

# Comparative analysis of methods of best regression structure search

Volodymyr Stepashko, Iana Bondarska

**Abstract.** Most critical part of linear regression algorithms is selection of structures to be checked. One way to reduce the number of such structures is to make some primary estimations of structures quality and to divide the structures into groups according to this estimation. One technique that uses this approach is La Motte-Hocking algorithm.

This paper represents research results of La Motte-Hocking method efficiency in comparison with the combinatorial algorithm.

Results of research and some possible ways to improve results are presented in this paper. The outcomes are planned to be used for updating the combinatorial GMDH algorithm.

## Keywords

Linear regression, best subset search methods, La Motte-Hocking method

## 1 Introduction

This paper presents results of comparison between methods of best subset search. Search of best structure is the most critical and time consuming part of modeling algorithm. There are approaches that avoid checking of all models and they are considered below.

Research was done in the next directions:

- comparison between speed of combinatorial and La Motte-Hocking algorithms;
- modification of La Motte-Hocking algorithm to improve speed of method

Results of research and some possible ways to improve results are presented in this paper. The outcomes are planned to be used for updating the combinatorial GMDH algorithm.

## 2 Theoretical Part

### 2.1 Overview of problem of best linear regression subset search

The problem of best linear regression subset search can be described as search for model with minimal value of quality criteria among models from set  $\Omega$

$$d^* = \min_{\Omega_m} (CR(M(d))), \quad (1)$$

where

$m$ -total number of arguments,

$d[1 \times m]$ -structure vector,

$M(d)$  -- partial model  $d$ ,

$CR(M(d))$  -- value of quality criteria for model  $M(d)$ ,

$\Omega_m$  -- set of structures.

Structure vector  $d$  consists of 0 and 1 and defines what arguments have non-zero coefficient in a model.

Partial model does not include all arguments and is defined by structure vector  $d$ , parameter vector  $a$  and value of quality criteria on the available data set. Model which includes all arguments is named full model.

Complexity of model is defined by number of arguments in it.

The simplest method of best subset search is checking of all possible structures, combinatorial algorithm. But it cannot be used for models with great number of arguments. Other approaches are:

- dividing structure set to groups, excluding groups with worse quality criteria value from search (La Motte-Hocking method);
- step-by-step including and excluding regressors from model according to change of quality criteria (stepwise regression).

## 2.2 Overview of La Motte-Hocking method

Method is based on next property of RSS:

$$\begin{aligned} R1, R2 - \text{partial structures} \\ \text{if } R1 \subseteq R2, \text{ then } \text{RSS}_{R1} \geq \text{RSS}_{R2} \end{aligned} \quad (2)$$

It is obvious from (2) that if there is some partial model  $R^*$  and there is some subset of models  $\{R\}_{R^*}$ , created by excluding some additional arguments from  $R^*$ , value of RSS of model  $R^*$  will be lower bound of values of RSS for models from  $\{R\}_{R^*}$ .

In La Motte-Hocking method first step is dividing a whole set of models into groups. Then search is performed in each group separately.

Additionally parameter  $k$  and complexity of model  $s$  are specified. Method consists of following steps:

1. All possible  $k$ -models (structures with  $k$  arguments excluded) are generated. Then they are ordered according to their RSS value (ascending).
2. Subgroup of models is built on the basis of  $k$ -model with best value of RSS. From  $k$ -model additionally ( $m-k-s$ ) arguments are excluded. Search for best model in this group is performed.
3. RSS of best model in the group is compared with RSS of next  $k$ -model. If RSS of model is not greater than RSS of next  $k$ -model, than best model has been found. If not, steps 2,3 are repeated for next  $k$ -model.

RSS is used as quality criteria. Most of criterias have following structure:

$$QR = g(s)RSS + f(s), \quad (3)$$

where

$s$  – complexity of model,  $g(s), f(s)$  – some strictly increasing functions.

If complexity of models is fixed, model with best value of RSS will be best model for any quality criteria that has structure given in (3).

Method's advantages:

- result of method is same as for combinatorial one;
- method generates much less models than combinatorial one.

Disadvantages:

- slow performance for small set of arguments;
- absence of procedure for defining parameter  $k$ ;
- impossible to estimate number of structures that will be generated.

## 2.3 Ways for enhancing La Motte-Hocking method

For great complexities method slows down quickly. There are two main reasons for this:

- increasing of number of  $k$ -models that should be built and estimated;
- increasing number of models in each group.

Possible ways for enhancing La Motte-Hocking method:

1. do not build all k-models, only some number of best k-models (proposed by method authors);
2. do not check all models in group, use more efficient methods of search.

The main disadvantage of first approach is probability that search should be performed in all groups defined by k-models. Another problem is absence of procedure for defining parameter k for each stage of method.

Second approach is using more efficient methods of search in each group, e.g. stepwise regression, backward regression. Backward regression is step-by-step excluding arguments from model Those arguments excluding of which leads to least increasing of quality criterion are excluded. The problem of such approach is that results of backward regression not always same as results of full check, especially for small complexities. That is why for small complexities this method is not accurate.

Second approach was implemented and investigated in current article.

### 3 Research Results

#### 3.1 Goal and Subject of Research

The goal of current research is comparison between different methods of best subset search. Subjects of comparison are speed of method and accuracy of results. Next methods were investigated: combinatorial method, La Motte-Hocking method, modification of La Motte-Hocking method, stepwise regression.

Parameters of input sample:

- sample size 25;
- level of noise 50% of maximal value of output parameter.

Value of parameter k for La Motte-Hocking method and its modification is 3. Maximal complexity of model that can be found using these methods is (m-k-1).

#### 3.2 Comparison of Methods Accuracy

Accuracy of method can be defined as level of difference between result of this method and result of combinatorial

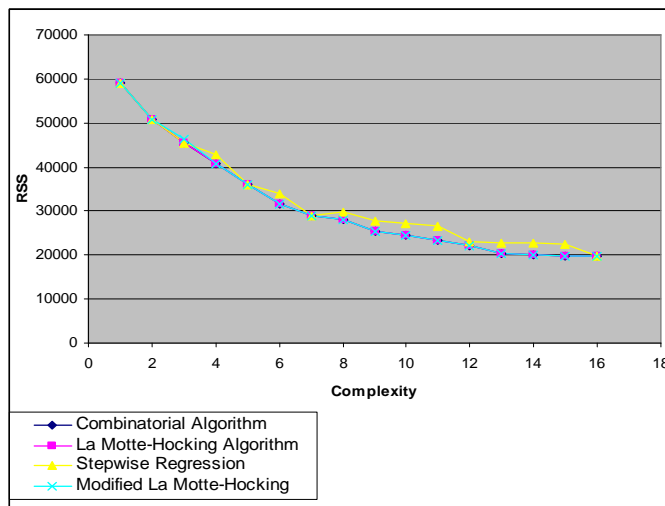


Fig. 1 Comparison of accuracy for best subset search on fixed levels of complexity

method. Difference can be measured as number of different arguments in structures or difference between values of quality criterias, e.g. RSS.

It's better to compare accuracy of result for each level of complexity separately. Since best model for whole set of structures can be found by full check among best models for each level of complexity, if method is accurate on each level of complexity, it's accurate on the whole set of structures.

As it seen from figure 1, for small complexities both modification of La Motte-Hocking method and stepwise regression are not accurate. Accuracy of La Motte-Hocking method can be improved by fitting value of parameter k. Possible algorithm for choosing value of k:

1. Optimal value of parameter

$$k = ((m - s) / 2) \pm 2 \quad (4)$$

2. If number of structures of complexity  $k \geq 2^{20} - 1$ , method will not be efficient
3. Value of parameter  $k$  is chosen according to formula (4)

From step 3 of above algorithm it's clear that method can't be efficient for small complexities of models (e.g if there is need to find structure of complexity 3 for total number of arguments 20, combinatorial algorithm will check 1140 structures. La Motte-Hocking method will first build all possible structures of  $(20-3)/2=8$  -- 125 970 structures ).

### 3.3 Comparison of numbers of checked models

Number of checked models is main criteria of method's speed. For combinatorial method number of checked

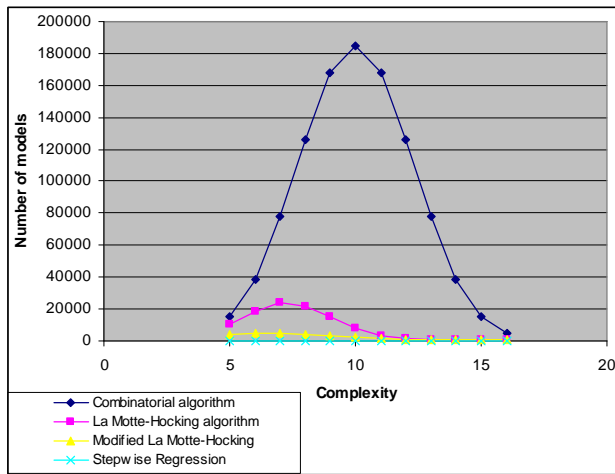


Fig. 2 Number of checked structures on fixed levels of complexity

models for one level of complexity can be found using following formula:

$$C_m^s = \frac{m!}{(m-s)!s!}, \quad (5)$$

where  $m$ -total number of arguments,  $s$ -complexity of model.

Maximal value of  $C_m^s$  is reached for  $s=m/2$ . For La Motte-Hocking method and its modification number of checked structures is not stable and cannot be estimated. Figure 2 presents number of checked structures for all methods on fixed levels of complexity.

Modification of La Motte-Hocking method is more efficient than standard La Motte-Hocking method for  $s$  near to  $m/2$ . Results for check of all structures from complexity 1 to maximal possible complexity  $(m-s-1)$  are given on figure 3. They show that for full check modification of La Motte-Hocking method is not more efficient than standard La Motte-Hocking method. But for large complexities there is some economy.

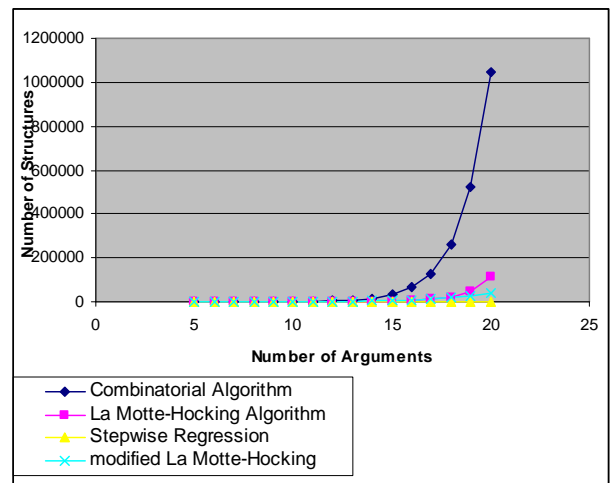


Fig. 3. Number of checked structures for full check

### 3.4 Comparison of Running Time of Methods

Running time of methods is also very important indicator of method speed, it takes into account all additional calculations done in method. Figures 4,5 show comparison between running time of investigated methods. Graphics are proportional to number of checked structures and one can make conclusion that methods are not slowed down because of large number of additional calculations. However calculations in modified La Motte-Hocking method should be optimized.

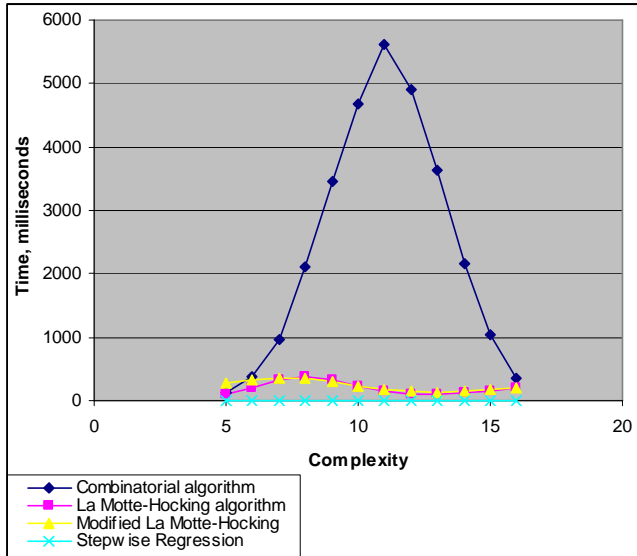


Fig. 4 Comparison of running time for fixed levels of complexity

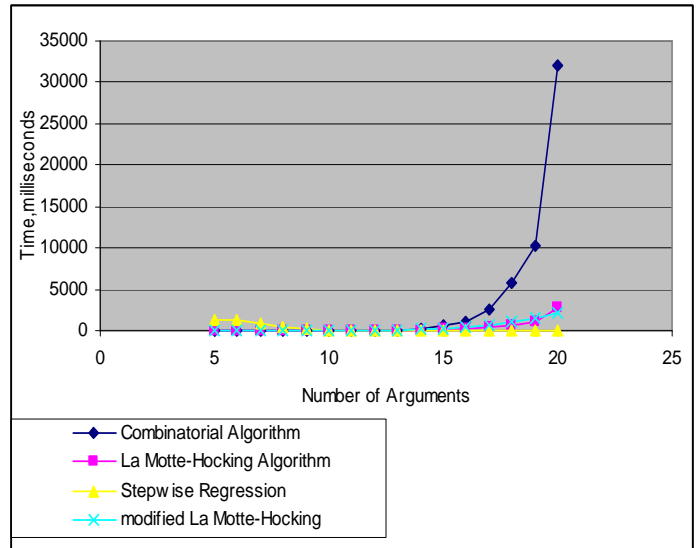


Fig. 5 Comparison of running time for full check

### 3.5 Comparison of number of structures for different values of parameter s

Efficiency of La Motte-Hocking method depends on values of parameter k. If great value of parameter k is specified, search within each group will be done very quickly but number of k-models that should be generated and estimated, will increase greatly. On the other hand for small values of k first stage of method when k-models are generated will be done quickly but second part, search in k-groups, will be very slow.

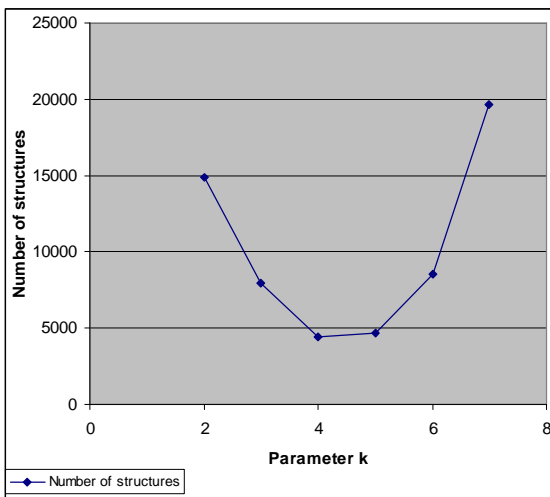


Fig. 6 Comparison of models number for different values of parameter k. Complexity of model 10, total number of arguments 20

Another point that should be considered when parameter k is defined is that to finish search, k-model should be compared with model of much smaller complexity. So, when difference between s and m-k is great, it will be hard to find group of complexity k that has RSS value smaller than RSS value of group of complexity m-k (e.g. when models of complexity 17 and 5 are compared).

Figure 6 introduces dependency between different values of parameter k and number of models that is checked. It's obvious that optimal value of parameter k is near  $(m-s)/2$

### 3.6 Investigation of efficiency of La Motte-Hocking algorithms and it's modification for different number of arguments.

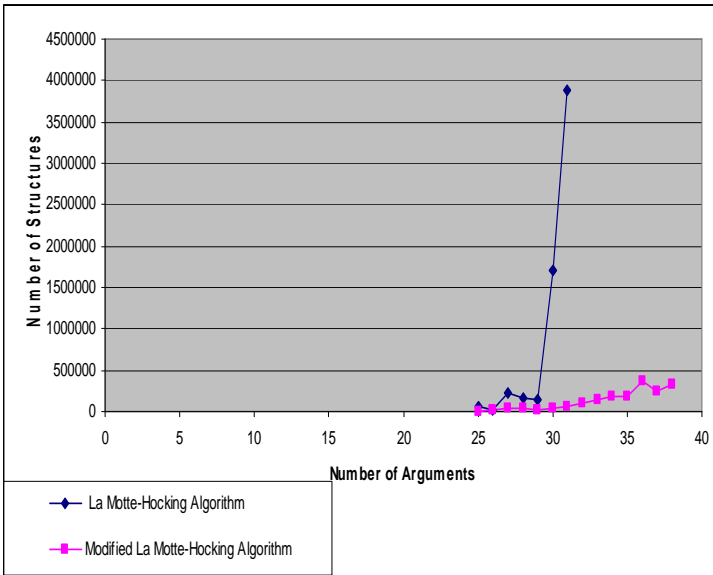


Fig. 7 Comparison of efficiency of La Motte-Hocking method and its modification

Previous research presents some findings according algorithm's accuracy. Now it is also important to check how La Motte-Hocking method and its modification work for large complexities, above 20.

Investigation of efficiency was done in next way: maximal number of structures is for  $s=m/2$ . Both algorithms found models of such complexity for different values of  $m$ . Method considered ineffective if number of structures to be checked  $S \geq C_{20}^{10} = 184756$ .

Results of investigation are presented on figure 7

Number of structures that is checked in modified method of La Motte-Hocking is much more less than in standard method. Also increasing of number of structures is very slow, so method can be considered efficient for complexities near to 35.

## 4 Conclusions

Results of investigation of method for best subset search were introduced in article. Detailed results for each method are below.

La Motte-Hocking method. Method is accurate, result is always same as for full check, but for small complexities it is not as efficient as full check method. Method is efficient for complexities near 30 arguments. Reason for slowing down is great increasing of number of k-models.

Modification of La Motte-Hocking method. Method not always gives same result as full check. For improving accuracy of method it's possible to fit parameter  $k$  to number of arguments to be excluded. Method is more efficient for complexities near to  $m/2$ . Method is efficient for number of arguments up to 35.

Optimal algorithm of subset search can be built by combination of investigated methods and using different methods according to number of structures that should be checked.

## References

1. Statistical Methods for Digital Computers. Edited by Kurt Enslein, Anthony Ralston and HERBERT S. Wilf. John Wiley, New York, 1977