Multifactor Modelling with Regularization

Dr. Ventsislav Nikolov
Senior Software Developer
Eurorisk Systems Ltd.
31, General Kiselov Str., 9002 Varna, Bulgaria
E-mail: vnikolov at eurorisksystems dot com

Keywords: *Multifactor, Polynomial Formula, Basis Functions, Genetic Algorithm, Least Squares Regression, Regularization*

1. INTRODUCTION

Suppose we are given a finite number of discrete time series xi called factors. They can represent arbitrary physical, social, financial or other indicators. All factors are with equal length and their values correspond to measurements performed in equal time intervals. One of the series is chosen to be a target factor and some of the others are chosen to be explanatory factors. The aim is to create a formula by which a series can be generated, using the explanatory factors for the given historical period, that should be as close as possible to the given target series, using a chosen criterion [4]. For simplicity such a criterion can be the Euclidean distance between the target and generated factor for all data points. Such a created formula can be used for different purposes in the financial instruments modelling, sensitivity analysis, etc. In the case of predictable explanatory factors and unpredictable target factor analysis can be performed about the influence of the explanatory factors changes to the target factor.
The formula can be created in different forms but simplifying the solution the following polynomial form is used:

$$y = \beta_1 f_1(x_1) + \beta_2 f_2(x_2) + ... + \beta_m f_m(x_m) + \beta_{m+1} \quad (1)$$

where f_1, f_2, …f_m are arbitrary basis functions, and β_1, β_2, … β_m are regression coefficients, β_{m+1} is a free term without explanatory factor.

2. FORMULA GENERATION

First of all, the target factor is selected according to the specific purposes. After that the explanatory factors are selected amongst the all available series. In our solution a few alternative approaches can be used as selection of the most correlated factors to the target factor or minimal correlated each other or so on. When both the target and explanatory factors are selected the automatic modelling, stage is performed by repeating the stages of applying basis functions to explanatory factors and after that calculation of the regression coefficients.
Taking into account that for all selected factors all basis functions can be applied, there are k^m combinations, where k is the number of the basic functions and m is the number of the explanatory factors. Usually in the practice the factors are a few hundred and the functions are a few dozen. Thus, the brute force searching of the best basis functions combination is practically impossible. That is why for

that purpose we chose to apply heuristic approach by usage of a genetic algorithm. It is realized as a software library written in Java.

a. Finding the best combination of the basic functions

Initial population

The genetic algorithm is used to determine the combination of the basic functions to the explanatory factors. And a function can be used for more than one factor. Thus, an individual in terms of the genetic algorithms is a sequence of integer values representing the indices of the basic functions and the goodness of fit is the distance between the generated and the given target factor [2]. In the realized system a random integer sequence generator was created to generate the initial population of the sequences. Applying the functions to the explanatory factors and calculating the regression coefficients produces a set of target factors which are compared to the given target in order to select the best individuals.

Selection

Given a set of the generated individuals the best of them should be selected according to their goodness of fit. We have implemented two alternative approaches: roulette wheel and truncation selection [3]. The first one is preferred as default because it allows every individual to continue even with less chance.

Recombination and mutation

The recombination is performed by splitting the selected L individuals in a given point and randomly combining their parts. In our implementation the splitting point is randomly generated at every step within the interval from 25% to 75% of the individuals length rounded to the nearest integer.

Coefficients determination

The calculation of the regression coefficients is done for every combination of basic functions. In our case the ordinary least squares error is used according to which the coefficients are obtained in matrix form calculating the following matrix equation [1]:

$$B = (A^T A)^{-1} A^T Y \tag{2}$$

where B is the matrix of the regression coefficients, A is the matrix of factors with applied functions and Y is the target factor.
Having B calculated the generated target factor is:

$$\hat{Y} = A \times B \tag{3}$$

and the distance between the generated and given target is:

$$d = \left\| Y - \hat{Y} \right\| \tag{4}$$

Coefficients reduction

The formula terms with small coefficients can be removed because they do not significantly influence the formula results. Removing or not the small coefficients is an optional setting in our system and if it is chosen the second regression coefficients calculation must be performed at every step after the reduction.

Calibration

Using the generated formula for future calculations and modelling must be periodically reconsidered and the formula must be calibrated because its accuracy decreases. This can be done either by using the same explanatory factors or by other factors.

Regularization

The multifactor formula provides good results in the cases when there are explanatory factors similar to the target factor. Otherwise often the future calculations are not very accurate because of the overfitting. In order to avoid overfitting a regularization parameter is used in the following form:

$$B = (A^T A + \lambda I)^{-1} A^T Y \qquad (5)$$

where I is the identity matrix and λ is the regularization parameter.

This is L2-regularization or ridge regularization [5]. In the formula searching stage the set of the data points is separated in training and validation subsets. The formula functions and coefficients are determined using the training set but the error is calculated using the validation set. In order to separate these two sets the factors values are shuffled together and the last, for example, 20% or 30% of the length are used as validation set. When the training and validation sets are determined, and the basis functions are fixed to explanatory factors an appropriate value of λ should be found. Our investigation shows that there is a single global minimum of the validation error which allows searching it with adaptive step starting from a random point.

3. CONCLUSIONS AND FUTURE WORK

The built software prototype system can be seen on fig. 1. The experimental results show that the best results are obtained when the number of the explanatory factors is near to, but not exceeding, the number of the historical dates.

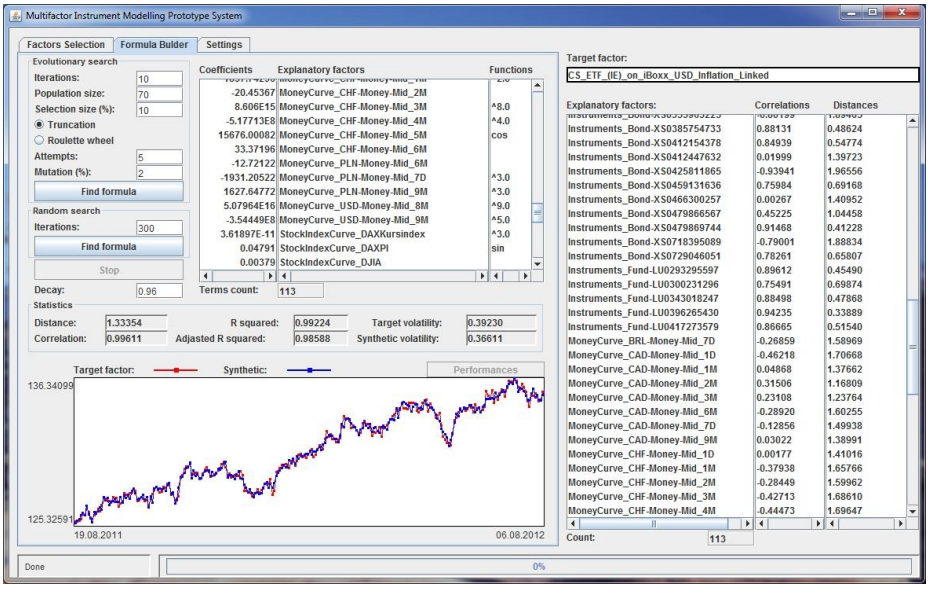

Fig. 1. The multifactor modelling prototype

The system also confirms that the greater the regularization parameter is the greater the penalization is which produces better results in the future calculations with generated formula in cases when the target factor is different to some extent than anyone of the explanatory factors. But this is not a general rule and taking into account that often in practice there are indicators with similar behavior sometimes the regularization parameter should not be used.

4. REFERENCES

1. Hamilton, J. Time Series Analysis. Princeton University Press, 1994.
2. Koza, J. Genetic Programming. MIT Press, 1992.
3. Mitchell, M, An Introduction to Genetic Algorithms. MIT Press, 1999.
4. Rosen, K. Discrete Mathematics and Its Applications, Fourth Edition. AT&T, 1998.
5. Rosenberg, A. Machine Learning Lectures, CUNY Graduate Center, 2009. (http://eniac.cs.qc.cuny.edu/andrew/gcml/lecture5.pdf)