Novel Combined Variable Selection Approach Using Memetic Algorithm With Complex Harmonic Regularization

XIAO-YING LIU
Computer Engineering Technical College, Guangdong Polytechnic of Science and Technology, Zhuhai 519090, China
e-mail: 631218194@qq.com
This work was supported in part by the Key Project for University of Department of Education of Guangdong Province of China Funds (Natural) under Grant 2019GZDXM005, and in part by the Chinese Ministry of Education’s Tian Cheng Hui Zhi Innovation and Education Improvement Funds under Grant 2018A01014.

ABSTRACT Variable selection has been highly successful in big data analyses, and regularization approaches are commonly used methods, which can automatically select important variables while constructing machine learning models. Due to the real datasets have complex relationships between relevant variable groups, many group-sparsity regularization approaches are proposed recently. However, these approaches usually are non-convex and sensitive to hyper-parameters. Therefore, optimization of the regularizations is a challenging task. In this paper, we present a novel memetic algorithm with the complex harmonic regularization (MA-CHR), which combines EC algorithm for hyper-parameters (global search) and complex harmonic regularization with path seeking strategy for the group-sparsity variable selection (local search). We further introduced a novel genetic individual representation (intron+exon) to efficient obtain the global optimal solution of this group-sparsity regularization. Simulation and five real data experiments demonstrate that the proposed MA-CHR method performs better than the state-of-the-art regularization methods in selecting groups of relevant variables and classification.

INDEX TERMS Variable selection, memetic algorithm, complex harmonic regularization.

I. INTRODUCTION Variable selection is one of the important issues in high-dimensional and massive data analysis, which has attracted more attention in machine learning. Regularizations are commonly used methods for variable selection, which can automatically select important variables while constructing machine learning model, and their applications have become increasingly popular. Lasso [1] is one of the most popular regularization methods. In order to obtain a more sparse solution, a growing number of sparse regularization methods are proposed, such as smoothly clipped absolute deviation penalty (SCAD) [2], adaptive Lasso [3], minimax concave plus method (MCP) [4], $L_p$ penalty [5], smoothed $L_p$ regularization [6], $L_{1/2}$ penalty [7], [8] and harmonic regularization [9].

However, beyond the sparsity of data, the real datasets generally have complex relationships between relevant variable groups, which are widely referred to as grouping effect [10]–[18]. That is, the datasets typically have high correlations or grouping effects between variables, especially in biological data, gene regulation networks and protein-protein interactions. Using prior knowledge, regularization methods can rapidly generalize to new tasks containing some correlation information with supervised learning. For instance, group Lasso [10] can be used for multivariate analysis based on the variance model, where each of these variables can be represented by a set of dummy variables and has multiple levels. In addition, some regularization methods can generate group-sparsity learning models without prior knowledge. For example, Elastic Net (EN) regularization [11] linearly combines the $L_1$ penalty with the $L_2$ penalty. Similarly, Zou and Zhang [14] studied an adaptive elastic network using the $L_2$ penalty in adaptive weighted Lasso. By incorporating the $L_2$ penalty, the SCAD-$L_2$ penalty [15] inherits good performance of SCAD and can be extended to select group variables. Huang et al. [16] studied a new regularization method,
which linearly combines the $L_{1/2}$ penalty and the $L_2$ penalty with logistic regression model for gene selection, termed as HLR. However, one limitation of the aforementioned methods ($L_2$ and $L_p$ with a fixed value of $p$) is that they often require the data to satisfy a certain probability distribution [17]. Very recently, Liu et al. [18] proposed a complex harmonic regularization (CHR), which can produce solutions that closely approximate to the combination of $L_p$ ($1/2 \leq p \leq 1$) and $L_q$ ($1 < q < 2$) regularizations with adjustable $p$ and $q$ to evaluate the sparsity and grouping effect of data.

However, most existing group-sparsity regularization methods inevitably involve hyper-parameters, like the control parameters to balance loss and penalty functions in machine learning models, and the mixing parameters to balance grouping effect and sparsity in datasets. The grid search method and $k$-fold cross-validation (CV) are typically used to adjust these hyper-parameters. Generally, the performance of the regularization methods is extremely susceptible to the hyper-parameters. This tends to raise their application difficulty and reduce their performance stability in practical problems.

Recently, evolutionary computation (EC) algorithms have been widely used for variable selection that benefits from their global optimization capability. Chen et al. [19] mentioned that more than five hundred related literatures on this topic have been published in recent years. Based on which aspect is enhanced using evolutionary search strategies, existing EC algorithms for variable selection can be categorized into the following perspectives: 1) filtering methods: including genetic algorithm (GA), genetic programming (GP), particle swarm optimization (PSO), ant colony optimization, differential evolution and evolutionary strategy (ES) [20]–[25]; 2) wrapper methods: including GA, GP, PSO, ES, ant colony optimization (ACO), differential evolution (DE), and estimation distribution algorithm (EDA) [26]–[32]; 3) embedded methods: including GP, etc. [19], [33]–[35]; 4) combination methods: including GA, PSO, ACO, DE and memetic algorithm (MA) [36]–[46].

In the combination methods, many evolutionary computation methods for variable selection, which combine filter and wrapper methods, provide an opportunity for population-based optimization with local search. For instance, Zhu et al. [40] proposed a filtering variable ordering method in MA, which is used to balance local and global search to improve the optimization results and efficiency. However, these two-stage memetic-based methods have the potential limitation that the filter phase may eliminate potentially useful variables regardless of their performance during the wrapper phase. On the other hand, the wrapper methods usually involve a series of evaluations, and each evaluation usually consumes a lot of time, especially when the dimension of variables is large.

To alleviate the limitations of the existing hybrid evolutionary-computation methods for variable selection, and help to globally optimize hyper-parameters of the group-sparsity regularization methods, we present a novel memetic algorithm with the complex harmonic regularization (MA-CHR) for variable selection in this paper. The proposed method combines the evolutionary optimization of the hyper-parameters (global search) with the complex harmonic regularization of the group-sparsity variable selection (local search). The memetic optimization procedure is represented by the genetic chromosomes, including introns (hyper-parameters) and exons (variable coefficients in learning models). Simulation and real data experiments substantiate the capability of the proposed method for prediction and variable selection.

The remaining sections are organized as follows. Section II reviews the related work of regularization methods and evolutionary computation algorithms for variable selection. Section III introduces the proposed method as well as the detailed algorithm. Section IV presents the results of simulation and real data experiments. Finally, the conclusion is given in Section V.

II. RELATED WORKS

A. REGULARIZATION METHODS

Regularizations are important embedded methods of variable selection. Suppose that dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $n$ represents the number of samples, $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ represents the $i$th sample with $t$ dimensions, and $y_i$ represents the variable of corresponding dependent. The linear regression with the regularization term can be expressed as:

$$L(\lambda, \beta) = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y - X'i\beta)^2 + \lambda P(\beta) \quad (1)$$

where $P(\beta)$ represents the penalty function, and $\lambda$ represents the hyper-parameter for controlling the complexity of the model. As the most commonly used regularization method, Lasso is given as $P(\beta) = ||\beta||_1$, which can construct the model while selecting the variables. In order to obtain more sparse solutions, a growing number of sparse regularization methods based on $L_1$ norm are proposed, such as SCAD [2], adaptive Lasso [3], MCP [4]. The $L_0$ regularization yields the most sparse solutions, which is given as $P(\beta) = ||\beta||_0$, but it faces the NP-hard combinatorial optimization problem.

In order to improve the prediction accuracy of the machine learning model, the researchers studied $L_p$ ($0 < p < 1$) regularization with a fixed $p$, especially $p = 1/10, 1/2, 2/3, 9/10$ [21]. The work in [7], Xu et al. proposed the $L_{1/2}$ regularization, which can be used as a representative of the $L_p$ ($0 < p < 1$) penalty. They analyzed its threshold representation [8] and demonstrated that solving the $L_{1/2}$ regularization is easier than solving the $L_1$ regularization. Furthermore, they also proved that the $L_{1/2}$ regularization possesses unbiasedness and Oracle properties [8], which make the $L_{1/2}$ regularization an effective tool for solving high-dimensional variable selection problems [47].

The aforementioned sparse regularization methods, however, do not consider the grouping effects. If faced with a set of variables with high correlation, the sparse regularization method tends to select only one variable to represent the
The complex harmonic regularization (CHR) can be expressed as:

\[ \hat{\beta} = \arg\min_{\beta} \left\{ L(\beta) + \lambda_1 \sum_{j=1}^{k} m(\beta_j) + \lambda_2 \sum_{j=1}^{k} n(\beta_j) \right\} \]

where

\[ m(\beta) = \sqrt{\frac{2}{a(a + 1)}} |\beta| + \left( \frac{1 - a}{a} \right)^2 - \frac{1 - a}{a} \]
\[ n(\beta) = \sqrt{\frac{2}{b(b + 1)}} |\beta| + \left( \frac{1 - b}{b} \right)^2 - \frac{1 - b}{b} \]

\(a\) and \(b\) represent the hyper-parameters of the CHR penalty. The first part \(m(\beta)\) of the CHR penalty approximates to the \(L_1\) regularization. When the parameter \(a\) is close to 1, \(m(\beta) \approx |\beta|^{1/2}\) (Figure 1-(b)), the first part \(m(\beta)\) of the CHR penalty approximates to the \(L_{1/2}\) regularization. When the parameter \(b\) is close to 0, \(m(\beta) \approx |\beta|^2\) (Figure 1-(c)), the first part \(m(\beta)\) of the CHR penalty approximates to the \(L_2\) regularization. When the parameter \(b\) is close to 1, \(n(\beta) \approx |\beta|\) (Figure 1-(f)), the second part \(n(\beta)\) of the CHR penalty approximates to the \(L_1\) regularization. Moreover, the CHR penalty has the property that its first derivative is finite at origin, which implies that the corresponding regularization problem can be efficiently solved via the direct seeking techniques.

In theory, the sufficient condition for the grouping effect of variables is the strict convex penalty function, and the \(L_q (1 < q < 2)\) penalty ensures that the function has strict convexity [11]. Therefore, the second part of the CHR penalty, \(n(\beta)\), approximates the \(L_q (1 < q < 2)\) penalty while inducing a grouping effect. The experimental results of [18] show that the CHR method of artificial and real gene expression data is very promising.

**B. EVOLUTIONARY COMPUTATIONS FOR VARIABLE SELECTION**

Evolutionary computation (EC) randomly generates an initial population of candidate solutions in the search space, and then iteratively updates through intersection, mutation, and selection operators to obtain the optimal solution. In recent years, more and more researchers have applied the local search (LS) technology to the random search process of EC to improve optimization efficiency [22]. These hybrid algorithms are often referred to as the Memetic Algorithm (MA). The MA algorithm combines wrapper and filter variable evaluation metrics for variable selection, providing a local search opportunity for population-based optimization. For example, Zhu et al. [40] proposed a filtering variable ordering method in MA, which is used to balance local and global search to improve optimization results and efficiency. Then, Zhu et al. [36] integrated the Markov coverage method into MA to simultaneously identify all relevant or partially related variables. The work in [44] proposed another two-stage variable selection algorithm in which each variable is sorted using the Relief-F algorithm, and then the top-ranked variable is used as the input to the memetic package variable selection algorithm. Other researchers have introduced a heuristic blending mechanism that combines filter ranking scores to guide the GA and PSO search process for packaging variable selection [26]–[29]. In addition, MA for variable selection has been used to solve practical application problems such as optimal controller design [48], motif discovery in DNA, microRNA and protein sequences [49], [50].

As is shown above, in most hybrid evolutionary computation methods for variable selection, the EC stage is used for wrapper feature selection, and the filtering-based LS algorithm helps to achieve a local optimal solution. However, these “wrapper + filter” two-stage memetic methods cannot
guarantee that the selected variables in the filtering phase are the best candidates for the EC phase. Therefore, the filter phase of the MA algorithm has the potential to eliminate potentially useful variables, regardless of their performance during the wrapper process.

III. MEMETIC ALGORITHM WITH COMPLEX HARMONIC REGULARIZATION

To alleviate the limitations of the existing hybrid evolutionary computation methods for variable selection, and help to globally optimize hyper-parameters of the group-sparsity regularization methods, we present a novel memetic algorithm with complex harmonic regularization (MA-CHR) for variable selection. The proposed method combines the evolutionary optimization of the hyper-parameters (global search) with the complex harmonic regularization of the group-sparsity variable selection (local search). The memetic optimization procedure is represented by the genetic chromosomes, including introns (hyper-parameters) and exons (variable coefficients in learning models). In the first step of the MA-CHR method, the evolutionary population is randomly initialized with each chromosome encoded by an intron and an exon part. Subsequently, the CHR regularization (local search) is performed on the exon parts under the fixed intron parts to achieve a local optimal solution or to improve the adaptability of individual in the search population. Genetic operators such as mutations and crossovers are performed on the intron parts of the chromosome. After that, the selection operator is used to produce the next population. This process is repeated until the stop condition is satisfied. Each component is explained below.

A. CHROMOSOME REPRESENTATION: INTRON AND EXON

In the proposed MA-CHR method, a representation for the four hyper-parameters $\lambda_1$, $\lambda_2$, $a$, $b$ and the coefficients $\beta_1, \beta_2, \ldots, \beta_t$ of the candidate variable subset can be encoded as a chromosome: intron+exon=($\lambda_1$, $\lambda_2$, $a$, $b$, $\beta_1, \beta_2, \ldots, \beta_t$). The length of the chromosome is denoted as $t + 4$, where $t$ represents the total number of variables. The chromosome is represented as a string that is globally optimized by its EC operator for its intron parts. Although the search space of the intron portion has non-convex and multi-peak properties, the EC can find a global optimal solution, because the dimension of the intron is very low. On the other hand, the exon part is optimized by the regularization method for synchronizing variable selection and building a learning model. In the exon part, the value of $\beta_i$ is nonzero means that the model has selected the corresponding variable. Conversely, if the $\beta_i$ coefficient of the candidate variable is equal to zero, it means that the candidate variable has been rejected. $T$ represents the maximum non-zero $\beta_i$ exon number allowed in each chromosome. When the optimal number of variables has relevant prior knowledge, we can limit $T$ to no more than the predetermined value; otherwise $T = t$.

B. OBJECTIVE FUNCTION

The objective function is defined by:

$$\text{Fitness}(\text{chromosome}) = \text{Accuracy of the classification model with } (\lambda_1, \lambda_2, a, b, \beta_1, \beta_2, \ldots, \beta_t)$$

where nonzero $\beta_i$ denotes the corresponding selected variable subset encoded in the exon part of the chromosome. The objective function evaluates the significance of the given variable subset. In this paper, the fitness of the objective function is specified as the classification accuracy of the logistic regularization model with the chromosome ($\lambda_1, \lambda_2, a, b, \beta_1, \beta_2, \ldots, \beta_t$), using the complex harmonic regularization method. Note that when two chromosomes are found to have similar fitness, i.e. the difference between their fitness is smaller than the small value of $e$, the chromosome with a smaller number of selected variables is more likely to be selected for the next generation.

C. LOCAL SEARCH IMPROVEMENT PROCEDURE

(LS)-PATH SEEKING APPROACH FOR THE SPARSE LOGISTIC REGRESSION WITH THE CHR PENALTY

In this section, we consider the use of complex harmonic regularization and path search algorithms as memes or local search methods for our proposed MA-CHR method. Usually, the computation time of the path search algorithm [51] increases linearly with the dimension of the variable selection problem. Therefore, this method is an effective way to solve the problem of regularization.

When applying $Lp(0 < p < 1)$ regularization, the path search algorithm cannot be used directly. Therefore, we also need to solve the optimization problem of $Lp(0 < p < 1)$ penalty. Fortunately, we can use the direct path finding method to overcome the CHR penalty. Popular path search methods based on squared errors include partial least squares regression (PLS) [52], forward stepwise regression [53], minimum angle regression [12], piecewise linear path [54] and gradient enhancement [55]. Friedman [56] proposed a generalized path search that produces solutions that are very close to any convex loss function and non-convex constraint. The advantage of the pathfinding method is that we provide a new method for solving non-convex penalty regularization. Let $v$ measure the length along the path, $\Delta v > 0$ is a small increment.

$$\phi_j(v) = \begin{cases} \delta \left( \sum_{j=1}^{k} m(\beta_j) + (1 - \gamma) \sum_{j=1}^{k} n(\beta_j) \right) & \beta = \hat{\beta}(v) \\ \frac{\gamma}{\sqrt{2a(a - 1) |\beta_j| + (1 - a^2)^2}} + \frac{2(1 - \gamma) |\beta_j|}{\sqrt{2b(b - 1) |\beta_j| + (1 - b^2)^2}} & \beta = \hat{\beta}(v) \end{cases}$$
\[ \phi_j(v) = - \frac{\delta L(\beta)}{\delta \beta_j} \bigg|_{\beta = \hat{\beta}(v)} = - \delta_1 \sum_{i=1}^{h} (y_i - \sum_{j=0}^{k} \beta_j x_{ij})^2 \bigg|_{\beta = \hat{\beta}(v)} = - \frac{2}{h} \sum_{i=1}^{h} x_{ij}(y_i - \sum_{j=0}^{k} \beta_j x_{ij}) \bigg|_{\beta = \hat{\beta}(v)} \]

Here \( \phi_j(v) \), \( \phi_i(v) \) are the \( j \)th gradient components of the logistic loss function and penalty function respectively, and \( \tau_j(v) = \phi_j(v) / \phi_i(v) \). Then, we give the path seeking algorithm for the CHR regularization as follows.

**Algorithm 1** The Path Seeking Algorithm for CHR Penalty

1: Initialize: \( v = 0 \).
2: repeat
3: Compute \( [T_j(v)]^k \).
4: \( S = \{ j \mid T_j(v) \cdot \hat{\beta}(v) < 0 \} \).
5: if \( S \) is empty then
6: \( j^* = \arg \max_j |T_j(v)| \).
7: else
8: \( j^* = \arg \max_{j \in S} |T_j(v)| \).
9: end if
10: \( \hat{\beta}_j^*(v + \Delta v) = \hat{\beta}_j^*(v + \Delta v \cdot sign(T_j^*(v))) \).
11: \( \{ \hat{\beta}_j^*(v + \Delta v) = \hat{\beta}_j^*(v) \}_{j \neq j^*} \).
12: \( v \leftarrow v + \Delta v \).
13: until \( T(v) = 0 \).

After finishing initialization of the path and computation of vector \( \tau(v) \), we need to identify that those signed coefficients \( \hat{\beta}_j(v) \neq 0 \) are opposite to that of the corresponding \( \tau_j(v) \). If the set \( S \) becomes empty, the selected coefficients of the absolute value need to correspond to the maximum element \( \tau_j(v) \) at Step 6. Moreover, if there are one or more than one elements in the set \( S \), the coefficient with corresponding maximum \( |\tau_j(v)| \) within this subset will be selected. Then, the selected coefficient \( \hat{\beta}_j(v) \) will be increased a bit in the direction of its signed correspond \( \tau_j(v) \), and keeping all other coefficients not change as well as producing the result for the next path point \( v + \Delta v \). Next, the iterations keep running until all components of \( \tau(v) \) become 0 and the algorithm gets the regular result of the CHR penalty. Therefore, MA-CHR has the capability to construct the learning model and select the relevant variables which can group effect efficiently and synchronously.

**D. EVOLUTIONARY OPERATORS**

In the MA-CHR, there is an evolution process has standard genetic operators which includes fitness proportionate selection, one-point crossover and uniform mutation operators. In addition, if the prior knowledge about optimal number of variables is available, the nonzero-number of \( \beta \) in every exon of the chromosome will maybe constrained to the largest value of \( T \) during evolution process.

1) Selection operator: The standard roulette-wheel selection applies in generating the next generation which includes the parent-generation and child population. The probability of the chromosome selection is proportional to its adaptive faculty directly. In the genetic selection phase, the candidate chromosomes with lower MSEs may be unlikely to be erased but there is still has the chance to implement it.

2) Crossover operator: First, we select two parents \((p_c, p_f)\) from current population for producing children randomly. Then, execute the crossover operation to produce offsprings that inherit characteristics from both parents. Next, in order to generate a single crossover point on the intron of both \( p_c \) and \( p_f \) chromosomes among the hyper-parameter \( \lambda_1, \lambda_2, a \) and \( b \), these four hyper-parameters on both sides of the point are exchanged in the intron of the parent’s chromosomes to produce \( c_e \) and \( c_f \) in the intron part of the offspring’s chromosomes. Finally, the exon part \( \beta \) of chromosomes of the two offspring is evaluated by the local optimization strategies.

3) Mutation operator: The mutation operator allows for greater diversity of populations diversity and search space. In this stage, we select one of the hyper-parameters \( \lambda_1, \lambda_2, a \) and \( b \) with a mutation probability \( p_m \) \((p_m = 0.1 \) in our experiments\) to mutate its value randomly. Then, local optimization strategy was used to evaluate the heterozygosity and exon \( \beta \) of the new chromosome produced by mutation operation.

**IV. RESULTS AND DISCUSSION**

To evaluate the performance of the proposed method, we implement experiments on simulated datasets with three scenarios and five publicly available microarray gene expression datasets. To compare with the proposed method, we apply GA, GP, MA, Elastic Net (EN), SCAD-L2, the hybrid \( L_{1/2} + L_2 \) regularization (HLR) and the complex harmonic regularization (CHR).

**A. ANALYSIS OF SIMULATED DATA**

The simulated data is generated from the logistic regression model, and has the following form:

\[ \log \left( \frac{y}{1-y} \right) = X\beta + \sigma \varepsilon, \quad \varepsilon \sim N(0,1) \]

where \( X \sim N(0,1) \), \( \varepsilon \) is the independent random error and \( \sigma \) is the parameter which controls the signal to noise. Three scenarios are simulated in the simulated experiments. In each scenario, the dimension of variables is 2000. The notation / represents the number of observations in the training and test sets, respectively, e.g. 100/100. Here are the details of the three scenarios.
(a) In scenario 1, the dataset consists of 300/300 observations, the noise control parameter \( \sigma \) is set to be 0.1 and
\[
\beta = \begin{bmatrix} 1, 1, \ldots, 1, 1, 0, \ldots, 0, 1, -1, \ldots, -1, 0, \ldots, 0, \\
2, 2, \ldots, 2, 0, \ldots, 0, 2, -2, \ldots, -2, 0, \ldots, 0, \\
100, 400, 100, 400, 100, 400, 100, 400 \\
\end{bmatrix}.
\]

We did the simulation on a grouped variable situation
\[
x_i = \rho \times x_1 + (1 - \rho) \times x_j, \quad i = 2, 3, \ldots, 100;
\]
\[
x_i = \rho \times x_{501} + (1 - \rho) \times x_i, \quad i = 502, 503, \ldots, 600;
\]
\[
x_i = \rho \times x_{1001} + (1 - \rho) \times x_i, \quad i = 1002, 1003, \ldots, 1100;
\]
\[
x_i = \rho \times x_{1501} + (1 - \rho) \times x_i, \quad i = 1502, 1503, \ldots, 1600.
\]
where \( \rho \) is the correlation coefficient of the grouped variables. An ideal sparse regression method would select only the 400 true variables and set the coefficients of the 1600 irrelevant variables to zero.

(b) Scenario 2’s definition is similar to Scenario 1. However, Scenario 2 except that if the case has other independent factors, which also contribute to the decision variable \( y \):
\[
\beta = \left( \frac{1, -1, \ldots, 1, -1, 1, \ldots, 1, -1, 1, \ldots, 1, -1, 1, \ldots, 1, -1, 1, \ldots, 1, -1}{100, 5 \times 20, 800, 5 \times 20, 800} \right)
\]
\[
x_i = \rho \times x_1 + (1 - \rho) \times x_i, \quad i = 2, 3, \ldots, 100;
\]
\[
x_i = \rho \times x_{501} + (1 - \rho) \times x_i, \quad i = 502, 503, \ldots, 600;
\]
\[
x_i = \rho \times x_{1001} + (1 - \rho) \times x_i, \quad i = 1002, 1003, \ldots, 1100;
\]
\[
x_i = \rho \times x_{1501} + (1 - \rho) \times x_i, \quad i = 1302, 1303, \ldots, 1400.
\]
In this case, the correlated variables consist of three groups, 400 correlated variables and 300 independent variables. An ideal sparse regression method would select only the 1000 real variables and the coefficients of the 1000 irrelevant variables would be set to 0.

(c) In Scenario 3, the number of real variables were increased to 1000 totally, \( \sigma = 0.1 \), and the dataset is constituted by 500/100 observations, and
\[
\beta = \left( \frac{1, -1, \ldots, 1, -1, 1, \ldots, 1, -1, 1, \ldots, 1, -1, 1, \ldots, 1, -1, 1, \ldots, 1, -1}{100, 5 \times 20, 300, 5 \times 20, 100} \right)
\]
\[
x_i = \rho \times x_1 + (1 - \rho) \times x_i, \quad i = 2, 3, \ldots, 100;
\]
\[
x_i = \rho \times x_{1001} + (1 - \rho) \times x_i, \quad i = 1002, 1003, \ldots, 1100;
\]

In this example, there are two correlated variables’ groups and 200 single independent variables. An ideal sparse regression method would select the 400 real variables and the coefficients of the 1600 irrelevant variables would be set to 0.

### Table 1: The parameter setting of the EC algorithms for seven methods.

| Parameter        | Value     |
|------------------|-----------|
| population size \( (P) \) | 200       |
| crossover probability \( (p_c) \) | 0.85      |
| mutation probability \( (p_m) \) | 0.1       |
| stopping criterion \( (G) \) | 2000      |
| \( x_1 = \rho \times x_1 + (1 - \rho) \times x_i, \) | \( i = 2, 3, \ldots, 100; \) |
| \( x_1 = \rho \times x_{501} + (1 - \rho) \times x_i, \) | \( i = 502, 503, \ldots, 600; \) |
| \( x_1 = \rho \times x_{1001} + (1 - \rho) \times x_i, \) | \( i = 1002, 1003, \ldots, 1100; \) |
| \( x_1 = \rho \times x_{1501} + (1 - \rho) \times x_i, \) | \( i = 1302, 1303, \ldots, 1400. \) |

In this case, the real variables were set to 0.1, 0.4, 0.7 singly. GA, MA, Elastic net (EN), SCAD-L2, HLR, CHR and MA-CHR are adopted the logistic classification approach. In GA, it adopts the multi-tree classifier. In each iteration of GA and MA, the number of selected variables based on the information gain filter is set to 1000. Table 1 lists the configuration parameters used by the EC algorithm in these eight methods. The model parameters of EN, SCAD-L2, and HLR are adjusted by 10-fold cross-validation. It is worth noting that EN and HLR are tuned by the 10-fold cross-validation on the two-dimensional parameter surfaces. Moreover, the model parameters of SCAD-L2 and CHR are tuned by the 10-fold cross-validation on the three-dimensional and the four-dimensional parametric surfaces, respectively. We repeated simulations of all the competing methods 100 times and calculated the average classification accuracy on the test set. In order to evaluate the variable selection performance of all competing methods, we adopt the sensitivity and specificity to evaluate the variable selection performance, which can be defined as follows [21]:

\[
\text{True Negative(TN)} := \left| \hat{\beta} \right|, \\
\text{False Positive(FP)} := \left| \hat{\beta} \right|.
\]
TABLE 2. Performance comparison on simulated datasets. The best results are highlighted in bold.

| Scenario | Sensitivity of variable selection | Specificity of variable selection | Accuracy of classification |
|----------|-----------------------------------|-----------------------------------|--------------------------|
| 0.1      | GA                                | 0.857                            | 0.564                    | 0.607                    |
|          | GP                                | 0.872                            | 0.710                    | 0.662                    |
|          | MA                                | 0.884                            | 0.729                    | 0.672                    |
|          | EN                                | 0.910                            | 0.825                    | 0.736                    |
|          | SCAD-L2                           | 0.911                            | 0.825                    | 0.736                    |
|          | HLR                               | 0.917                            | 0.871                    | 0.782                    |
|          | CHR                               | 0.926                            | 0.897                    | 0.810                    |
|          | MA-CHR                            | 0.941                            | 0.935                    | 0.851                    |
| 0.4      | GA                                | 0.742                            | 0.693                    | 0.712                    |
|          | GP                                | 0.753                            | 0.708                    | 0.724                    |
|          | MA                                | 0.759                            | 0.714                    | 0.730                    |
|          | EN                                | 0.811                            | 0.742                    | 0.755                    |
|          | SCAD-L2                           | 0.842                            | 0.783                    | 0.768                    |
|          | HLR                               | 0.801                            | 0.866                    | 0.792                    |
|          | CHR                               | 0.903                            | 0.981                    | 0.926                    |
|          | MA-CHR                            | 0.978                            | 0.954                    | 0.943                    |
| 0.7      | GA                                | 0.823                            | 0.797                    | 0.815                    |
|          | GP                                | 0.834                            | 0.815                    | 0.828                    |
|          | MA                                | 0.865                            | 0.833                    | 0.866                    |
|          | EN                                | 0.865                            | 0.842                    | 0.871                    |
|          | SCAD-L2                           | 0.870                            | 0.865                    | 0.872                    |
|          | HLR                               | 0.796                            | 0.780                    | 0.822                    |
|          | CHR                               | 0.872                            | 0.797                    | 0.822                    |
|          | MA-CHR                            | 0.834                            | 0.765                    | 0.776                    |

False Negative (FN) := \( | β^* \setminus \hat{β} | \),
True Positive (TP) := \( | β \cap \hat{β} | \),
Sensitivity := \( \frac{TP}{TP + FN} \),
Specificity := \( \frac{TN}{TN + FP} \).

where the \( \cdot \) is the element-wise product, and \( | \cdot |_0 \) calculates the number of non-zero elements in a vector, \( \hat{β} \) and \( \hat{β} \) are the logical “non” operators on the true coefficients vector \( β \) and the simulated \( \beta \).

Table 2 shows the variable selection and classification performance of all the competing methods in all the scenarios. It can be observed that the smaller correlation coefficient \( ρ \), the better performance of the methods. In Table 2, the proposed method always selects the most correct relevant variables in different scenarios. MA-CHR obtains the highest sensitivities and specificities of variable selection, which means that MA-CHR can select the most relevant variables while discarding the most irrelevant variables. The classification accuracy of MA-CHR is also superior to other competitive methods.

B. ANALYSIS OF REAL DATA

In this section, we choose five publicly available gene expression microarray datasets to further estimate the performance of the proposed method. Five publicly available datasets are AML, DLBCL, Prostate, Lymphoma and Lung cancer, respectively. The AML dataset was mentioned by Bullinger et al. [57] in the first time and this dataset included 116 patients with 6283 genes. The DLBCL was first proposed by Rosenwald et al. and it has almost 240 samples’ information [58] which every sample contains the gene-number of expression data is 8810. The Prostate dataset is a dataset includes 12,600-gene expression profiles for 50 normal tissues and 52 prostate tumour tissues and it was mentioned by Singhet et al. in the first time [59]. The Lymphoma dataset includes 77 microarray gene expression profiles for two of the most common adult lymphoid malignancies which contains 58 samples of diffuse large B-cell lymphomas and 19 Follicular Lymphomas (FL) [60]. It should be noted that the original Lymphoma data includes 7,129 gene expression values. The Lung cancer data set consists of 164 samples of 87 lung adenocarcinoma and 77 adjacent normal tissues with 22,401 microarray gene expression profiles [61]. The Lung cancer dataset is a public dataset and it can be downloaded at www.ncbi.nlm.nih.gov/geo with through access number (GSE40419). A summary of these datasets is shown in Table 3.

In order to effectively evaluate the performance of all the competing methods, the real data set is randomly divided into two parts: 2/3 samples are in the training set for model estimation, and the remaining 1/3 data are used to test the estimation performance. The model parameters are turned by 10-fold cross-validation (CV).

Table 4 reports the average training and test accuracy calculated by all competing methods over 100 repetitions. It can be easily seen that the performance of the MA-CHR method is better than other seven competitive methods. Moreover, the number of selected genes by all competing methods is also shown in Table 4. Compared to the other competitive methods, the number of genes selected by the proposed method is the smallest. In group-sparsity regularization methods, such as Elastic net, SCAD-\( \ell_2 \), HLR, and CHR, the performance of CHR is better than that of Elastic net, SCAD-\( \ell_2 \) and HLR in gene selection. On the other side, in EC algorithms, such as GA, GP and MA, the performance of GP is better than that of GA and MA in gene selection and classification. It is shown that the proposed method has better performances in gene selection and prediction classification compared with seven competitive methods.

C. DISCUSSION

Table 5 lists the results of the biological analysis. The table shows the top 10 genes got by different methods in the AML data set. Compared to existing variable selection methods, the MA-CHR model is able to select unique genes such as SFRP1 and SFRP2, which are members of the Sfrp family.
and a signal transduction protein. The Sfrp family of proteins transmits TGF-β signals from cell surface receptors to the nucleus. This gene is very important in the mutation or deletion of AML disease and has been shown to cause pancreatic cancer [61]. We consider it may be closely related to AML disease. Among other genes selected by the MA-CHR method, the gene FLT3 has been found to be up-regulated in several different types of AML diseases [62]. This gene can stimulate the motility of AML disease. It has been reported that the protein encoded by the NPM1 gene is very similar to the tumour suppressor of Drosophila, a highly related gene of AML disease [63]. In addition, MA-CHR can also find some related genes chose by other regularization models with Elastic Net, SCAD-$\ell_2$, HLR and CHR methods, such as SFRP5 and GSTM1. They are significantly associated with AML disease and have been discussed in [57]. To explore the interaction network between these genes selected by eight methods, we built up the integration of intergenic interactions using cBioPortal [64] by integrating biological interactions from the public acute myeloid leukemia (TCGA, NEJM 2013) database. Figure 2 shows the maximum interactive network view of the top 10 genes selected by eight different methods. For Elastic Net, in the largest network view, only two genes, PTPN6 and CDKN2A, are linked to other frequently altered genes. SFRP1, CDKN2B, DAPK1 and PTPN6, which are linked to other frequently altered genes from the acute myeloid leukemia (TCGA, NEJM2013) database, were selected by GP. For GA, MA and SCAD-$\ell_2$, 6 genes were identified by these three methods, which are related to other frequently altered genes. For the proposed MA-CHR method, it identified 8 genes in the same interactive network. This result indicates that MA-CHR performs better than other methods in selecting related genes.

Similar results were obtained from the analysis of four other real gene expression data sets. Biological analysis shows that MA-CHR model can not only find the related genes selected by other variable selection methods but also find some unique genes. These genes are not selected by other models but are significantly related to the disease. Therefore, MA-CHR method can be used to identify related genes accurately and effectively.

### TABLE 4. Performance comparison on five publicly available gene expression microarray datasets. The best results are highlighted in bold.

| Dataset | Method | Training accuracy (10-CV) | Test accuracy | No. of selected genes |
|---------|--------|--------------------------|---------------|-----------------------|
| AML     | GA     | 0.953                    | 0.918         | 37                    |
|         | GP     | 0.972                    | 0.922         | 25                    |
|         | MA     | 0.965                    | 0.911         | 29                    |
|         | SCAD-L$_2$ | 0.962                   | 0.929         | 26                    |
|         | HLR    | 0.976                    | 0.936         | 25                    |
|         | CHR    | 0.979                    | 0.958         | 23                    |
|         | MA-CHR | 0.984                    | 0.943         | 18                    |
|          | **MA-CHR** | **0.972**               | **0.937**     | **11**                |
| DLBCL   | GA     | 0.954                    | 0.896         | 45                    |
|         | GP     | 0.935                    | 0.912         | 29                    |
|         | MA     | 0.934                    | 0.903         | 27                    |
|         | SCAD-L$_2$ | 0.955                   | 0.921         | 19                    |
|         | HLR    | 0.968                    | 0.931         | 17                    |
|         | CHR    | 0.971                    | 0.928         | 14                    |
|         | MA-CHR | 0.985                    | 0.943         | 25                    |
| Lymphomas | GA     | 0.903                    | 0.883         | 40                    |
|         | GP     | 0.926                    | 0.913         | 39                    |
|         | MA     | 0.957                    | 0.908         | 37                    |
|         | SCAD-L$_2$ | 0.952                   | 0.928         | 25                    |
|         | HLR    | 0.915                    | 0.916         | 21                    |
|         | CHR    | 0.943                    | 0.930         | 22                    |
|         | MA-CHR | 0.975                    | 0.941         | 20                    |
| Prostate | GA     | 0.914                    | 0.873         | 49                    |
|         | GP     | 0.925                    | 0.884         | 43                    |
|         | MA     | 0.942                    | 0.893         | 42                    |
|         | SCAD-L$_2$ | 0.958                   | 0.927         | 34                    |
|         | HLR    | 0.963                    | 0.912         | 32                    |
|         | CHR    | 0.972                    | 0.913         | 31                    |
|         | MA-CHR | 0.986                    | 0.934         | 29                    |

**FIGURE 2.** Maximum integrative network view of the genes is selected by different methods in AML dataset. The nodes with a thick border indicate the identified genes by the classifier. And the rest of the nodes indicate the genes that are frequently altered and are known to interact with the selected genes by the classifier. The orange hexagons indicate FDA-approved drugs.

**TABLE 5.** The most frequently selected 10 genes found by the seven variable selection approaches from the AML dataset.

| Rank | GA     | GP     | MA     | SCAD-L$_2$ | HLR    | CHR    | MA-CHR |
|------|--------|--------|--------|------------|--------|--------|--------|
| 1    | TNCXEB | ALR    | PTPN6 | GSK3A1     | CAV11  | SFRP2 | CAV11  |
| 2    | TP73   | PNP1   | MEI1N | CDK2NA     | RASA2  | TP73  | CDK2NA |
| 3    | GLIPR1 | SFRP5  | ENZ    | CDK2NA     | HHT1   | CDK2N | HHT1   |
| 4    | SFRP1  | PTPN6  | CDK2NA | CDK2N     | CDK2N  | CD2N  | CD2N   |
| 5    | PTPN6  | FFRP5  | FFRP5 | CDK2N     | CDK2N  | CD2N  | CD2N   |
| 6    | PHDX2  | A4AGLT | CAV11  | CDK2N     | DNM11  | MEI1N | SFRP1  |
| 7    | VEGFA  | DAPKI  | DNM1   | DNM11     | HOX9   | CUX3  | CUX3   |
| 8    | FHT    | PROX2  | JUNB   | PTPN6     | WNT5A  | SFRP5 | ALOU12 |
| 9    | GKI    | CDK2N  | PNL12  | GSTM1     | PNP1   | SFRP1 | GKI    |
| 10   | ARSKA2 | HHT    | SFRP5  | ABCA8     | GSTM1  | SFRP1 | GLIPR1 |
|      |        |        |        |           |        |        |        |

VOLUME 8, 2020 122441
V. CONCLUSIONS
The primary purpose of this paper is trying to propose a new memetic algorithm MA-CHR. This method combines an evolutionary optimization method with a complex harmonic regularization method to deal with high-dimensional and small-sample gene expression data. This paper first uses the evolutionary computation method to globally optimize the hyper-parameters in the non-convex machine learning model, and robust selection of related variable sets can be achieved by using complex harmonic regularization methods. Next, build a classification model. The effective path search algorithm is used to optimize complex harmonic regularization in order to generate a global optimal of the regularized algorithm is used to optimize complex harmonic regularization, model, and robust selection of related variable sets can be achieved by using complex harmonic regularization methods. Next, build a classification model.

The experimental results show that the MA-CHR model proposed in this paper can use more suitable sparse and group regularization methods to learn machine learning models more accurately. In addition, the reliability and stability of the method are relatively high. We believe that this method can be an effective tool for analysing gene expression data by means of adaptive sparse and group regularization.

REFERENCES
[1] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. Roy. Stat. Soc. B, Methodol., vol. 58, no. 1, pp. 267–288, Jan. 1996.
[2] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” J. Amer. Stat. Assoc., vol. 96, no. 456, pp. 1348–1360, Dec. 2001.
[3] H. H. Zhang and W. Lu, “Adaptive lasso for Cox’s proportional hazards model,” Biometrika, vol. 94, no. 3, pp. 691–703, Aug. 2007.
[4] C.-H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” Ann. Statist., vol. 38, no. 2, pp. 894–942, 2010.
[5] X. He, W. Lu, and T. Chen, “A note on adaptive lp regularization,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jun. 2012.
[6] T. Zhang, “Analysis of multi-stage convex relaxation for sparse regularization,” J. Mach. Learn. Res., vol. 11, pp. 1081–1107 and 523, Mar. 2010.
[7] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, “L1/2 regularization,” Sci. China Inf. Sci., vol. 53, no. 6, pp. 1159–1169, 2010.
[8] Z. Xu, X. Chang, F. Xu, and H. Zhang, “L1/2 regularization: A thresholding representation theory and a fast solver,” IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 7, pp. 1013–1027, Jul. 2012.
[9] G.-J. Chu, Y. Liang, and J.-X. Wang, “Novel harmonic regularization approach for variable selection in Cox’s proportional hazards model,” Comput. Math. Methods Med., vol. 2014, Nov. 2014, Art. no. 857398.
[10] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” J. Roy. Stat. Soc., B Stat. Methodol., vol. 66, no. 1, pp. 49–67, Feb. 2006.
[11] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” J. Roy. Stat. Soc., B Stat. Methodol., vol. 67, no. 2, pp. 302–318, Apr. 2005.
[12] R. Tibshirani, I. Johnstone, T. Hastie, and B. Efron, “Least angle regression,” Ann. Statist., vol. 32, no. 2, pp. 407–499, Apr. 2004.
[13] J. Fan and R. Li, “Variable selection for cox’s proportional hazards model and frailty model,” Ann. Statist., vol. 30, pp. 74–99, 2002.
[14] H. Zou and H. H. Zhang, “On the adaptive elastic-net with a diverging number of parameters,” Ann. Statist., vol. 37, no. 4, pp. 1733–1751, Aug. 2009.
[15] L. Zeng and J. Xie, “Group variable selection via SCAD-L2,” Statistics, vol. 48, no. 1, pp. 49–66, Jan. 2014.
[16] H.-H. Huang, X.-Y. Liu, and Y. Liang, “Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2/2 regular- ization,” PLoS ONE, vol. 11, no. 5, May 2016, Art. no. e0149675.
[17] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A survey on evolutionary computation approaches to feature selection,” IEEE Trans. Evol. Comput., vol. 20, no. 4, pp. 606–626, Aug. 2016.
L. Bullinger, K. Döhner, E. Bair, S. Fröhling, R. F. Schlenk, R. Tibshirani, Y. Wen and H. Xu, “A cooperative coevolution-based pittsburgh learning classifier system embedded with memetic feature selection,” in Proc. IEEE Cong. Evol. Comput. (CEC), Jun. 2011, pp. 2415–2422.

D. Meng, Q. Zhao, and L. Jiang, “What objective does self-paced learning indeed optimize?” 2015, arXiv:1511.06049. [Online]. Available: https://arxiv.org/abs/1511.06049

X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang, and H.-S. Ye, “A hybrid genetic algorithm with wrapper-embedded approaches for feature selection,” IEEE Access, vol. 6, pp. 22863–22874, 2018.

Y. Liang, C. Liu, X. Z. Luan, K. S. Leung, T. M. Chan, Z. B. Xu, and H. Zhang, “Sparse logistic regression with a L 1/2 penalty for gene selection in cancer classification,” BMC Bioinf., vol. 14, no. 1, 2013, Art. no. 198.

R. N. Khushaba, A. Al-Ani, and A. Al-Jumaily, “Feature subset selection using differential evolution and a statistical repair mechanism,” Expert Syst. Appl., vol. 38, no. 9, pp. 11515–11526, Sep. 2011, doi: 10.1016/j.eswa.2011.03.028.

K. Holm, C. Hegardt, J. Staafl, J. Vallon-Christersson, G. Jonsson, H. Olsson, Å. Borg, and M. Ringnér, “Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns,” Breast Cancer Res., vol. 12, no. 3, p. R36, Jun. 2010.

A. Villanueva, Y. Hoshida, C. Battiston, V. Tovar, S. Sia, and C. Alsinet, “Combining clinical, pathology, and gene expression data to predict recurrence of hepatocellular carcinoma,” Gastroentenology, vol. 140, no. 5, pp. 1501-1512, 2011.

D. L. Donoho and X. Huo, “Uncertainty principles and ideal atomic decomposition,” IEEE Trans. Inf. Theory, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

G. H. Golub and C. Van Loan, “An analysis of the total least squares problem,” SIAM J. Numer. Anal., vol. 17, no. 1, pp. 88–893, 1980.

P. Zhao and B. Yu, “Stagewise Lasso,” J. Mach. Learn. Res., vol. 8, pp. 2701–2726 and 564–565, Dec. 2007. 565.

S. Rosset and J. Zhu, “Piecewise linear regularized solution paths,” Ann. Statist., vol. 35, no. 3, pp. 1012–1030, Jul. 2007.

Y. Chen, Z. Jia, D. Mercola, and X. Xie, “A gradient boosting algorithm for survival analysis via direct optimization of concordance index,” Comput. Math. Methods Med., vol. 2013, Oct. 2013, Art. no. 873595.

J. H. Friedman, “Fast sparse regression and classification,” Int. J. Forecasting, vol. 28, no. 3, pp. 722–738, Jul. 2012.

L. Bülling, K. Döhner, E. Bair, S. Fröhling, R. F. Schlenk, R. Tibshirani, H. Döhner, and J. R. Pollack, “Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia,” New England J. Med., vol. 350, no. 16, pp. 1605–1616, Apr. 2004.

A. Rosenwald et al., “The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma,” New England J. Med., vol. 346, no. 25, pp. 1937–1947, 2002.

D. Singh, P. G. Febo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D’Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, “Gene expression correlates of clinical prostate cancer behavior,” Cancer Cell, vol. 1, no. 2, pp. 203–209, Mar. 2002.

M. A. Ship, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, “Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning,” Nature Med., vol. 8, no. 1, pp. 68–74, Jan. 2002.

J.-S. Seo, Y. S. Ju, W.-C. Lee, J.-Y. Shin, J. K. Lee, T. Bleazard, J. Lee, Y. J. Jung, J.-O. Kim, J.-Y. Shin, S.-B. Yu, J. Kim, E.-R. Lee, C.-H. Kang, I.-K. Park, H. Rhee, S.-H. Lee, J.-I. Kim, J.-H. Kang, and Y. T. Kim, “The transcriptional landscape and mutational profile of lung adenocarcinoma,” Genome Res., vol. 22, no. 11, pp. 2109–2119, Nov. 2012.

B. A. Boone, S. Sabbaghian, M. Zenati, J. W. Marsh, A. J. Moser, A. H. Zareikat, A. D. Singh, H. J. Zeh, and A. M. Krasinskas, “Loss of SMAD4 staining in pre-operative cell blocks is associated with distant metastases following pancreaticoduodenectomy with venous resection for pancreatic cancer,” J. Surgical Oncol., vol. 110, no. 2, pp. 171–175, Aug. 2014.

H. Vuong, F. Cheng, C.-C. Lin, and Z. Zhao, “Functional consequences of somatic mutations in cancer using protein pocket-based prioritization approach,” Genome Med., vol. 6, no. 10, p. 81, Dec. 2014.

J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal,” Sci. Signaling, vol. 6, no. 269, pp. pH1–pH2, Apr. 2013.