Feature selection in high dimensional regression problems for genomic

Julie Hamon, Clarisse Dhaenens, Gaël Even, Julien Jacques

Feature selection in high dimensional regression problems for genomics

Julie Hamon[1,2,3], Clarisse Dhaenens[1,2], Gaël Even[3], and Julien Jacques[1,4]

[1] Inria Lille - Nord Europe, France
julie.hamon@inria.fr
[2] LIFL / Université Lille 1, France
clarisse.dhaenens@lifl.fr
[3] Gènes Diffusion 3595 Route de Tournai, Douai, France
g.even@genesdiffusion.com
[4] Laboratoire Paul Painlevé / CNRS & Université Lille 1, France
julien.jacques@lifl.fr

Abstract. In the context of genomic selection in animal breeding, an important objective consists in looking for explicative markers for a phenotype under study. In order to deal with a high number of markers, we propose to use combinatorial optimization to perform variable selection. Results show that our approach outperforms some classical and widely used methods on simulated and "closed to real" datasets.

Keywords: Feature selection, combinatorial optimization, regression, genomic.

1 Introduction

Genomic selection of animal breeding deals with a genetic evaluation of animals from their DNA (extracted using biological samples such as blood or hairs, or biopsy), based on a huge number of markers covering the whole genome. The basic principle was established by Meuwissen, Hayes and Goddard in 2001 [18]. This approach has become feasible thanks to the large number of single nucleotide polymorphisms (SNPs) that can be genotyped for a reasonable price. Hence, with the development of new technologies such as high-throughput genotyping and sequencing, it is possible to conduct such studies and read genomic information on around 800,000 markers on more and more subjects. In this context, an important objective of genomic selection in animal breeding consists in looking for explicative markers for a phenotype (quantitative trait characterizing an animal) under study. Such quantitative trait locus associated with the phenotype is detected thanks to adjacent genotyped markers in linkage disequilibrium. In their editorial of the special issue of *Animal frontiers* on *application of genomic tools in different livestock species*, Bagnato and Rosati [1] explain the importance of genomic in animal selection. They indicate, for example, that "genomic information may reduce costs and accelerate genetic gain by reducing generation intervals". Moreover, they also indicate that "genomic selection may

allow the identification of superior individuals for traits not currently considered in animal breeding plans because of technical difficulties". All these reasons lead genomic selection to be a real challenge for industry.
One of the important insight for this domain is to establish predictive models using genomic information. However, in addition to biological constraints (such as sample storage, time consuming and costly experiments, etc.) data analysis needs to be improved and original methods have to be proposed to take into account all the specificities of these data. Up to now, significant marker identification studies have mostly been proposed for qualitative traits, often binaries (disease or not) [12, 16]. SNP markers have also been used for other purposes such as animals identification, for example in Heaton *et al.* [10].

The challenge of this work is to find a predictive model based on a reasonable number of markers allowing to select the best animals for a given phenotype, in order to produce small size chips for the phenotype (trait) under study. We propose to deal with this problem of markers selection with a combinatorial optimization approach. In this combinatorial optimization context, one of the specificities of this work is to deal with quantitative traits such as milk production or meat quality.

The remaining of the paper is as follows: In the next section, we introduce genotyping data and their specificities. In section 3 we model the problem and introduce some classical statistical approaches and their limits. In section 4, our approach based on a cooperation between statistics and combinatorial optimization is exposed. Section 5 is dedicated to the validation on simulated data, and "closed to real" data from the literature, and to the comparison with classical approaches. We conclude in section 6.

2 Data

Markers used in our study are single nucleotide polymorphisms (SNPs) which are DNA sequence variations at a single base pair location. To have a complete explanation of the use of SNPs as genetic markers, the reader may refer to Vignal *et al.* [24].

Let us give the main principle: Figure 1 shows an extract of DNA of two subjects. The second subject differs from the first one by a single nucleotide (polymorphism C/T). Hence, if a marker is located at this place, the difference between the two subjects will be put in light. Each subject will be genotyped using a 54K chip, meaning that each animal will be described according to the 54,609 SNPs: At each position, its sequence will be known in the form AT, AA, TT, CG, etc. Data is then a matrix subjects $\times$ SNPs. In order to be analysed, this complex matrix is usually recoded either in {1, 0, -1} as in Ogutu *et al.* [19] or in {0, 1, 2} as in Usai *et al.* [23], representing homozygous major, heterozygous and homozygous minor respectively. In this work, we decide to use the encoding {0, 1, 2} as it is usually used in Gènes Diffusion, the company we are working with. Typical data that has to be analyzed is composed of around $n = 1,000$ to $3,000$ animals and $p = 40,000$ markers (after filtering on the $54,609$ ones genotyped).

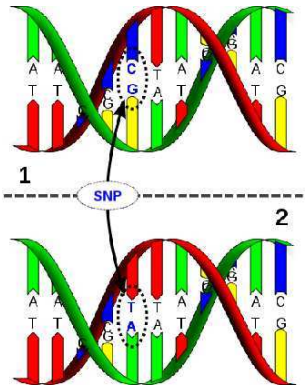

Fig. 1. Single nucleotide polymorphism [5]

Associated with this matrix, for each subject the value of a quantitative trait (phenotype) is indicated. The objective is then to find a model based on a reasonable number of markers, able to predict, for a new animal, a quantitative trait from this large set of quantitative features.

3 A high-dimensional regression problem

The problem of predicting a phenotype value from genomic information may be modeled as a regression problem as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} (\beta_j x_{ij}) + \epsilon_i, \tag{1}$$

where y_i is the trait of interest ($y_i \in \mathbb{R}$), x_{ij} are the studied SNPs (in $\{0, 1, 2\}$), and ϵ_i are uncorrelated Gaussian residuals with zero mean and variance σ^2.

The objective is to estimate the parameters β_j and the most popular method, assuming that observations are independent, is *ordinary least squares* (OLS) which consists in minimizing the residual sum of squares (RSS):

$$\beta^{OLS} = argmin_\beta \{RSS(\beta)\} \tag{2}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ and $RSS(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$.

The usual solution of this problem $\hat{\beta}^{OLS} = (X^t X)^{-1} X^t \mathrm{y}$ is intractable when $n \ll p$, as in typical problems we are interested in (problems where about $n = 1,000$ to $3,000$ animals are described by about $p = 40,000$ markers).

[5] Figure from `http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism`

To overcome this problem, several statistical methods have been introduced along the years (see [9] for a review). The common idea is to reduce the number of variables, either by defining new variables as combination of initial ones (partial least squares [14] and principal components regression [5]), or by selecting the most important variables with an iterative procedure (stepwise selection).
More recently, new approaches have been introduced by *shrinkage methods*, as the well known *lasso* [22] and *ridge* [11] regression, which both introduce a penalty in the OLS problem (2)

$$\beta = argmin\{RSS(\beta) + \lambda||\beta||\} \tag{3}$$

where $||.||$ is the L^1-norm for the lasso and the L^2-norm for ridge. If ridge is particularly effective to deal with correlated covariates, lasso has the advantage to perform variable selection. In order to combine the advantage of the lasso and ridge methods, Zou and Hastie [25] have proposed a regularization method called the elastic net (EN), combining the penalties of the lasso and ridge. These methods are considered as state-of-the-art for the problem of regression in high dimension (see for instance [19]).

On the other hand, optimization methods can be efficient ways to carry out variable selection for regression problems. For instance, [17] consider simulated annealing to proceed to variables selection in marketing applications, with an evaluation of the model adequacy by the mean of the AIC criterion [2]. [13] considers additionally a genetic algorithm and the BIC criterion [20]. In [8] an iterative local search is considered in a gene expression context and the models are evaluated thanks to their posterior probabilities. Following these successes we propose to adopt such type of approaches for the context of regression in high dimension.

4 Proposed approach

In this section we present the approach we propose for the prediction of a quantitative phenotype using a reasonable number of SNPs.

4.1 The statistical model

As the model exposed in (1) is intractable due to the number of variables (p) which is larger than the number of subjects (n), we introduce in this model additional binary parameters z_j indicating whether the SNP j is selected or not:

$$y_i = \beta_0 + \sum_{j=1}^{p} (\beta_j z_j x_{ij}) + \epsilon_i. \tag{4}$$

The objective here is to estimate the parameters $\beta = (\beta_0, \ldots, \beta_p)$ and $z = (z_1, \ldots, z_p)$. As z is a discrete parameter belonging to $\{0, 1\}^p$, determining the z_j values is equivalent to determining variables that participate to the regression model. This problem is a typical feature selection problem, which can be seen as a combinatorial problem. Hence it can be addressed by combinatorial optimization methods, and we choose an iterated local search as described above.

4.2 Iterated local search principle

As indicated in the survey proposed by Corne *et al.* [4], optimization methods are an efficient way to deal with feature selection problems. In order to tackle with a high number of features, metaheuristics (local search, evolutionary algorithm, etc.) have proved their efficiency. In our context of feature selection for regression, we propose to use an iterated local search (ILS), in which the solution evaluation will be performed with classical model selection criteria for multivariate regression.

One of the fastest local search is the descent method, as it only works on the current solution and its neighbors. A classical descent method optimizing function f is described by Algorithm 1. Such a method starts from an initial solution and replaces it by an improving neighbor. This method stops when a local optimum (a solution that does not have any improving neighbor) is found.

Algorithm 1 Descent method

$z = z_0$ /* Generate initial solution */
while has.neighbor(z) **do**
  Choose $z^{'}$ in $N(z)$ /* $z^{'}$ in the neighborhood */
  if $f(z^{'})$ is better than $f(z)$ **then**
    $z = z'$ /* Replace current solution by its neighbor */
  end if
end while
return Final solution found (local optima)

An important limit of this algorithm, is that the quality of the local optimum obtained may depend on the initial solution. To overcome this problem, the answer is to continue the search after a local optimum is reached. That is why it is embedded in the iterative scheme of an ILS.
An ILS method is based on a succession of local searches and perturbations (to have a deeper explanation of such a method the reader may refer to [21]). The main principles are described on Figure 2: starting from a given initial solution, the ILS applies a descent method. When a local optimum is reached, it is perturbed and the local search is restarted from this perturbed solution, until a stopping criterion is reached.

4.3 Iterated local search components

Performing such an ILS requires to define several components such as the representation of a solution, the initialization, the neighborhood relation... This may depend on the problem to be solved [21]. Hence we give below information about the choices we made for the problem under study.

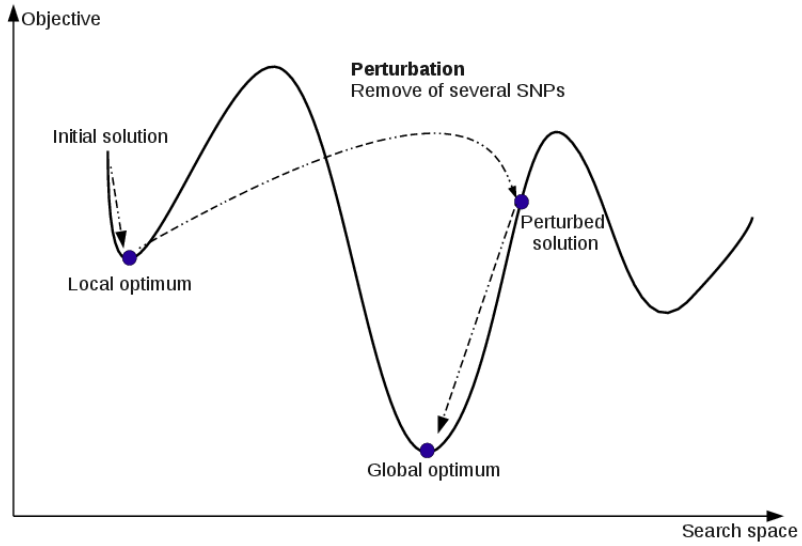

Fig. 2. Iterated local search

– *Encoding:* The encoding of a solution plays a major role in the efficiency and effectiveness of a metaheuristic as it influences the choices of operators and evaluation function. Several encodings could be used for the feature selection problem: Binary vector indicating selected features, real vector indicating the weight of each feature (β_j), list of features selected, etc. We choose to use a binary vector indicating whether a feature is selected (1) or not (0) as it is very close to the statistical model presented in the previous part (is equivalent to the z vector). Moreover, this encoding allows to design a simple but efficient neighborhood.

– *Initialization:* The initial solution which corresponds to a first subset of selected features, is done randomly but driven by the correlation of the features with the trait to explain. Therefore, a roulette wheel selection (random selection according to a multinomial distribution) is proposed. It assigns to each feature a selection probability that is proportional to its relative correlation with the trait. Such an initialization has not much influence on the quality of results obtained, but can speed up the search starting from solution of medium quality.

– *Neighborhood:* The neighborhood function assigns to each solution a set of neighbor solutions obtained by the application of a *move* operator. The neighborhood solution plays also a crucial role in the performance of a local

search, as it will define the set of solutions to explore at each step. The commonly used neighborhood function for binary vectors is the *bit flip* operator. This operator randomly selects a feature and modifies the corresponding bit in the vector. Hence, if this feature is selected in the current solution it will not be selected in its neighbor, otherwise it becomes selected. Common neighborhood exploration strategies are to choose either the best neighbor, or the first one that improves the quality. Choosing the best one, requires to generate at each step the whole neighborhood, that may be time consuming. Hence, as in practice, on many applications, it has been observed that the first improving strategy leads to the same quality of solutions than the best improving strategy, we choose to use the first improvement strategy.

- *Perturbation:* The key idea of an ILS is that the perturbation method should be more effective than a random restart approach. Therefore, the perturbation is often based on the neighborhood function and may consist in several applications of it. Hence we propose to perturb the solution by flipping several features. The perturbed solution is accepted even if it does not improve the quality of the current solution.

- *Stopping criterion:* As this method is iterative, it does not stop by itself. Hence, a stopping criterion has to be defined. Here we propose to let the method converge, but stopping it when it has not produced any improvement on the best solution for a given number of iterations or when a maximum number of evaluations is reached.

- *Optimization function:* In order to compute the quality of a solution (subset of features), we calculate the prediction error of the regression model defined on this set of features using statistical criteria. This is the aim of next part.

Let us remark that because of their popularity, several authors such as Kapetanios *et al.* [13] or Meiri *et al.* [17] proposed to use a simulated annealing or a genetic algorithm to deal with the variable selection problem indifferent types of application. However, as the ILS is simplest, it uses less components and parameters than other metaheuristic approaches. This allows to better evaluate the relevance of using a combinatorial approach combined with several evaluation functions. This is the first step before designing more sophisticated approaches.

4.4 Fitness of a solution

The aim of the optimization method is to explore efficiently the large search space of solutions corresponding here in all the possible feature subsets. Therefore such a method uses an evaluation criterion (fitness function) able to associate to each solution a quality measure that represents, in fine, the goal to achieve. In the present context, if the goal is clearly to identify the best subset of features, that is the one that will produce the best predictive model, computing the quality of

such a subset is not straightforward and should be discussed. In order to compute the quality of a solution (subset of features), we calculate the prediction error of the regression model defined thanks to equation (4) where vector z describes features belonging to the solution. One difficulty well known in Datamining is to be able to assess the quality of the model on data that has not been used for the computation of this model (validation set).
Therefore, in an earlier work [7], several approaches have been compared on the basis of simulation studies to evaluate their ability to estimate this prediction error: BIC [9], k-fold cross-validation and leave-one-out cross-validation. Following this study, the retained approach is 3-fold cross-validation.

5 Experimental results

Our model is validated using two simulation studies. In the first one a high-dimensional regression model is considered, whereas the second one is a "closed to real" data simulation used as challenging dataset in the XV^{th} QTLMAS workshop [6]. Application on real data will be presented at the conference, using datasets from our company Gènes Diffusion which is specialized in genetic and animal reproduction.

- Simulation 1
  Datasets with $p = 1,000$ features are generated. We generate $X \sim N(0, \Sigma)$, with Σ such that $Cov(X_i, X_j) = 0.5^{|i-j|}$, and then compute y $= \beta X + \epsilon$ where $\beta \sim N(0, 1)$ for 50% of variables, $\beta \sim N(5, 1)$ for 25% of variables and $\beta \sim N(-5, 1)$ for the remaining 25%. The model error ϵ is assumed to be Gaussian $\epsilon \sim N(0, 1)$.
  In order to assess the quality of solutions found by the method, training and validation datasets are generated:
  - 20 training datasets with $n = 100$ subjects (recall $p = 1,000$ features).
  - A validation dataset with $n = 1,000$ subjects to fairly evaluate the model on non previously seen subjects.

- XV^{th} QTLMAS dataset
  This dataset is composed of a total of 3,000 individuals genotyped for 9,990 SNPs. They are separated in a training dataset with 2,000 subjects and a validation dataset with 1,000 subjects. The trait studied y is a quantitative phenotype. A preprocessing is used to remove SNPs that have a same value for all individuals leading to 7,121 SNPs to study.

The algorithm stops when it reaches 1,000 iterations (100 for QTLMAS) without improving the best solution or $1,000,000$ evaluations ($50,000$ for QTLMAS). For QTLMAS, we choose lower values of stopping criteria in order to keep a reasonable runtime (10 minutes) compared to classical methods.

For each dataset, the proposed method is compared to ridge, lasso and elastic net regressions as well as stepwise selection as these methods are widely used in this context of genomic selection for regression. As the proposed method tends

to select as many variables as the number of individuals leading to overfitting, we decided to fix a maximum number of variables regarding how much classical methods select. So we fix 30 variables at most for simulated data and 50 for QTLMAS.

This method is implemented in C++ and uses the implementation of the ILS algorithm available in the paradiseEO plateform [3]. Others methods are computed using **R** with the package *lars* for stepwise, the *lm.ridge* procedure for ridge and the *glmnet* procedure for lasso and EN. We select the shrinkage parameters λ of lasso and EN using the *cv.glmnet* procedure. In addition, we set the α parameter of EN (which can take values in $[0:1]$) by 10-fold cross-validation (as recommended).

5.1 Results and discussion

As the proposed approach is stochastic, to attest its performance, we perform 30 runs on each datasets and report, within a boxplot, obtained results.

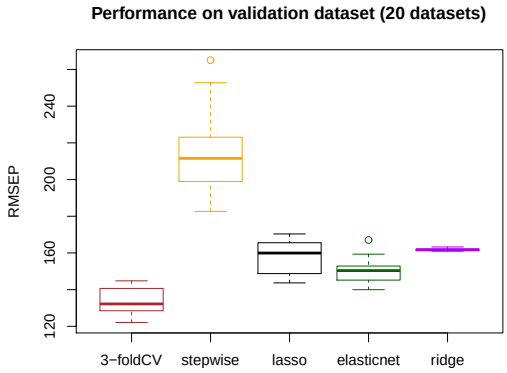

Fig. 3. Comparison with classical statistical approaches: Simulation 1

Figure 3 plots the performance (on validation dataset) of our method in term of root mean square error of prediction (RMSEP), and compare it with classical approaches on the 20 datasets of data from Simulation 1.

Figure 4 shows results on the QTLMAS dataset, in order to analyze how the method behaves with "closed to real" data.

Results show first that, lasso, elastic net and ridge are efficient methods compared to stepwise. Moreover, results also show that the proposed approach

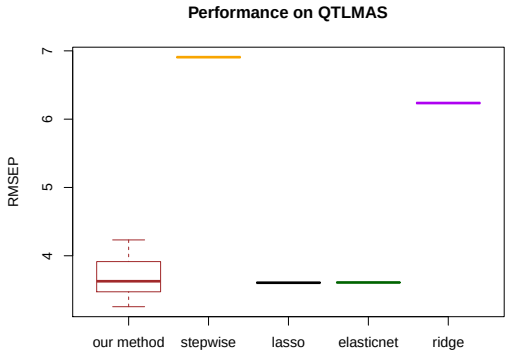

Fig. 4. Comparison with classical statistical approaches: QTLMAS dataset

manages to outperform these state-of-the-art methods for almost every runs on datasets from Simulation 1. Regarding, the QTLMAS dataset our method obtains results as good as the most efficient methods in average, and manages to outperform them for some runs.

To complete this comparison, we may refer to the article [15] which reports best results obtained during the challenge QTLMAS. Around 15 approaches are compared there, and the results obtained with our method belong to the best ones.

6 Conclusion

In this article, we addressed the problem of selecting the subset of relevant variables that will allow to obtain the best predictive model for a quantitative trait to explain. We are in a high-dimensional regression problem ($n \ll p$), that makes the problem more difficult as the risk of overfitting is high. We propose to combine a combinatorial optimization algorithm with a statistical evaluation criterion to address this problem.
The comparison with other classical approaches shows that our method gets good results compared with state-of-the-art approaches used in genomics (lasso, elastic net or ridge).
Moreover, this context of genomic selection in animal studies involves a lot of characteristics such as familial relationship between animals, and an interesting thing to note, is that the scheme proposed is very general, and can fit with many models. In particular, it will be possible to introduce these familial relationship, that have been left out for the moment, by adapting the method probably using a mixed model approach. An other perspective of this work is to accelerate the algorithm by parallelizing the evaluation of solutions as in [8], and to extend

this local search to more sophisticated method (such as genetic algorithms) to be able to explore larger search space, corresponding to a larger number of SNPs.

References

1. A. Bagnato and A. Rosati, *From the Editors - Animal Selection: The genomics revolution.*, Animal Frontiers, vol. 2, n. 1, pp. 1–2, 2012.
2. H. Akaike, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, vol. 19, n. 6, pp. 716–723, 1974.
3. S. Cahon, N. Melab and E-G. Talbi, *Paradiseo: A framework for the reusable design of parallel and distributed metaheuristics.*, Journal of Heuristics, vol 10, n. 3, pp. 357–380, 2004.
4. D. Corne, C. Dhaenens, and L. Jourdan, *Synergies between operations research and data mining: The emerging use of multi-objective approaches.*, European Journal of Operational Research, vol. 221, n. 3, pp. 469–479, 2012.
5. K.L Coxe, N.L. Johnson, S. Kotz, *Principal components regression analysis.* Encyclopedia of Statistical Science. New York: Wiley, pp. 181–184, 1986.
6. JM. Elsen, S. Tesseydre, O. Filangi, P. Le Roy, O. Demeure, *XVth QTLMAS: simulated dataset.*, in BMC Proceedings 2012, 6(Suppl 2), S1.
7. J. Hamon, C. Dhaenens, G. Even and J. Jacques, *Feature selection for high dimensional regression using local search and statistical criteria.*, International Conference on Metaheuristics and Nature Inspired Computing, 2012.
8. C. Hans, A. Dobra and M. West, *Shotgun stochastic search for "large p" regression.*, Journal of the American Statistical Association, vol. 102, n. 478, pp. 507–516, 2007.
9. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning.*, Springer, 2009.
10. M. P. Heaton, G. P. Harhay, G. L. Bennett, R. T. Stone, W. M. Grosse, E. Casas, J. W. Keele, T. P. L. Smith, C. G. Chitko-McKown, and W. W. Laegreid, *Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle.*, Mammalian Genome, vol. 13, n. 5, pp. 272–281, may 2002.
11. A. E. Hoerl and R. Kennard, *Ridge regression: Biaised estimation for nonorthogonal problems.*, Technometrics, vol. 12, pp. 55–67, 1970.
12. L. Jourdan, C. Dhaenens, and E.-G. Talbi, *Linkage disequilibrium study with a parallel adaptive GA.*, International Journal of Foundations of Computer Science, 2004.
13. G. Kapetanios, *Variable selection in regression models using nonstandard optimisation of information criteria.*, Computational Statistics & Data Analysis, vol. 52, pp. 4–15, 2007.
14. K. -A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, *Sparse PLS for variable selection when integrating omics data.*, Stat Appl Genet Mol Biol, vol. 7, Article 35, 2008.
15. P. Le Roy, O. Filangi, O. Demeure, JM Elsen, *Comparison of analyses of the XVth QTLMAS common dataset III: Genomic Estimations of Breeding Values.*, BMC Proceedings 2012, 6(Suppl 2):S3, 2012.
16. Y. Liang and A. Kelemen, *Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases.*, Statistics Surveys, vol. 2, pp. 43–60, 2008.
17. R. Meiri, J. Zahavi, *Using simulated annealing to optimize the feature selection problem in marketing applications.*, European Journal of Operational Research, vol. 171, pp. 842–858, 2006.

18. T. H. E. Meuwissen, B. J. Hayes and M. E. Goddard, *Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.*, Genetics Society of America, 2001.
19. J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho, *Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions.*, BMC Proceedings, vol. 6, Suppl 2, S10, 2012.
20. G. E. Schwarz, *Estimating the dimension of a model.*, Annals of Statistics, vol. 6, n. 2, pp. 461–464, 1978.
21. E.-G. Talbi, editor. *Metaheuristics.*, John Wiley and Sons, Inc., 2009.
22. R. Tibshirani, *Regression Shrinkage and Selection Via the Lasso.*, Journal of the Royal Statistical Society, Series B, vol. 58, pp. 267–288, 1994.
23. M. G. Usai, M. E. Goddard and B. J. Hayes, *LASSO with cross-validation for genomic selection.*, Genet Res (Camb)., vol. 91, n. 6, pp. 427–36, 2009.
24. A. Vignal, D. Milan, M. SanCristobal, and A. Eggen, *A review on SNP and other types of molecular markers and their use in animal genetics.*, Genetics Selection Evolution, vol. 34, n. 3, pp. 275, 2002.
25. H. Zou and T. Hastie, *Regularization and Variable Selection via the Elastic Net.*, Journal of the Royal Statistical Society, Series B, vol. 67, Part 2, pp. 301–320, 2005.