mathrm{RP}}^{N}(f)$, where N is sorted in ascending order with respect to the mean position. Therefore, extreme individuals can be set aside.

The fourth GDM, proposed by Ursem [17], is the distance-to-average-point measure (D_{DTAP}^{N}), and it represents the mean radius of the population. In this paper, a modified normalization version of this GDM is also considered, i.e., D_{DTAP}^{N2}, which is presented as the fifth GDM in Table II. With this form, the LD normalization factor is replaced by NMDF. This expression can also be considered as the normalization alternative to the D_{DTAP} measure proposed by Abbass and Deb [18]. No justification was provided in [17] or [18] to justify the usefulness of D_{DTAP}, except its intuitive formulation meaning.

The sixth GDM, proposed by Olorunda and Engelbrecht [20], defined a measure considering the average of the average distance around the individuals of the population (D_{ALL}). In this formulation, the center is represented by individuals i. D_{ALL} was defined to give an indication of the dispersion of the individuals with respect to each other. In fact, with normalization, D_{ALL}^{N} becomes identical to D_{PW}^{N} (ninth GDM in Table II), but its formulation is more computationally intensive than the latter. Therefore, D_{ALL}^{N} is not considered further in this paper.

In order to reduce the calculation time associated with pairwise measurements, which is $O(n{\cdot}N^2)$, to a linear relation $O(n.N)$, Wineberg and Oppacher [22] propose a measure, named "true diversity" (D_{TD}), which represents the average standard deviation of each gene in the population. The "true diversity" normalized with NMDF is given by the expression D_{TD}^{N}, which corresponds to the seventh GDM in Table II.

Following the computational improvement idea, Morrison and De Jong [19] proposed the moment of inertia measure (D_{MI}), which leads to D_{MI}^{N} (the eight GDM in Table II) when normalized with NMDF. As with the physical concept, the remote points (outliers) should have greater influence on this measurement. The development of this GDM was justified by the goal of having a unique diversity measure, whatever binary or real-coded representation is used.

The mean of the pairwise distance among individuals in the population (D_{PW}) is an intuitive GDM. This corresponds to the ninth entry in Table II. Even though this measure may be more time-consuming, it could be quite effective for describing population diversity. Moreover, it is worth making the point that it is better to use a slower, but effective measure than an indicator that is fast, but prone to be inaccurate. For this paper, the NMDF normalization factor is used for D_{PW}^{N}.

Herrera *et al.* [25] proposed two GDMs as input to their fuzzy logic system: the variance average chromosomes (D_{VAC}) and the average variance alleles (D_{AVA}), both of which are defined for real-coded representations. The latter is not presented in Table II, since it is equivalent to $D_{\mathrm{MI}}/(n.N)$ and the term $n.N$ remains constant in the evolution process considered. D_{VAC}^{N} is normalized by NMDF and it is the tenth GDM in Table II. No justification was provided for characterizing the usefulness of these GDMs, except the fact that they are indifferent to the mutual exchange of individuals in a population and they take a low value when the population moves toward a genotypic convergence state.

The last GDM, based on the distance measure described in this paper, is represented by the 11th entry in Table II. It is D_{ED}^{N}, proposed by Herrera and Lozano [10] without any justification. This diversity measure requires the preidentification of the fittest individual in the population, since it uses this individual as a reference to measure the distance from the other individuals. Other variants of this GDM are possible. Nevertheless, as will be explained in the next section, a major flaw can be seen in this kind of measurement.

In terms of GF measures, the Shannon entropy (GF_S^{N}) [26] is the best-known method employed as a GDM. It is intuitive, since entropy defines the level of disorder in a population [29]. The normalization of GF_S requires its maximum value. This is obtained when the gene frequencies are similar, which means that $p_{m,k} = 1/M$. However, it is important to note that this is true only if $M \leq N$. Otherwise, the maximum value is obtained when $p_{m,k} = 1/N$. In these cases, the most uniformly spread out distribution is $1/N$. Thus, replacing $p_{m,k}$ in the GF equation by one of these two upper bounds leads to that maximum value. This observation is valid for all GF measures, and the expressions 12–15 in Table II present the normalized version of the GF, where $u = \min\{M, N\}$. The Havrda and Charvát entropy ($\mathrm{GF}_{\mathrm{HC}}$) [27] is another important GF measure. This descriptor has been well analyzed by Nayak [30]. The following conditions are required for this family: $\alpha > 0$ and $\alpha \neq 1$. It is interesting to note that, when $\alpha = 2$, $\mathrm{GF}_{\mathrm{HC}}$ reduces to the Gini–Simpson index [31], [32]. Good [33] offers an excellent historical perspective on this index, and Rényi [28] has proposed another entropy family (GF_R). It is worth noting that, as $\alpha \to 1$, $\mathrm{GF}_{\mathrm{HC}}$ and GF_R tend toward GF_S [30]. Finally, Wineberg and Oppacher [22] published a GF that was developed for the same reasons as D_{TD}. This GF is designed to work with a finite-sized alphabet, which means that it can be used in the present context, where the total number of intervals on a gene (M) depicts the alphabet. This GDM is designated $\mathrm{GF}_{\mathrm{PW}}$. These authors have shown that $\mathrm{GF}_{\mathrm{PW}}$ is correlated to GF_S [22]. In fact, by means of a Taylor expansion of the second term of GF_S ($\log(p_{m,k})$), they demonstrated that the last term of $\mathrm{GF}_{\mathrm{PW}}$ ($1 - p_{m,k}$) constitutes the first term of this series, and dominates all the other terms. The normalization process for $\mathrm{GF}_{\mathrm{PW}}$ is identical to that of the other GF measures. However, Wineberg and Oppacher added a correction term ($r = N \bmod M$) to account for the cases where

TABLE II
GDMs USED FOR THE COMPARATIVE STUDY

No.	GDM Formulation	Ref.
1.	$D_{DP}^{N}=\frac{1}{LD}\max_{(i\neq j)\in(1,2,\ldots,N)}\left(\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-x_{j,k}\right)^2}\right)$	
2.	$D_{RP}^{N}=\frac{1}{LD}\max_{i\in(1,2,\ldots,N)}\left(\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-\overline{x}_k\right)^2}\right)$	[20]
3.	$D_{RP}^{N}(f)=\frac{1}{LD}\max_{i\in(1,2,\ldots,f\cdot N)}\left(\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-\overline{x}_k\right)^2}\right)$	
4.	$D_{DTAP}^{N}=\frac{1}{LD\cdot N}\sum_{i=1}^{N}\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-\overline{x}_k\right)^2}$	[17]
5.	$D_{DTAP}^{N2}=\frac{\frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-\overline{x}_k\right)^2}}{\mathrm{NMDF}}$	[18]*
6.	$D_{ALL}^{N}=\frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-x_{j,k}\right)^2}\right)}{\mathrm{NMDF}}$	[20]*
7.	$D_{TD}^{N}=\frac{\frac{1}{n}\sqrt{\sum_{k=1}^{n}\left(\overline{x_k^2}-\left(\overline{x}_k\right)^2\right)}}{\mathrm{NMDF}}$ where $\overline{x_k^2}=\frac{1}{N}\sum_{i=1}^{N}x_{i,k}^2$	[22]*
8.	$D_{MI}^{N}=\frac{\sum_{k=1}^{n}\sum_{i=1}^{N}\left(x_{i,k}-\overline{x}_k\right)^2}{\mathrm{NMDF}}$	[19]*
9.	$D_{PW}^{N}=\frac{\frac{2}{N(N-1)}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-x_{j,k}\right)^2}}{\mathrm{NMDF}}$	[21]*
10.	$D_{VAC}^{N}=\frac{\frac{1}{N}\sum_{i=1}^{N}(\overline{x}_i-\overline{x})^2}{\mathrm{NMDF}}$ where $\overline{x}_i=\frac{1}{n}\sum_{k=1}^{n}x_{i,k}$ $\overline{x}=\frac{1}{n\cdot N}\sum_{i=1}^{N}\sum_{k=1}^{n}x_{i,k}$	[25]*
11.	$D_{ED}^{N}=\frac{\overline{d}-d_{\min}}{d_{\max}-d_{\min}}$ where $\overline{d}=\frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-x_{best,k}\right)^2}$ $d_{\max}=\max_{i\in(1,2,\ldots,N)}\left(\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-x_{best,k}\right)^2}\right)$ $d_{\min}=\min_{i\in(1,2,\ldots N),i\neq best}\left(\sqrt{\sum_{k=1}^{n}\left(x_{i,k}-x_{best,k}\right)^2}\right)$	[10]
12.	$\mathrm{GF}_{S}^{N}=-\frac{1}{n\cdot\log(u)}\sum_{k=1}^{n}\sum_{m=1}^{M}p_{m,k}\log\left(p_{m,k}\right)$	[26]*
13.	$\mathrm{GF}_{HC}^{N}(\alpha)=\frac{1}{n\left(1-u^{1-\alpha}\right)}\sum_{k=1}^{n}\left(1-\sum_{m=1}^{M}p_{m,k}^{\alpha}\right)$	[27]*
14.	$\mathrm{GF}_{R}^{N}(\alpha)=\frac{1}{n\cdot\log(u)}\sum_{k=1}^{n}\frac{\log\left(\sum_{m=1}^{M}p_{m,k}^{\alpha}\right)}{1-\alpha}$	[28]*
15.	$\mathrm{GF}_{PW}^{N}=\beta\cdot\sum_{k=1}^{n}\sum_{m=1}^{M}p_{m,k}\left(1-p_{m,k}\right)$ where $\beta=\begin{cases}\frac{M}{n\left((M-1)-\frac{r(M-r)}{N^2}\right)}, & \text{if } M<N\\ \frac{N}{n(N-1)}, & \text{otherwise}\end{cases}$	[22]

M is not a common divisor of N, and is therefore applied when $M < N$.

To the authors' knowledge, all the published GDMs for real-coded representations available in the literature have been presented here.

D. Prior Observable Flaws on Certain GDMs

Before moving further in the comparative study of GDMs, it is useful to eliminate those that present observable flaws in their formulations. This applies to D_{DP}^{N}, D_{RP}^{N}, $D_{\mathrm{RP}}^{N}(f)$, D_{DTAP}^{N}, and D_{ED}^{N}.

Actually, D_{DP}^{N} is not an appropriate GDM for two reasons. First, the diversity estimate of the population is led by only the two most distant individuals, and this is the case whatever the scattering of the remaining individuals. Furthermore, the maximum value obtained by D_{DP}^{N} is when these two individuals are located on the extreme corners of the landscape, which is not, in any case, a sign that the population is fully diverse.

The formulation of D_{RP}^{N} shows similar flaws, as the diversity is based on the location of the individual farthest from the center of mass of the population. Therefore, a fully diverse population will be described by this indicator with a value near

0.5. The true diversity state of the population is misleading, as the value goes toward 1. In fact, this indicates that the population converges near a landscape corner, whereas an outlier exists near the opposite corner of the landscape.

$D^N_{\mathrm{RP}}(f)$ was introduced to reduce the potential impact of outliers on the preceding GDM. However, the factor f has to be properly defined, and, even though it increases robustness, it inevitably generates information leakage. Moreover, this indicator faces the same issue as D^N_{RP} with respect to coverage of the diversity range.

D^N_{DTAP} copes with the same issue as the three preceding GDMs, in terms of the diversity range coverage. This aspect is related to the LD used as the normalization factor. Furthermore, it is worth noting that the LD makes the diversity evaluation very sensitive to the landscape dimensionality, as the distance between the extreme corners of the landscape increases with the number of dimensions.

In contrast, D^N_{ED} is unable to describe the population diversity, since its normalization term decreases with its numerator, when the population moves toward convergence. Therefore, over a linearly convergent process, this indicator will remain constant, even if the population shows a linear reduction in its diversity.

In the next section, we present and discuss the comparative studies available in the literature.

III. Review of Comparative Studies

Gouvêa, Jr., and Araújo [16] presented five GDMs that can be used with real-coded representations: D^N_{DTAP}, GF_S, $\mathrm{GF}_{\mathrm{PW}}$, and $\mathrm{GF}_{\mathrm{HC}}(2.0)$ (Gini–Simpson index). The GDM not listed is a GF measure developed in [34] for binary representations, and adaptable to real-coded representations. In fact, it uses D^N_{DTAP} for the intervals in a formulation similar to the Shannon entropy. Preliminary tests conducted in this study show that this descriptor is not adequate for the diversity evaluation of real-coded representations, and so it is not considered here. Gouvêa, Jr., and Araújo promoted the use of $\mathrm{GF}_{\mathrm{HC}}(2.0)$ with $M = 10$, and consider only one representative gene. However, they did not provide any clear justification for doing so. They developed their EA adaptive control with this measure, and compared the resulting performance with Ursem [17] and a standard genetic algorithm (SGA) on three dynamic environment problems. They concluded that their method outperformed the other two.

Olorunda and Engelbrecht [20] compared six GDMs (D_{DP},D_{RP}, D_{DTAP}, D^{N*}_{DTAP}, D_{ALL}, and swarm coherence) on four synthetic test functions treated with a particle swarm optimization (PSO) approach. D^{N*}_{DTAP} is a normalized version of D_{DTAP} which is different from D^N_{DTAP} and D^{N2}_{DTAP}. It considers the population diameter instead of the diagonal of the search space. Olorunda and Engelbrecht referred to this measure as the one used by Riget and Vesterstrom [35]. However, Riget and Vesterstrom clearly state that the normalization of their measure was achieved with the LD. In contrast, the swarm coherence measure requires the velocity of the swarm, which makes it PSO-specific. Olorunda and Engelbrecht also showed that it can produce ambiguous results. Consequently, swarm coherence was not included in Section II. Finally, the authors only include the D^{N*}_{DTAP} results in their paper, which makes the analysis close to an intuitive comparison. Nevertheless, they rank the measures according to their sensitivity to outliers. From the most sensitive to the most robust, the classification is as follows: D^{N*}_{DTAP}, D_{DP}, D_{RP}, D_{DTAP}, and D_{ALL}. They recommend D_{DTAP} based on this ranking and on the computation time.

As mentioned above, Wineberg and Oppacher [22] showed that $\mathrm{GF}_{\mathrm{PW}}$ is actually an approximation of GF_S, and, since D_{TD} corresponds to the average standard deviation of each gene, they all seem to be the same measure. These authors claim that, as a result, experiments were not required to choose the best GDM. However, in this paper, we will demonstrate that this belief appears to be a mistaken one, at least for real-coded representations.

IV. Benchmark

The EA domain offers recognized benchmarks, such as CEC'05 [36] and BBOB'09 [37], for single objective environment test cases. Nevertheless, for GDMs comparison purpose their usefulness can be problematic owing to two major reasons. First, since the use of a particular EA dictated the EEB over the optimization process, the diversity level of the population is biased by the underlying choice of EA parameters. Therefore, no information about the real diversity state of a population is available, except the one from the GDM comparison. This leads to an ill-defined problem, as we get different estimations from the GDMs without being able to say which one best reflects the true diversity value. The second aspect is related to the benchmark definition. Indeed, genotypic diversity is only concerned with the location of the individuals over a landscape, and not with its associated fitness function. Therefore, the sole requirement is to provide an environment for the GDMs where the population moves from a fully scattered state to a fully converged one. The number of optima over the landscape should also have an impact on the GDMs. A well-defined benchmark has to be able to simulate the modality influence.

In contrast, it could be interesting to link the GDM analysis to EA convergence tools as the takeover time concept, which is the time required by the best individual to populate the entire population [38]. Within this framework, we will obtain a reference boundary between a fully scattered population (first generation) and a fully converged population (takeover time generation) for any landscape. However, as will be clearly seen in Section V, the most important zone where the behavior of the GDMs can be discriminated is between those boundaries where any convergence tools remain silent about EA behavior, and this is because of the stochastic nature of EAs.

We believe that an appropriate benchmark problem should present a population diversity that is known quantitatively, or at least qualitatively, throughout the evolution process. This section presents such benchmark problems for both unimodal and multimodal landscapes with two and four optima. These modality choices are made with the aim of visualizing the effect of GDM behavior on different landscape

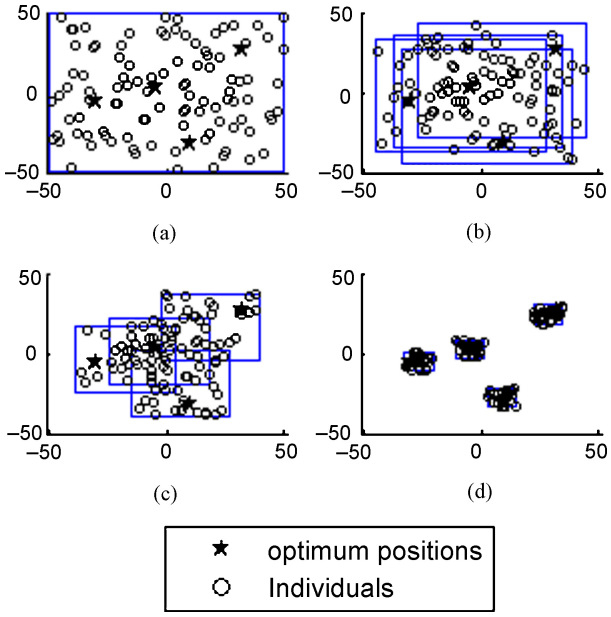

Fig. 1. Population of uniformly distributed individuals (N = 100) on four optima positions in a 2-D landscape at a given iteration. (a) Iteration 1. (b) Iteration 15. (c) Iteration 30. (d) Iteration 45.

structures. The main steps of this benchmark are presented in Algorithm 1.

For quantitative comparison purposes, the simplest benchmark decreases the diversity linearly from a fully scattered population to a fully converged one. This is achieved by creating a uniformly distributed random population over the search space (line 29 of Algorithm 1) and reducing the available hyperspace toward a given location at a constant rate (line 9 of Algorithm 1). This simulates convergence toward an optimum position. It is important to mention that this involves no evolution operators, since a new population is generated within the converging population bounds of the genotypic space for each iteration. The reduction rate chosen per iteration is 2% of the distance between the landscape frontiers and the optimum position. The process then requires 51 iterations to converge, and ensures a clearly observable GDM behavior. To avoid the introduction of any bias, the optimum position is randomly generated on the landscape at each repetition (line 4 of Algorithm 1). For all experiments presented in this paper, the genes ($x_{i,k}$) range from −50 to +50.

The multimodal landscape is similar to the unimodal one. However, since many optimum positions are fixed randomly at each repetition, the population is distributed uniformly or with a predefined ratio inside the respective bounds (line 28 of Algorithm 1). For example, Fig. 1 shows four optima on a 2-D landscape. The population is uniformly attributed to each optimum position. In this example, the square boxes represent the space boundaries for each optimum at a given iteration.

As mentioned by Olorunda and Engelbrecht [20], different GDMs may have different sensitivity to outlier individuals, which means that the proposed benchmark must be adapted to reflect this aspect. For outlier influence simulation purposes, the initial benchmark remains unchanged up to the tenth iteration. Then, a fraction of the population (N_{outlier})

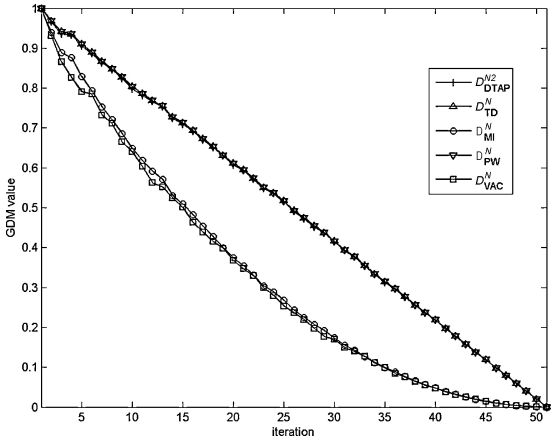

Fig. 2. Mean GDM values of D^{N2}_{DTAP}, D^{N}_{TD}, D^{N}_{MI}, D^{N}_{PW}, and D^{N}_{VAC} for the unimodal benchmark.

is generated randomly (line 36 of Algorithm 1) within the hyperspace comprising the first to tenth iterations (lines 13–20 of Algorithm 1). Consequently, the outliers are free to move in a restrictive zone, while remaining at a significant distance from the optimum. Outliers bring exploration capabilities to the population. Nevertheless, their influence on GDMs has to be related to their number. In other words, the outliers should increase the diversity evaluation, but never dominate the measure.

V. Results

A default configuration allowing analysis of all GDMs on a similar basis is employed: the population size (N) is 100, and the number of intervals (M) for GF measures is fixed to 100 for each gene. The benchmark is defined on 2-D landscapes. Finally, the results are averaged over 50 repetitions.

The first section presents the behavior of all the GDMs on a unimodal landscape. Thereafter, the GDMs are studied on multimodal landscapes.

A. Unimodal Landscape Experiment

Figs. 2–4 show the GDMs response on the unimodal landscape. Fig. 2 indicates that D^{N2}_{DTAP}, D^{N}_{TD}, and D^{N}_{PW}, with overlaid curves, precisely describe the linear relation intended by this benchmark. D^{N}_{MI}, even though showing a quadratic shape, still offers good discrimination of the diversity state. D^{N}_{VAC} acts similarly to D^{N}_{MI}. The behavior of these measures is expected to be quadratic, since they are based on genotypic variance. A linear trend could be achieved by taking their square root. However, this is not considered here, as this paper is limited to GDMs that have already been suggested.

Figs. 3 and 4 present the GF diversity measures. Given that all these measures share common properties, they are combined in the following discussion. First, the parameter a has a greater impact on $\text{GF}^{N}_{\text{HC}}$ than on GF^{N}_{R}, making this latter GDM more reliable. In fact, α has an inverse influence on the two measure families. Also, the Gini–Simpson index ($\text{GF}^{N}_{\text{HC}}(2.0)$) appears to be similar to $\text{GF}^{N}_{\text{PW}}$. These measures were found to have a major drawback, however, which is

Algorithm 1: benchmark(*repetition, modality, iteration, pop_size, Noutlier, it_outlier*)

Input: number of repetition, modality of the landscape, maximum number of iterations for the entire process, population size, number of simulated outliers, and iteration where outliers appear.

Output: optimum positions and populations genotype for the complete process of all repetitions.

```
for r = 1,..., repetition do
    //*creation of the optimum positions*//
    for peak = 1,..., modality do
        Generate random optimum position inside landscape frontiers
    end for
    for it = 1,..., iteration do
        //*definition of iteration bounds for each optimum*//
        for peak = 1,..., modality do
            Evaluate the lower bound and the upper bound for each dimension with respect to it and peak.
        end for
        //*definition of the outlier parameters*//
        if Noutlier > 0 and it ≥ it_outlier then
            N = pop_size – Noutlier   //*evaluate the non outlier population size*//
            if it == it_outlier then
                //*initialize outlier bounds*//
                Set the outside lower bound and the outside upper bound of the outliers with the landscape frontiers
                for peak = 1,..., modality do
                    Set the inside lower bound and the inside upper bound of the outliers by the lower and upper bounds of peak for the current it
                end for
            end if
        else
            N = pop_size
        end if
        //*creation of the individuals*//
        for ind = 1,..., N do
            for peak = 1,..., modality do
                //*distribution of the individuals uniformly inside each peak boundaries*//
                if ind > ⌊(peak - 1) * N / modality⌋ and ind ≤ ⌊peak * N / modality⌋ then
                    Generate random individual inside the boundaries of peak for the current it
                end if
            end for
        end for
        //*creation of the outliers*//
        if Noutlier > 0 and it ≥ it_outlier then
            for ind = 1,..., Noutlier do
                Generate random outlier located between the inside and outside outlier bounds of all peak.
            end for
        end if
    end for
end for
Return optimum positions for each repetition, and the population genotype of each iteration for each repetition
```

that they remain very close to their maximum values for a long period during the process. In other words, they provide the worst discriminating diversity evaluations. Their formulations place the emphasis on crowded species or intervals [39]. Therefore, diversity changes begin to be measured only when all the individuals pile up in a small number of intervals, which happens close to when the convergence state is reached.

Complementary information about this GF drawback is presented in Fig. 5. In this figure, the black and empty circles represent two different populations. Each contains ten individuals, and a 10 by 10 grid is used for interval control. The black circle population is obviously more scattered than the empty circle one. However, the diversity evaluations for all the GF measurements indicate that these two populations are equally distributed. In contrast, distance-based measurements demonstrate the difference between them. For instance, D_{PW} indicates that population 2 (empty circles) is about 68% less diversified than population 1 (black circles). The nondiscrimination phenomenon observed for all GF measures can be explained by the fact that all GF measures are based on the proportion of individuals resident in the various intervals for each gene, and there is no consideration at all of the location of these intervals over the gene axis. This is a major weakness,

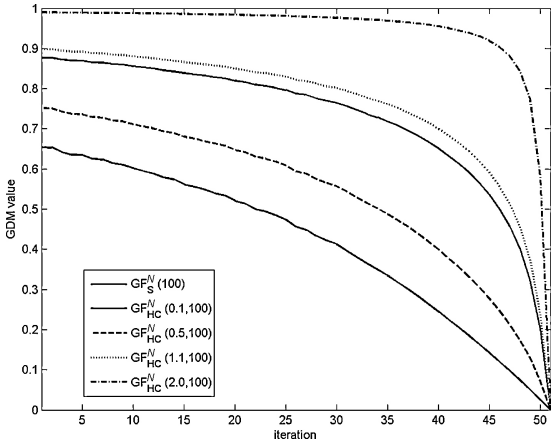

Fig. 3. Mean GDM values of $\mathrm{GF}_S^N(M)$ and $\mathrm{GF}_{\mathrm{HC}}^N(\alpha, M)$ for the unimodal benchmark $\alpha = \{0.1, 0.5, 1.1, 2.0\}$.

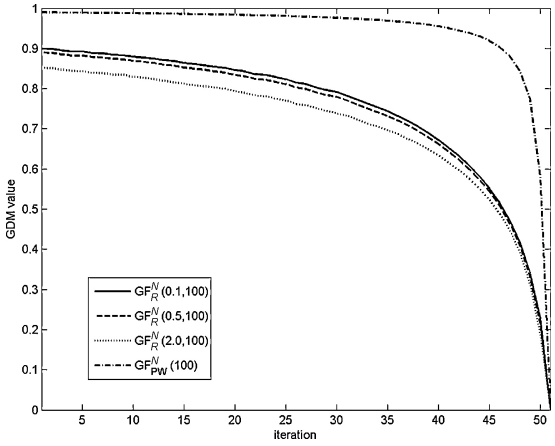

Fig. 4. Mean GDM values of $\mathrm{GF}_R^N(\alpha, M)$, and $\mathrm{GF}_{\mathrm{PW}}^N(M)$ for the unimodal benchmark $\alpha = \{0.1, 0.5, 2.0\}$.

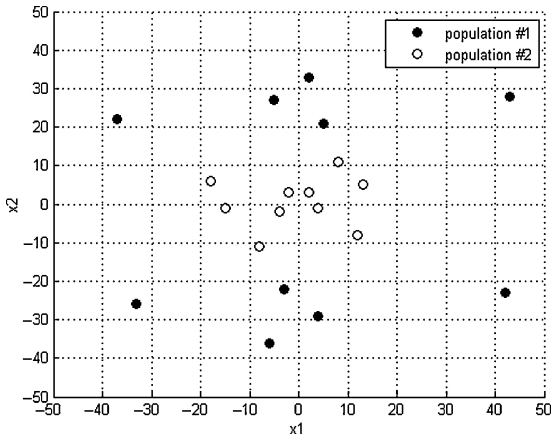

Fig. 5. Simulation with two different populations (black and empty circles).

which, as illustrated, could rapidly result in a misleading diversity analysis.

B. Multimodal Landscape Experiment

This section presents the response of selected GDMs to the multimodal benchmarks. Fig. 6 shows the evaluation of the five GDMs normalized with NMDF on the multimodal

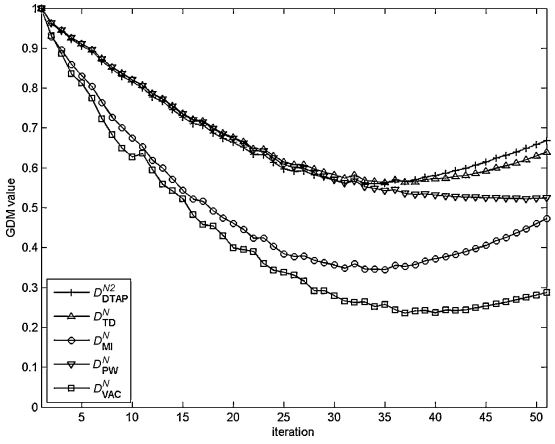

Fig. 6. Mean GDM values of D_{DTAP}^{N2}, D_{TD}^{N}, D_{MI}^{N}, D_{PW}^{N}, and D_{VAC}^{N} on a two optima benchmark.

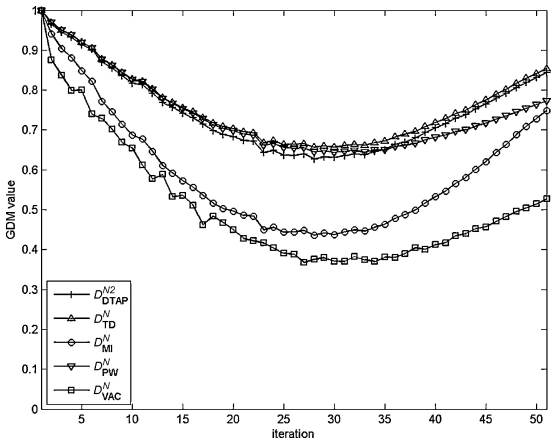

Fig. 7. Mean GDM values of D_{DTAP}^{N2}, D_{TD}^{N}, D_{MI}^{N}, D_{PW}^{N}, and D_{VAC}^{N} on a four optima benchmark.

benchmark with two optima. Fig. 7 displays the same GDMs on a four optima landscape.

A general quadratic shape with a minimum somewhere in the process appears with these GDMs. The trend is accentuated as the modality increases. This phenomenon is explained as follows: at the beginning of the process, all the attracting pool boundaries share the entire landscape. As the process goes on, every bounded space shrinks around its respective optimum. As long as the boundaries overlap, diversity decreases, but then starts to increase with the separation of the bounded hyperspaces (Fig. 1). Moreover, the rises in measured diversity depend on the ratio of the number of individuals converging to each optimum and the distance between the optima. Fig. 8 illustrates the ratio effect with two different GDMs on the four optima landscape. The comparison is performed for a uniform ratio (25% of N attached to each optimum) and a monopolizing optimum (70% of N to the dominant point, and the remaining 30% equally distributed among the other three optima). In light of Fig. 8, the influence of the ratio becomes obvious; the nonuniform case behaves as a unimodal landscape with the less attractive points acting as outlier clusters.

Fig. 9 presents the characteristic GF pattern. In reality, the figure is restricted to the GF_S^N response for the unimodal and

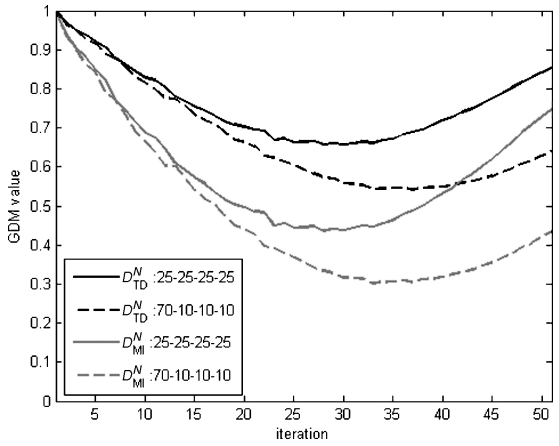

Fig. 8. Effect of the ratio of individuals associated with each optimum on a four optima benchmark.

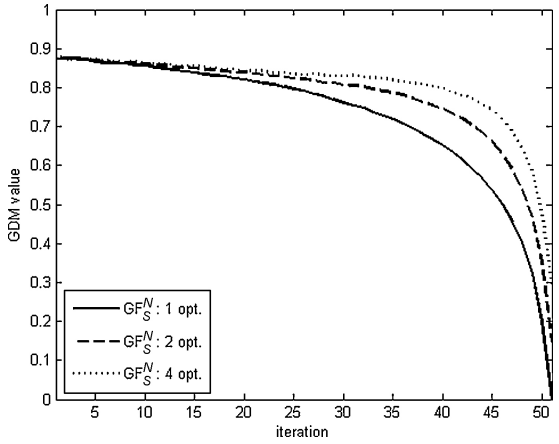

Fig. 9. Mean GDM values of GF^N_S on the unimodal, two optima, and four optima benchmarks.

multimodal landscapes with two or four optima. The curves clearly indicate that, as modality increases, the discriminating GF power deteriorates. This is because the convergence is less concentrated in a few intervals.

Since only two (four) optima locations are represented by the population at the end of the multimodal process, GF-based measurement seems to provide a better estimate of the real diversity than distance-based measurement. The difficulty of the latter is probably due to the nonspecial treatment afforded to duplicated individuals. Nevertheless, this phenomenon is not studied further in this paper, as no better discrimination capability could be found among the indicators compared. At the same time, it is difficult for GF measures to adequately describe the diversity of the population throughout the majority of the process, as no consideration is given to the location intervals. Consequently, this experiment has demonstrated that none of the GDMs is capable of reflecting the true diversity over a multisite convergence process.

C. Stability Analysis

To further discriminate the power of the various GDMs, a stability analysis is produced in this section, followed by a sensitivity analysis and an outlier study.

TABLE III
STABILITY ANALYSIS − UNIMODAL LANDSCAPE, WITH $n = 2$

GDM	Population Size (N)			
	50	100	300	500
D^{N2}_{DTAP}	0.110	0.086	0.049	0.040
D^N_{TD}	0.093	0.074	0.042	0.034
D^N_{MI}	0.119	0.096	0.055	0.045
D^N_{PW}	0.096	0.076	0.043	0.035
D^N_{VAC}	0.200	0.162	0.094	0.076

GDMs should be stable in their measurement of the diversity value over similarly scattered populations. This property could be analyzed by looking at the dispersion of the 50 repetitions for a given iteration. Because the samples do not follow a normal distribution, the standard deviation is not a suitable indicator. Indeed, the normality assumption associated with the samples was tested and invalidated in this paper using the Kolmogorov–Smirnov test (0.05 significance level). Stability is therefore evaluated by considering the dispersion range among 96% of the repetition data, which provides the same stability basis for all GDMs. That means that the difference between the second highest diversity value and the second lowest diversity value of the repetition at each iteration is computed. To present this analysis in a comprehensible manner, the dispersion values are averaged over the whole process.

Table III presents the stability computed for the five GDMs normalized with NMDF. Only the unimodal landscape is processed, since random positioning among optima on multimodal landscapes makes the stability analysis unreliable. However, the analysis is presented over four commonly used population sizes in EAs: $N \in \{50, 100, 300, 500\}$. This allows the sampling error to be visualized, since stability improves as the population size increases. By considering the largest population size, the sampling error is minimized. Thus, for this configuration ($N = 500$), four GDMs (D^{N2}_{DTAP}, D^N_{TD}, D^N_{MI}, and D^N_{PW}) have an average dispersion value under 0.05, which can be qualified as stable. The remaining GDM (D^N_{VAC}) could be considered less stable. The classification of the five GDMs, presented, in increasing order of stability, is D^N_{VAC}, D^N_{MI}, D^{N2}_{DTAP}, D^N_{PW}, and D^N_{TD}.

The high degree of stability of most GDMs justifies our presentation of the above experiments, which shows that the mean curves of the GDM behavior are representative. It is also interesting to note that, even if most GDMs are stable and some have demonstrated a similar trend in terms of their mean curves ($\{D^{N2}_{DTAP}, D^N_{TD}, D^N_{PW}\}$, $\{D^N_{MI}, D^N_{VAC}\}$, $\{GF^N_{HC}(2.0), GF^N_{PW}\}$), preliminary statistical tests based on the Wilcoxon-signed rank indicated that none of them is built from the same diversity distribution over the 50 repetitions.

D. Sensitivity Analysis

The robustness of the various GDMs with respect to the underlying parameters of the analyses (n and N) is also a concern. A one-at-a-time sensitivity analysis based on the Friedman statistical test allows a good definition of robustness. This is a nonparametric statistical test with the implicit assumption that the samples are related. It could be viewed as

TABLE IV
SENSITIVITY ANALYSIS − LANDSCAPE DIMENSIONALITY
{2, 10, 30}, WITH N = 100

GDM	% p-Values < α		
	One Optimum	Two Optima	Four Optima
D^{N2}_{DTAP}	21.57	39.22	27.45
D^{N}_{TD}	23.53	11.76	0
D^{N}_{MI}	19.61	11.76	0
D^{N}_{PW}	21.57	15.69	1.96
D^{N}_{VAC}	11.76	21.57	0

a nonparametric version of the repeated-measures analysis of variance. The null hypothesis is that the sample distributions are the same, while the alternative is that their medians are different, at least for one sample [40]. The application of this test is justified for two reasons. First, as previously mentioned (see Section V-C), the sampling considered does not follow a normal distribution. Second, the same GDM is compared for different repeated simulations (sensitivity with respect to n or N), and they are thus related. More details on this statistical test in the EA context are provided in an excellent description by Garcia *et al.* [41].

Before the results of the statistical test are presented, one question remains to be answered. It is related to the composition of the sampling used for comparison, since 50 repetitions were conducted during a 51-iteration process. Should the sampling be formed with the mean of the 50 repetitions at each iteration (51 points in each sample and one p-value), or should a test be conducted for each iteration with the 50 repetition values (50 points in each sample and 51 p-values)? The second option appears to be the more relevant one, as comparing the mean of the repetitions at each iteration would cloud the analysis, and the null hypothesis would be rejected if the median of the mean values were statistically different for the samples compared. For example, if two simulations were to monotonically decrease over the convergence process, the statistical test would be based only on the difference in their mean values calculated exactly at the central iteration of the whole process. In contrast, the use of the 50 repetition values raises another question: how should we treat the 51 p-values (each related to a different iteration) to accept or reject the null hypothesis? In this paper, we decided to rely on the percentage of p-values that fall below the predefined level of significance (α), which is fixed here at 0.05. Thus, the percentage value reflects the number of rejections of the null hypothesis over the 51-iteration process. A low percentage would indicate that most of the p-values were over the significance level, in which case the null hypothesis would not be rejected. A rejection then means that the GDM tested is sensitive to the scrutinized parameter. The default configuration described at the beginning of Section V serves as a reference for the fixed parameters. No potential cross-influences between factors are included in this analysis. First, the impact of landscape dimensionality (n) is studied, followed by the effect of population size (N). Algorithm 2 presents the general procedure for the statistical comparison.

Algorithm 2: statistical_comparison(*iteration*, *sample*, α)

Input: maximum number of iterations for the entire process, number of samples for the statistical comparison, and level of significance.
Ouput: percentage of rejection of the null hypothesis (H0) over the entire process.

```
number_reject_H0 = 0
for it = 1,..., iteration do
    for si=1,..., sample do
        Define the sample with the GDM repetition values
        for it and the benchmark parameter analyzed.
    end for
    p_value ←Evaluate H0 with the statistical test
    if p_value < α then
        number_reject_H0 = number_reject_H0 + 1
    end if
end for
Return (number_reject_H0 / iteration *100)
```

TABLE V
SENSITIVITY ANALYSIS − POPULATION SIZE
{50, 100, 300, 500}, WITH n = 2

GDM	% p-Values < α		
	One Optimum	Two Optima	Four Optima
D^{N2}_{DTAP}	50.98	37.25	11.76
D^{N}_{TD}	29.41	29.41	13.73
D^{N}_{MI}	27.45	29.41	13.73
D^{N}_{PW}	21.57	35.29	11.76
D^{N}_{VAC}	78.43	19.61	66.67

Table IV presents the statistical test results for three landscape dimensions: 2, 10, and 30. The robustness of a GDM with respect to the dimensionality of the landscape is synonymous with scalability, which is important in the EA context. In other words, it means that the GDM offers similar diversity estimation, whatever the dimensionality of the landscape. As this analysis indicates, all NMDF-normalized GDMs show a relatively high degree of robustness, since fewer than one-third of the iterations reject the similarity among the samplings. Furthermore, in general, the sensitivity decreases as the modality of the landscape increases. Based on this study, the classification, in terms of increasing order of robustness with respect to the dimensionality, is as follows: D^{N2}_{DTAP}, D^{N}_{PW}, D^{N}_{TD}, D^{N}_{VAC}, and D^{N}_{MI}.

Table V presents the sensitivity analysis results for the population size. The range was chosen to reflect common EA population sizes: $N \in \{50, 100, 300, 500\}$. No clear trend stands out from this analysis. However, we can see that D^{N}_{VAC} is very sensitive to population size, as is D^{N2}_{DTAP} for a low modality structure. Based on this study, the classification, in terms of increasing order of robustness with respect to the population size, is as follows: D^{N}_{VAC}, D^{N2}_{DTAP}, D^{N}_{TD}, D^{N}_{MI}, and D^{N}_{PW}.

E. Effect of Outliers

The following experiments illustrate the effect of outliers on the GDMs. Intuitively, the presence of outliers should increase diversity. Nevertheless, even though their number must be correctly reflected, outliers should never dominate the diversity evaluation, since, by definition, they correspond to a small portion of the population. The simulations were conducted

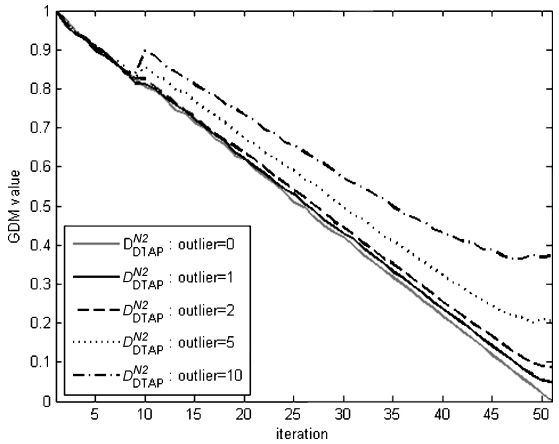

Fig. 10. Effect of outliers on D^{N2}_{DTAP} for the default configuration in a unimodal landscape. Outliers are introduced from the tenth iteration.

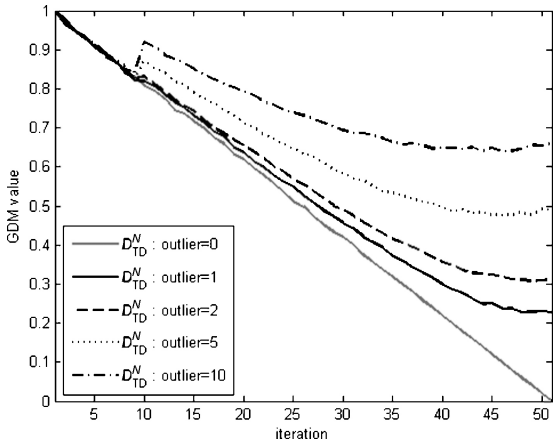

Fig. 11. Effect of outliers on D^{N}_{TD} for the default configuration in a unimodal landscape. Outliers are introduced from the tenth iteration.

with 1%, 2%, 5%, and 10% of outliers in the population. The following discussion uses the configuration presented at the beginning of Section V as a reference. In addition, other experiments were performed with different population sizes: $N \in \{300, 500\}$. The results indicate that the effect of outliers on diversity evaluation is the same for all these population sizes. Also, it could be shown that outliers have a similar influence in both unimodal and multimodal cases. Consequently, to abbreviate the description, the multimodal landscape results are not incorporated. Moreover, for the sake of conciseness, even though the discussion includes the five GDMs based on NMDF, Figs. 10–12 present only the three GDMs that show perfect identification of the diversity level over the unimodal benchmark (D^{N2}_{DTAP}, D^{N}_{TD}, and D^{N}_{PW}).

These three GDMs show adequate patterns. They present a translating trend with respect to their no outlier mean curve. This translation is proportional to the percentage of outliers. However, D^{N2}_{DTAP} (Fig. 10) and D^{N}_{TD} (Fig. 11) reveal the distinct influence of the number of outliers at the end of the process. This phenomenon is explained as follows: in most repetitions, the outliers are far from the population mean. As the process evolves, the difference between each individual and the center of the population becomes dominated by the outliers and culminates at the last iteration.

Table VI presents a comparison, with respect to D^{N}_{PW}, of the diversity value at the end of the process for each GDM based on NMDF. D^{N}_{PW} served as a reference because this GDM showed the most stable outlier evaluation (Fig. 12). The comparison is summarized with a robustness classification (in increasing order of robustness): D^{N}_{TD}, D^{N}_{MI}, D^{N}_{VAC}, D^{N2}_{DTAP}, and D^{N}_{PW}.

TABLE VI
END DIVERSITY RATIO, WITH RESPECT TO D^{N}_{PW}, IN THE PRESENCE OF OUTLIERS — UNIMODAL LANDSCAPE, WITH $n = 2$ AND $N = 100$

GDM	Outliers %				Mean Value
	1%	2%	5%	10%	
D^{N2}_{DTAP}	1.36	1.28	1.23	1.19	1.26
D^{N}_{TD}	6.39	4.56	2.93	2.12	4.00
D^{N}_{MI}	1.53	1.46	1.46	1.41	1.47
D^{N}_{PW}	1.00	1.00	1.00	1.00	1.00
D^{N}_{VAC}	1.22	1.31	1.49	1.25	1.32

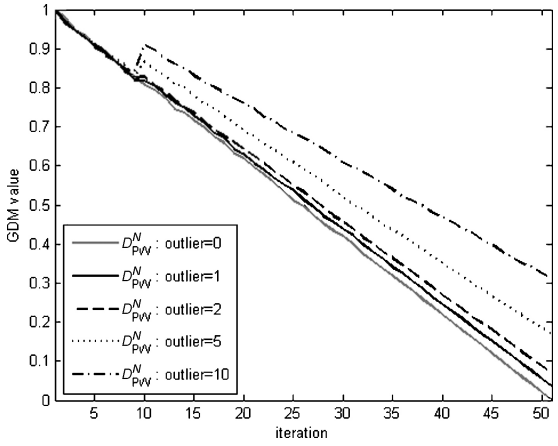

Fig. 12. Effect of outliers on D^{N}_{PW} for the default configuration in a unimodal landscape. Outliers are introduced from the tenth iteration.

VI. GDM COMPARISON OVER THE CEC'05 BENCHMARK

To strengthen the usefulness of this paper, all the GDMs presented were compared over the CEC'05 benchmark [36]. To accomplish this task, a state-of-the-art EA was used, which is G-CMA-ES[1] [42] and a particular EA specifically designed to promote diversity [43].

[1]CMA-ES version 3.51.beta was used to conduct this analysis. Available at http://www.lri.fr/∼hansen/cmaes.m

The former was considered the best algorithm of the 11 EAs over the CEC'05 benchmark [41], [44]. G-CMA-ES is an evolution strategy (ES) based on the covariance matrix adaptation (CMA) and a restart feature implemented to increase the exploration capability, as the population size is doubled at each restart. This feature is triggered by five independent convergence criteria related to CMA-ES parameters [42]. The parameters of G-CMA-ES used were the same as for CEC'05, except for the population size. Indeed, to make the observable behavior of the various GDMs clearer, and to have the same comparative basis as the paper presented in the previous

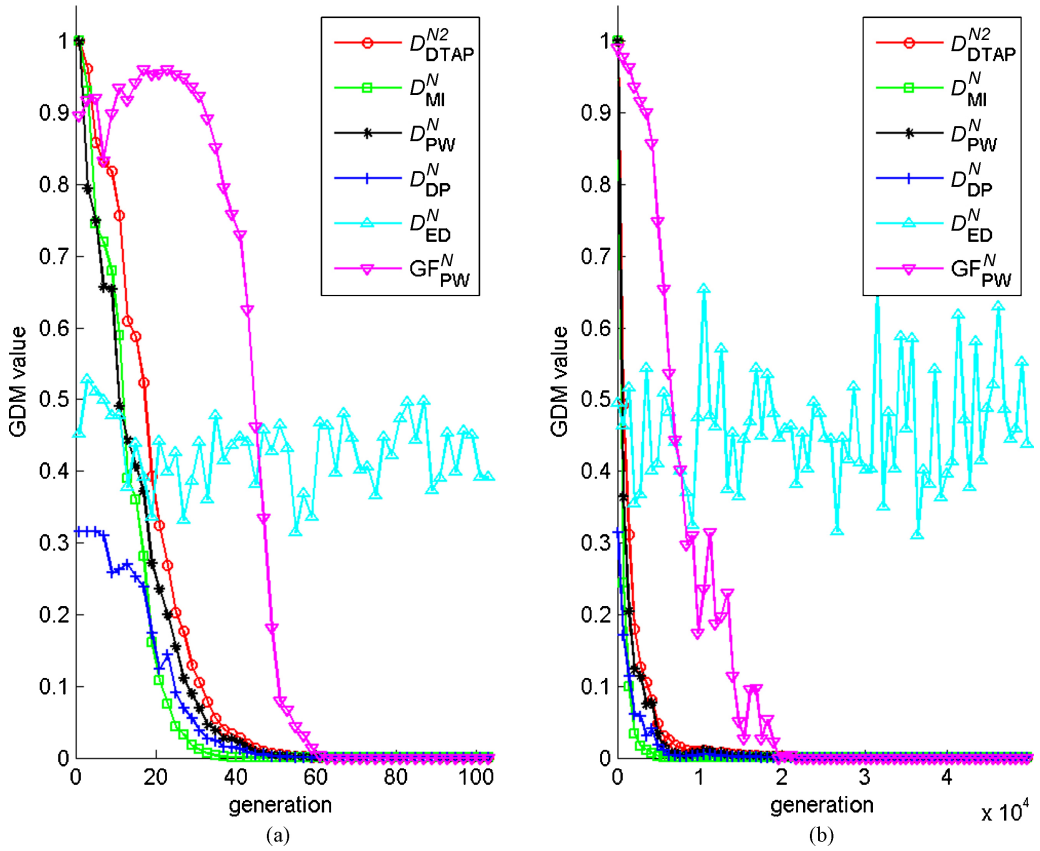

Fig. 13. Genotypic diversity level of various GDMs for the median repetition of the 10-D F2 function. (a) Solved with G-CMA-ES. (b) Solved with SSGA.

sections, an initial population size (N) of 100 was used, instead of $4 + \lfloor 3 \cdot \ln(n) \rfloor$. For the GF measures, $M = N$.

The latter EA is based on a real-coded steady-state genetic algorithm (SSGA), where the selection plan and genetic operators are specifically chosen to promote diversity. In fact, a negative assortative mating strategy is used, as well as BLX-0.5 and a BGA mutation operator. This combination was selected in a memetic algorithm (MA) context, where the main assumption is that an EA is responsible for focusing on exploration, and exploitation is driven by local search algorithms [43]. Nevertheless, the true behavior of the explorative search method is often only implicitly addressed. Therefore, the following experiment attempts to explicitly characterize the explorative capability of the chosen strategy by means of GDMs. The parameters used within this SSGA framework are the same as those defined by [43], except that the population size is fixed at 100 instead of 60, for the same reason as for G-CMA-ES.

A similar comparison was performed by Mattiussi *et al.* [45] for binary GDMs over the 2-D sine envelope sine wave function, and they did this using an SGA. They reported the average genotypic diversity over ten repetitions to demonstrate the similar behavior among different GDMs. However, due to the restart strategy of G-CMA-ES, and the fact that each repetition does not show the same convergence history, it is not helpful to compare the GDMs based on the average diversity obtained over the repetitions. Therefore, we have provided an analysis here for the median repetition of different CEC'05 benchmark functions.

To be concise, only the results of 10-D F2 (the shifted Schwefel problem 1.2) and 10-D F10 (the shifted rotated Rastrigin function) are presented, which are a unimodal and a multimodal landscape, respectively. For the median repetition, G-CMA-ES found the optimum within a 1e-6 tolerance in 8900 evaluations for F2, whereas the F10 optimum was achieved within a 1e-2 tolerance in 38 500 evaluations. In contrast, the SSGA implemented with diversity promoting features did not find the optimum within the CEC'05-prescribed tolerance, even after 100 000 evaluations.

Fig. 13 exposes the genotypic diversity history of F2, and Fig. 14 presents this history for F10. The restart strategy of G-CMA-ES is clearly observable over F10, where one restart was required, owing to the loss of all diversity without the global optimum being reached. To be comprehensive, only six GDMs are provided over these median runs; three of the most efficient measures (D^{N2}_{DTAP}, D^{N}_{MI}, and D^{N}_{PW}) and three of the worst descriptors (D^{N}_{DP}, D^{N}_{ED}, and $\mathrm{GF}^{N}_{\mathrm{PW}}$).

The discrimination problem of $\mathrm{GF}^{N}_{\mathrm{PW}}$, discussed in Section V-A, is clearly observable. In fact, this drawback, which characterizes all GF measures, can dramatically distort the conclusion drawn on the search algorithm behavior, as demonstrated for the SSGA simulation over F10. The normalization problem raised in Section II-D for LD-based measurements is noticeable with D^{N}_{DP}, and the inability of D^{N}_{ED} to describe genotypic diversity is demonstrated by its relatively constant value over the process. As a result, neither of these

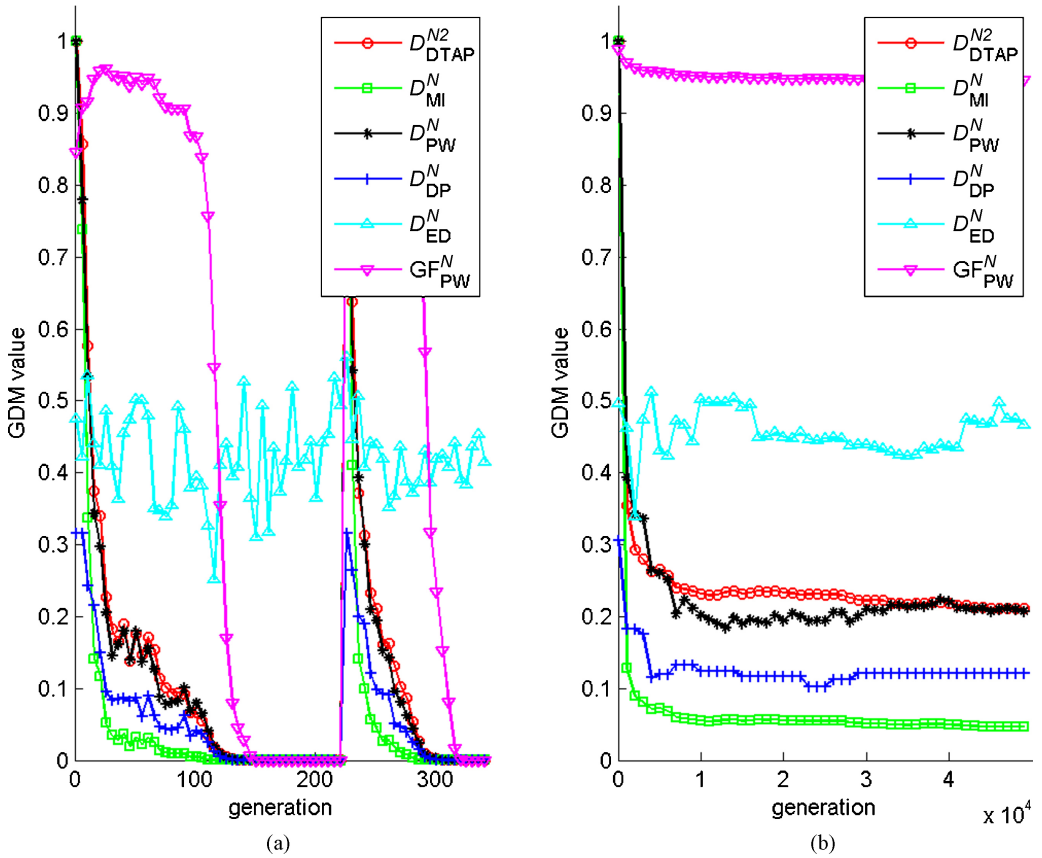

Fig. 14. Genotypic diversity level of various GDMs for the median repetition of the 10-D F10 function. (a) Solved with G-CMA-ES. (b) Solved with SSGA.

GDMs is helpful for estimating the genotypic diversity of a population.

In contrast, D^{N2}_{DTAP}, D^{N}_{MI}, and D^{N}_{PW} show comparable genotypic diversity monitoring without conclusive difference. For all the functions analyzed, they present a comparable monitoring trend with different diversity levels. This becomes quite obvious from observing SSGA history over F10 [Fig. 14(b)]. For the same function solved by G-CMA-ES, the maximum difference between D^{N2}_{DTAP} and D^{N}_{PW} is achieved at the third generation with a diversity gap of 0.34. Therefore, without any knowledge of the real diversity within the population, it is impossible to endorse the selection of any of these three GDMs. Furthermore, D^{N2}_{DTAP}, D^{N}_{MI}, and D^{N}_{PW} achieved a convergence state at the same evolutionary stage. For the F2 function, this behavior is expected, as it is characterized by only one convergence site. However, the multisite convergence phenomenon described in Section V-B is hidden from the multimodal F10 function because of the EA search bias. Indeed, G-CMA-ES converges toward a single location, which, by the way, proves the usefulness of a restart strategy. In contrast, at the end of the SSGA process, 90% of the individuals remain unique, and they do so with a radius threshold of 0.1 unit, or 1% of the distance between the F10 landscape frontiers. This is worth noting, considering the relatively low diversity of these three GDMs ($<$0.21) at the end of the process. In fact, we demonstrate that, even if no convergence status is monitored, all the individuals are neighbors. Furthermore, this happens quite rapidly during the process, as more than 75% of the generation stabilizes around this state. Therefore, it is possible that this particular SSGA strategy does not react as intended in the MA context. As a matter of fact, if the explorative strategy does not provide enough diversity, the occurrence of premature convergence could be exacerbated within an MA framework.

In summary, this analysis validated some of the GDM observations described in the previous sections. Nonetheless, the methodology has several limitations. The mere fact that each repetition has a different convergence history makes it impossible to use the mean GDM response that is necessary to reduce noise and produce sensitivity analyses that help to discriminate among GDMs. Also, the bias introduced by the EA does not allow multisite convergence search pattern to be visualized, which is of interest for GDM comparison purposes. By themselves, these shortcomings validate the formulation of a specific GDM comparative benchmark, as proposed in Section IV.

VII. Discussion

This paper has presented a detailed comparative study of more than 15 GDMs common in the EA domain. We define these measures as exploration descriptors, since they are related to the spatial location of individuals in a given population. In this investigation, the evolution process had to be controlled to ensure a population diversity that is known throughout the progression. This fact was reinforced by the analysis presented over the CEC'05 benchmark. We demonstrated that it is difficult to capture the fundamental

TABLE VII
QUALITATIVE RANKING OF THE DESCRIPTORS (0→UNRELIABLE, 1→WEAK, 2→GOOD, 3→EXCELLENT)

GDM	1-Single-Site Convergence	2-Multisites Convergence	3-Stability	4-Insensitivity with Respect to:			5-Outliers
				Dimension	Pop. Size	Interval	
D^{N2}_{DTAP}	3	0	3	2	1	–	3
D^{N}_{TD}	3	0	3	3	2	–	1
D^{N}_{MI}	1	0	3	3	2	–	2
D^{N}_{PW}	3	0	3	3	2	–	3
D^{N}_{VAC}	1	0	2	3	0	–	2
GF^{N}_{S}	1	1	3	3	0	0	1

properties of the various GDMs using an EA. This led to the development of a simple benchmark, which ensured the convergence of an initially fully scattered population in a chosen number of iterations. All the diversity measures were normalized to make it possible to compare them on the same basis. The results are summarized below.

Based on their formulation, five GDMs were eliminated prior to the comparative study: D^{N}_{DP}, D^{N}_{RP}, $D^{N}_{\mathrm{RP}}(f)$, D^{N}_{DTAP}, and D^{N}_{ED}. It was demonstrated that their underlying idea and/or their normalization method could be misleading in the genotypic diversity analysis. Therefore, these indicators are no longer recommended. Furthermore, D^{N}_{ALL} was not included in the comparative study, since its normalized version leads to D^{N}_{PW}.

Based on the GDM behavior requirements established in Section I, the five remaining distance-based GDMs (D^{N2}_{DTAP}, D^{N}_{TD}, D^{N}_{MI}, D^{N}_{PW}, and D^{N}_{VAC}) are capable of describing the intended diversity of the unimodal benchmark or single-site convergence problem, although some, because they are variance-based (D^{N}_{MI} and D^{N}_{VAC}), do so with more difficulty. In contrast, all the GF measures (GF^{N}_{S}, $\mathrm{GF}^{N}_{\mathrm{HC}}$, GF^{N}_{R}, and $\mathrm{GF}^{N}_{\mathrm{PW}}$) have the same shortcoming with respect to this benchmark, which is an inability to discriminate the diversity level until a nearly converged population state has been reached.

For the multisite convergence pattern, none of the available GDMs is capable of representing the diversity history. In fact, the multimodal experiments reveal that all distance-based GDMs (D^{N2}_{DTAP}, D^{N}_{TD}, D^{N}_{MI}, D^{N}_{PW}, and D^{N}_{VAC}) overestimate the end diversity, as no special treatment is afforded to duplicated individuals. Now, the GF measures have the same nondiscrimination issue throughout the scattered history of the population as in the case of the single-site convergence problem, even if they reach the intended convergence status level. It is worth noting that multisite convergence does not usually occur in conventional EAs, as the population is frequently steered one way or another toward only one convergence location. Incidentally, that is one of the root causes of premature convergence. Therefore, we shall account for multisite convergence with a GDM, in order to validate and appreciate new developments based on diversity promotion methods (such as niching methods [46]), or any other strategy aimed at improving EA performance.

That said, the available distance-based GDMs are at least potentially usable within standard EA frameworks. For a better depiction of the performance of the GDMs, the stability, sensitivity with simulation parameters, and consideration of outliers were also analyzed. From this stage onward, GF measurements were set aside in our presentation, owing to their poor power to discriminate diversity. All distance-based GDMs demonstrate stability characteristics that are good to excellent, like their insensitivity with respect to landscape dimensionality. In contrast, none of these GDMs provides excellent insensitivity with respect to population size. In fact, D^{N2}_{DTAP} and D^{N}_{VAC} could be considered very sensitive to this parameter. Finally, D^{N}_{PW} is the best GDM for adequately taking into account the presence of outliers.

The behaviors of GDMs are ranked qualitatively in Table VII, based on the comparative study results. GF^{N}_{S} is inserted as the representative GF measurement, with the aim of providing a global picture of the potential GDMs. This table clearly shows the multiobjective aspect of choosing the most interesting of them. Therefore, based on the dominance concept widely used to solve multiobjective optimization problems, we could assert that D^{N}_{PW} is the sole nondominated genotypic diversity indicator, which would make it the best available GDM. Nevertheless, as previously discussed, this GDM is not suitable for describing multisite convergence processes. As a result, a GDM formulation that is appropriate for dealing with any kind of search process remains an open question.

VIII. CONCLUSION

All things considered, this paper has demonstrated that no measurement was capable of reflecting the diversity of a population for any search process. Nonetheless, the development of this kind of measure may support the establishment of, for instance, the foundation for a feedback mechanism used in adaptive methods. In fact, these mechanisms are probably the most interesting application for diversity measures, as the GDM could be used to assess, in part, the quality of the EEB driving the optimization process.

REFERENCES

[1] K. A. De Jong, "An analysis of the behavior of a class of genetic adaptive systems," Ph.D. dissertation, Dept. Comput. Commun. Sci., Michigan Univ., Ann Arbor, 1975.
[2] M. L. Mauldin, "Maintaining diversity in genetic search," in *Proc. 4th Nat. Conf. Artif. Intell.*, 1984, pp. 247–250.
[3] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 1989.
[4] L. J. Eshelman and J. D. Schaffer, "Preventing premature convergence in genetic algorithms by preventing incest," in *Proc. 4th Int. Conf. Genet. Algorithms*, 1991, pp. 115–122.
[5] A. E. Eiben and C. A. Schippers, "On evolutionary exploration and exploitation," *Fundamenta Informaticae*, vol. 2, nos. 1–4, pp. 35–50, 1998.

[6] A. E. Eiben, R. Hinterding, and Z. Michalewicz, "Parameter control in evolutionary algorithms," *IEEE Trans. Evol. Comput.*, vol. 3, no. 2, pp. 124–141, Jul. 1999.
[7] A. K. Gupta, K. G. Smith, and C. E. Shalley, "The interplay between exploration and exploitation," *Acad. Manage. J.*, vol. 49, no. 4, pp. 693–706, Aug. 2006.
[8] S. Ishii, W. Yoshida, and J. Yoshimoto, "Control of exploitation-exploration meta-parameter in reinforcement learning," *Neural Netw.*, vol. 15, no. 4, pp. 665–687, Jun. 2002.
[9] J. Lee and Y. U. Ryu, "Exploration, exploitation and adaptive rationality: The neo-Schumpeterian perspective," *Simulat. Model. Practice Theory*, vol. 10, nos. 5–7, pp. 297–320, Dec. 2002.
[10] F. Herrera and M. Lozano, "Adaptation of genetic algorithm parameters based on fuzzy logic controllers," in *Genetic Algorithms and Soft Computing*, vol. 8, F. Herrera and J. L. Verdegay, Eds., 1st ed. Heidelberg, Germany: Physica-Verlag, 1996, pp. 95–125.
[11] E. K. Burke, S. Gustafson, and G. Kendall, "Diversity in genetic programming: An analysis of measures and correlation with fitness," *IEEE Trans. Evol. Comput.*, vol. 8, no. 1, pp. 47–62, Feb. 2004.
[12] N. F. McPhee and N. J. Hopper, "Analysis of genetic diversity through population history," in *Proc. 1st Genet. Evol. Comput. Conf.*, 1999, pp. 1112–1120.
[13] P. Monsieurs and E. Flerackers, "Reducing population size while maintaining diversity," in *Proc. 6th Eur. Conf. Genet. Program.*, vol. 2610. 2003, pp. 142–152.
[14] S. Lieberson, "Measuring population diversity," *Am. Sociol. Rev.*, vol. 34, no. 6, pp. 850–862, Dec. 1969.
[15] G. P. Patil and C. Taillie, "Diversity as a concept and its measurement," *J. Am. Stat. Assoc.*, vol. 77, no. 379, pp. 548–561, Sep. 1982.
[16] M. M. Gouvêa, Jr., and A. F. R. Araújo, "Diversity control based on population heterozygosity dynamics," in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2008, pp. 3671–3678.
[17] R. K. Ursem, "Diversity-guided evolutionary algorithms," in *Proc. 7th Conf. Parallel Problem Solving Nat.*, vol. 2439. 2002, pp. 462–471.
[18] H. A. Abbass and K. Deb, "Searching under multi-evolutionary pressures," in *Proc. 2nd Int. Conf. Multi-Criterion Optimiz.*, vol. 2632. 2003, pp. 391–404.
[19] R. W. Morrison and K. A. De Jong, "Measurement of population diversity," in *Proc. Select. Papers 5th Eur. Conf. Artif. Evol.*, vol. 2310. 2002, pp. 31–41.
[20] O. Olorunda and A. P. Engelbrecht, "Measuring exploration/exploitation in particle swarms using swarm diversity," in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2008, pp. 1128–1134.
[21] A. L. Barker and W. N. Martin, "Dynamics of a distance-based population diversity measure," in *Proc. IEEE Congr. Evol. Comput.*, Jul. 2000, pp. 1002–1009.
[22] M. Wineberg and F. Oppacher, "The underlying similarity of diversity measures used in evolutionary computation," in *Proc. Genet. Evol. Comput. Conf.*, vol. 2724. 2003, pp. 1493–1504.
[23] Y. Ichikawa and Y. Ishii, "Retaining diversity of genetic algorithms for multivariable optimization and neural network learning," in *Proc. IEEE Int. Conf. Neural Netw.*, Mar.–Apr. 1993, pp. 1110–1114.
[24] R. J. Collins and D. R. Jefferson, "Selection in massively parallel genetic algorithms," in *Proc. 4th Int. Conf. Genet. Algorithms*, 1991, pp. 249–256.
[25] F. Herrera, E. Herrera-Viedma, M. Lozano, and J. L. Verdegay, "Fuzzy tools to improve genetic algorithms," in *Proc. 2nd Eur. Congr. Intell. Tech. Soft Comput.*, 1994, pp. 1532–1539.
[26] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, Jul., Oct. 1948.
[27] J. Havrda and F. Charvát, "Quantification method of classification processes: Concept of structural a-entropy," *Kybernetika*, vol. 3, no. 1, pp. 30–35, 1967.
[28] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp.*, 1961, pp. 547–561.
[29] J. P. Rosca, "Entropy-driven adaptive representation," in *Proc. Workshop Genet. Program.*, 1995, pp. 23–32.
[30] T. K. Nayak, "On diversity measures based on entropy functions," *Commun. Statist.: Theor. Meth.*, vol. 14, no. 1, pp. 203–215, 1985.
[31] C. Gini, "Measurement of inequality of incomes," *Econ. J.*, vol. 31, no. 121, pp. 124–126, 1921.
[32] E. H. Simpson, "Measurement of diversity," *Nature*, vol. 163, p. 688, Apr. 1949.
[33] I. J. Good, "Comment: Diversity as a concept and its measurement by G. P. Patil and C. Taillie," *J. Am. Statist. Assoc.*, vol. 77, no. 379, pp. 561–563, Sep. 1982.
[34] L. Mei-Yi, C. Zi-Xing, and S. Guo-Yun, "An adaptive genetic algorithm with diversity-guided mutation and its global convergence property," *J. Cent. South Univ. Technol.*, vol. 11, no. 3, pp. 323–327, 2004.
[35] J. Riget and J. S. Vesterstrom, "A diversity-guided particle swarm optimizer: The ARPSO," EVALife Project Group, Aarhus Univ., Aarhus, Denmark, Tech. Rep. 2002-02, 2002.
[36] P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y.-P. Chen, A. Auger, and S. Tiwari, "Problem definitions and evaluation criteria for the CEC 2005 special session on real parameter optimization," School Elec. Electron. Eng., Nanyang Tech. Univ., Singapore, Tech. Rep. #2005005, May 2005.
[37] N. Hansen, A. Auger, S. Finck, and R. Ros, "Real-parameter black-box optimization benchmarking 2009: Experimental setup," Dept. Comput. Sci. Contr., INRIA, France, Res. Rep. RR-6828, May 2009.
[38] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," in *Proc. 1st Workshop Found. Genet. Algorithms*, 1990, pp. 69–93.
[39] N. I. Lyons and K. Hutcheson, "Comparing diversities: Gini's index," *J. Statist. Comput. Simul.*, vol. 8, no. 1, pp. 75–80, 1978.
[40] P. Sprent and N. C. Smeeton, *Applied Nonparametric Statistical Methods*. Boca Raton, FL: Chapman & Hall/CRC, 2001.
[41] S. Garcia, D. Molina, M. Lozano, and F. Herrera, "A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC'2005 special session on real parameter optimization," *J. Heuristics*, vol. 15, no. 6, pp. 617–644, Dec. 2009.
[42] A. Auger and N. Hansen, "A restart CMA evolution strategy with increasing population size," in *Proc. IEEE Congr. Evol. Comput.*, Sep. 2005, pp. 1769–1776.
[43] D. Molina, M. Lozano, C. Garcia-Martinez, and F. Herrera, "Memetic algorithms for continuous optimization based on local search chains," *Evol. Comput.*, vol. 18, no. 1, pp. 27–63, 2010.
[44] N. Hansen, "Compilation of results on the 2005 CEC benchmark function set," Computat. Lab., Inst. Computat. Sci., ETH Zurich, Switzerland, Tech. Rep., May 2006 [Online]. Available: http://www.lri.fr/∼hansen/cec2005compareresults.pdf
[45] C. Mattiussi, M. Waibel, and D. Floreano, "Measures of diversity for populations and distances between individuals with highly reorganizable genomes," *Evol. Comput.*, vol. 12, no. 4, pp. 495–515, 2004.
[46] S. Das, S. Maity, B.-Y. Qu, and P. N. Suganthan, "Real-parameter evolutionary multimodal optimization: A survey of the state-of-the-art," *Swarm Evol. Comput.*, vol. 1, no. 2, pp. 71–88, 2011.

Guillaume Corriveau received the B.E. degree in mechanical engineering from the École de Technologie Supérieure, Montreal, QC, Canada, in 2005, where he is currently pursuing the Ph.D. degree in engineering.

In 2010, he joined Bombardier Aerospace, Saint-Laurent, QC, to help develop multidisciplinary design optimization tools. His current research interests include evolutionary computation, adaptive systems, machine learning, structural optimization, and numerical simulations.

Raynald Guilbault received the Bachelors degree in 1994 and the Masters degree in 1995 from the Ecole Polytechnique Montréal, Montreal, QC, Canada, and the Ph.D. degree in 2000 from Laval University, Quebec City, QC, all in mechanical engineering.

He was a Mechanical Design Engineer with Alstom Power Canada, Inc., Edmonton, QC. He joined the Department of Mechanical Engineering, École de Technologie Supérieure, Montreal, QC, in 2001, as an Associate Professor. His current research interests include mechanical power transmission design, gear dynamics, contact mechanics, fatigue and wear, and shape optimization.

Dr. Guilbault is a Registered Professional Engineer in Quebec, Canada.

Antoine Tahan received the Masters degree in mechanical engineering and the Ph.D. degree in electrical engineering from Laval University, Quebec City, QC, Canada.

He is a Professor with the Department of Mechanical Engineering, École de Technologie Supérieure, Montreal, QC. His current research interests include industrial statistics, productivity improvement, quality control, and tolerance optimization.

Dr. Tahan is a Registered Professional Engineer in Quebec. He is a member of the American Society of Mechanical Engineers, the Canadian Machinery Vibration Association, and ASQ.

Robert Sabourin joined the Department of Physics, University of Montreal, Montreal, QC, Canada, in 1977, where he was responsible for the design, experimentation, and development of scientific instrumentation for the Mont Mégantic Astronomical Observatory.

In 1983, he joined the École de Technologie Supérieure, Université du Québec, Montreal, where he is currently a Full Professor. He is the author and co-author of more than 300 scientific publications, including journals and conference proceedings. His current research interests include handwriting recognition, signature verification, intelligent watermarking systems, and bio-cryptography.