MACQUARIE UNIVERSITY
SYDNEY ~ AUSTRALIA

Fuzzy linear regression with global continuous optimization

By

M. Hadi Mashinchi

This thesis is presented for
the degree of Doctor of Philosophy
Department of Computing, Faculty of Science
2012

© M. Hadi Mashinchi, 2012.

I certify that the work in this thesis entitled **"Fuzzy Linear Regression with Global Continuous Optimization"** has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in the thesis.

M. Hadi Mashinchi

Acknowledgements

I am very grateful to Prof. Mehmet Orgun for his guidance and support as my supervisor. His supervisory role, experience, and patience were inspirational to me. I also want to thank all the academics with whom I was in touch throughout my research career to date; Abhaya Nayak, Rajan Shankaran, Yan Wang, Mark Dras, Len Hamey, Scott McCallum, Xingquan (Hill) Zhu, Albert Zomaya, Frances Griffin, Mashaallah Mashinchi, Witold Pedrycz, Chi-Tsuen Yeh, and Nikola Kasabov. I am also grateful to David A. Barda who is both my friend and mentor and I have learnt a lot from him during my studies and working with him.

Special thanks go to the admin staff at Macquarie Univesity, graduate students and other people who have helped me throughout my research career, particularly Melina Chan, Jane Yang, Meredith McGregor, Donna Hua, Jackie Walsh, Sylvian Chow, Raina Kim, Aldrinine Creado, Peter Curran, Reza Sepahi, Armin Hezart, Lei Li and Kerstin Klemisch.

I would also like to thank Andy Connor, Sandhya Samarasinghe, Siti Mariyam HJ. Shamssudin, and Bernard DeBaets who hosted me at their institutions during my PhD studies. I am also very grateful to Macquarie University for funding my PhD research.

The last but not the least, my very special thanks go to my father, mother and brother whom without their never stopping friendship and support I would not be able to be what I am now.

List of Publications

- M. H. Mashinchi, M. A. Orgun, M. Mashinchi, and W. Pedrycz. *A tabu-harmony search based approach to fuzzy linear regression*. IEEE Transactions on Fuzzy Systems 19(3), 432-448 (2011).

- M. H. Mashinchi, M. A. Orgun, and W. Pedrycz. *Hybrid optimization with improved tabu search*. Applied Soft Computing 11(2), 1993-2006 (2011).

- M. H. Mashinchi, L. Li, M. A. Orgun, and Y. Wang. *The prediction of trust rating based on the quality of services using fuzzy linear regression*. In IEEE International Conference on Fuzzy Systems, 1953-1959, IEEE Computer Society Press (2011).

- M. H. Mashinchi, Mehmet A. Orgun, and M. R. Mashinchi. *A least square approach for the detection and removal of outliers for fuzzy linear regression*. In World Congress on Nature & Biologically Inspired Computing, 134-139, IEEE Computer Society Press (2010).

- M. H. Mashinchi, M. A. Orgun, and M. Mashinchi. *Solving fuzzy linear regression with hybrid optimization*. In 16th International Conference on Neural Information Processing, Lecture Notes in Computer Science (LNCS), 336-343, Springer- Verlag Berlin Heidelberg (2009).

Abstract

Finding the global optimum of an unknown system has attracted a great deal of interest in many engineering problems. In this setting, meta-heuristics are very common and efficient approaches for solving complex real-world problems in Global Continuous Optimization Problems (GCOPs) as they can approximate solutions without any need for mathematical assumptions such as differentiability. The application of global continuous optimization methods is essential in many engineering applications where an optimization problem has certain properties such as *unreliable derivatives* and/or *black-box nature*. Meta-heuristic based optimizations, as one of the promising approaches in global continuous optimization, have a slow rate of convergence. Hybridization frameworks are investigated as a potential way of enhancing the optimization speed, and the quality of solutions.

Fuzzy linear regression analysis is a powerful tool to model the input-output relationship for forcasting purposes or studying the behavior of the data. The existing challenges in fuzzy linear regression are, dealing with non-transparent fitness measures, outlier detection and spread increasing problem. The application of global continuous optimization is investigated to tackle these issues. We propose an Unconstrained Global Continuous Optimization (UGCO) method based on tabu search and harmony search to support the design of Fuzzy Linear Regression models (FLR). The proposed approach offers the flexibility of using any kind of an objective function based on the client's requirements or requests and the nature of the data set, and then attains its minimum error.

Fuzzy linear analysis may lead to an incorrect interpretation of data in case of being incapable of dealing with outliers. Both basic probabilistic and least squares approaches are sensitive to outliers. In order to detect outliers, we propose a two stage least squares approach based on global continuous optimization which outperforms some issues that exist in other methods. In both the first and second phases, the minimization of the model fitting measurement is achieved by hybrid optimization which gives us the flexibility of using any type of model fitting measures regardless of being continuous, differentiable, or transparent.

Some of the fuzzy linear regression models suffer from constantly increasing spreads of the outputs with the increase in the magnitude of the inputs. Such models are known to have the so-called spread increasing problem. We introduce a model, obtained by the application of hybrid optimization, which is capable of having variable spreads for different input variables regardless of their magnitute. The proposed approach is also compared and contrasted with other models in terms of the number of parameters, the flexibility of spreads, and errors.

Contents

Acknowledgements **iv**

List of Publications **v**

Abstract **vi**

List of Figures **xi**

List of Tables **xiii**

1 Introduction **1**

1.1 Overview 2

1.2 Contributions 4

1.3 Organization 5

2 Global continuous optimization **8**

2.1 Global continuous optimization 8

2.2 When to use global continuous optimization 10

2.3 Global continuous optimization approaches 12

2.3.1 Meta-heuristics based approaches 13

2.3.1.1 Genetic Algorithms 14

2.3.1.2 Tabu search 15

2.3.1.3 Harmony search 16

2.3.1.4 Simulated annealing 19

2.3.2 Direct search strategies 20

2.3.2.1 Nelder-Mead 20

2.3.2.2 Pattern search 22

2.3.3 Surrogate models 22

2.4 Hybridization methods in meta-heuristics based approaches 23

2.4.1 Hybridization frameworks in meta-heuristics 25

2.5 Application of global optimization for fuzzy regression 28

2.6 Summary 31

3 Fuzzy linear regression 33

3.1 Fuzzy logic 33

3.1.1 Fuzzy sets and fuzzy numbers 34

3.1.2 Fuzzy arithmetics 36

3.1.3 Fuzzy similarity measures 37

3.2 Fuzzy regression 39

3.2.1 Linear programming approaches 41

3.2.2 Least square approaches 44

3.2.3 Comparison between least square and linear programming approaches 46

3.3 Challenges in fuzzy linear regression 47

3.3.1 Fitness measure 47

3.3.2 Outlier detection 48

3.3.3 The spread increasing problem 49

3.4 Summary 49

4 Approaching fuzzy linear regression by global continuous optimization 50

4.1 Tabu-Harmony search optimization 51

4.1.1 The UGCO method 52

4.1.1.1 First Phase: Tabu Search 52

4.1.1.2 Second Phase: Improved harmony Search 55

4.2 Modeling fuzzy linear regression analysis as a global continuous optimization problem 59

4.3 Method's configurations 62

4.4 Summary 66

5 Performance evaluation of the proposed fuzzy linear regression analysis method 67

5.1 Numeric input-symmetric fuzzy output results 67

5.2 Symmetric fuzzy input-fuzzy output results 81

5.3 Numeric input-asymmetric fuzzy output results 85

5.4 Summary 85

6 **Approaching outlier detection and spread increasing problem by global continuous optimization** **87**
6.1 Outlier detection 88
6.1.1 Outlier detection approaches 89
6.1.2 Outlier detection with global continuous optimization 91
6.1.3 Results 93
6.1.3.1 small dataset with outliers happening in the spreads 94
6.1.3.2 large dataset with outliers happening in both centers and spreads 95
6.2 Spread increasing problem 100
6.2.1 Anti-spread increasing approaches 100
6.2.2 Drawbacks of the existing methods 104
6.2.3 Anti-spread increasing approach based on global continuous optimization 105
6.2.4 Experimental results 107
6.3 Summary 108

7 **Real-world applications of global continuous optimization and fuzzy linear regression analysis** **109**
7.1 Real-world application: The design of a pressure vessel 109
7.2 Application of fuzzy regression to prediction of quality of service . . 113
7.2.1 Why fuzzy linear regression for quality of service prediction . 113
7.2.2 Trust prediction based on fuzzy regression 115
7.2.3 Experiments 117
7.3 Summary 122

8 **Conclusions and future work** **123**
8.1 Conclusions of each chapter 123
8.2 Future research direction 125

A **Appendices** **128**
A.1 Optimization results for benchmark functions 128
A.2 Mathematical formulation of the benchmark function 140
A.3 Fuzzy outputs and the errors by Shakouri and Nadimi approach [1] . 142

Bibliography **144**

List of Figures

2.1 Hybridization strategies for diversification and intensification (taken from [2]) . . . 26

2.2 An example of the *simple divers-intens* drawback (taken from [2]) . . . 28

2.3 The computationally expensive example for *iterative divers-intens* hybridization strategy (taken from [2]) . . . 29

2.4 An example for *divers-semiintens-intens* hybridization strategy (taken from [2]) . . . 30

3.1 The fuzzy set of the comfortability of the temperature (derived from [3]) 34

3.2 An example of the non-intersected area of two fuzzy numbers which is used in the numerator of the objective function given in equation (6.3) (derived from [4]) . . . 38

3.3 Illustration of Tanaka et al. model [5] (derived from [4]) . . . 42

4.1 Generaing neighbors and tabulating areas based on DF and SF (taken from [6]) . . . 54

4.2 Modelling of a FLR problem regarded as an UGCO problem (taken from [6]) . . . 59

5.1 The sorted generated errors (4.9) for 1,000 separate runs for data set given in Table 5.1 obtained by the proposed method (taken from [6]) . 69

5.2 The estimated fuzzy function for the data given in Table 5.1: a comparative analysis involves six other models available in the literature (taken from [6]) . . . 70

5.3 The convergence of the error for the best model over the data set in Table 5.7 (taken from [6]) . . . 77

5.4 The convergence of the error for the best model over the data set in Table 5.11 (taken from [6]) . . . 77

5.5 The estimated fuzzy function for the given data in Table 5.16 (taken from [6]) . . . 83
5.6 The estimated fuzzy function for the data set given in Table 5.19 (taken from [6]) . . . 86

6.1 Comparison between Chen's estimated model [7] and our approach over Dataset 1 (taken from [8]) . . . 95
6.2 An example of a semi-linear fuzzy regression model (taken from [9]) . 106

7.1 The structure of the pressure design with its effective variables on total cost (taken from [2]) . . . 112
7.2 The comparison of the trust of two web services before defuzzification (taken from [10]) . . . 121
7.3 An example where a traditional method and an inexact method may vary in the result (taken from [10]) . . . 121

List of Tables

2.1 Timeline of innovation in global continuous optimization 13

4.1 Parameter settings for both phases of the proposed method 63

5.1 Tanaka et al. data set [11] 68
5.2 The experimental result for 1,000 runs for the data set given in Table 5.1 by the proposed method (taken from [6]) 68
5.3 Comparative analysis of fuzzy models in Table 5.1 (taken from [6]) . . 69
5.4 The average and generated error of the estimated models for the given data set in Table 5.1 (taken from [6]) 71
5.5 Continue from Table 5.4 72
5.6 Continue from Table 5.4 72
5.7 Kim and Bishu's data set [12] 73
5.8 The experimental result for 1,000 runs for the data set given in Table 5.7 (taken from [6]) 74
5.9 Models available in the literature for the data set in Table 5.7 (taken from [6]) 75
5.10 The average and generated error of the estimated models for the given data set in Table 5.7 (taken from [6]) 76
5.11 Tanaka et al. house price data set [5] 78
5.12 The experimental result for 1,000 runs for the data set given in Table 5.11 (taken from [6]) 79
5.13 Hong and Hwang's data set [13] 79
5.14 Estimated models and their errors for the data set given in Table 5.13 (taken from [6]) 80
5.15 The experimental result for 1,000 runs for the data set given in Table 5.13 by the proposed method (taken from [6]) 81

5.16 Sakawa and Yano's data set [14] 82
5.17 The experimental result for 1,000 runs for the data set given in Table 5.16 (taken from [6]) 82
5.18 Estimated models and generated errors for several models and data shown in Table 5.16 (taken from [6]) 84
5.19 Artificial asymmetric numeric input-fuzzy output (taken from [6]) . . 85
5.20 The experimental result for 1,000 runs for the data set given in Table 5.19 by the proposed method (taken from [6]) 86

6.1 Constant, increasing and decreasing spreads datasets with existence of outliers (taken from [8]) 94
6.2 The estimated error (4.9) of the best obtained model for the outliers in the large dataset (taken from [8]) 97
6.3 Comparison between different methods for the Datasets 1 and 2 (taken from [8]) 99
6.4 The summary of the studies to tackle SIP (taken from [9]) 103
6.5 Sought SIP-free models for different datasets in the literature 107
6.6 Comparison of the SIP-free errors with FLR model on different datasets in the literature 107

7.1 Comparison of the best solution for the pressure design vessel by different methods (taken from [2]) 111
7.2 Web-service dataset; attributes and descriptions (taken from [15]) . . 118
7.3 Accuracy of the classifiers after removing WSRF (taken from [10]) . . 120

A.1 Effect of STTL and LTTL sizes on the convergence, speed and robustness of the proposed method (taken from [2]) 130
A.2 List of various GCOMs used for the comparison (taken from [2]) . . . 133
A.3 The value of control variables (taken from [2]) 134
A.4 Comparison between the proposed method and other GCOMs for low dimensional benchmark functions ($n < 4$) with small or medium input domains ($|a_i - b_i| < 20,\ i = 1,\cdots,n$) (taken from [2]) 135
A.5 (Cont.) Comparison between the proposed method and other GCOMs for low dimensional benchmark functions ($n < 4$) with small or medium input domains ($|a_i - b_i| < 20,\ i = 1,\cdots,n$) (taken from [2]) 136

A.6 (Cont.) Comparison between the proposed method and other GCOMs for low dimensional benchmark functions ($n < 4$)with large input domains ($|a_i - b_i| \geq 20,\ i = 1, \cdots, n$) (taken from [2]) 137

A.7 Comparison between the proposed method and other GCOMs for high dimensional benchmark input domains ($n \geq 4$) (taken from [2]) . 138

A.8 (Cont.) Comparison between the proposed method and other GCOMs for high dimensional benchmark input domains ($n \geq 4$) (taken from [2])139

A.9 The incorrect reported fuzzy output given by Shakouri and Nadimi in [1] for data set given in Table 5.7 versus the actual fuzzy outputs (taken from [6]) 143