Regression of a Stock Market Dataset

Nov 2017

Eduardo Guilherme Ferreira Morais de Araújo
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
eduardo.araujo@tecnico.ulisboa.pt

1. Introduction

The current document presents the developed work in the scope of Intelligent Systems discipline, where a regression problem was solved using a fuzzy modelling approach, for which a fuzzy model was derived and its parameters optimized.

The dataset used was downloaded from the KEEL Dataset Repository and concerns daily stock prices for ten aerospace companies. The task was to approximate the price of the 10th company given the prices of the rest.

To model the fuzzy inference system a type-1 Takagi-Sugeno model was employed, using the Gustafson-Kessel (GK) as clustering algorithm.

The algorithmic work was implemented resorting to both Jáno's Fuzzy Clustering and Data Analysis Toolbox and Babuska's Fuzzy Identification Toolbox.

2. Regression Problem

2.1. Dataset Description

The dataset is composed of daily stock prices from January 1988 through October 1991, for ten aerospace companies.

There are 950 instances, with no data missing and 9 features labelled *Company i*, with i = 1, 2, 3 ... 9.

2.2. Dataset Partitioning

Since the problem at hand is a regression problem its dynamics in time had to be taken into account. To this end data was divided into groups of 9 equally spaced sets (Δt), which was then subdivided into two groups: training (70%) and testing (30%), as shown in figure 1.

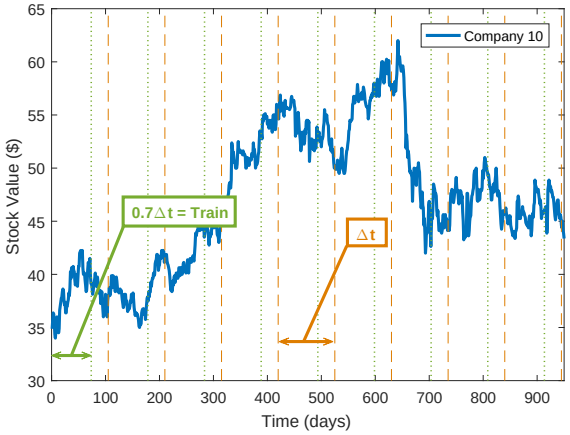

Figure 1. Data Partitioning

3. Fuzzy Modelling

3.1. Takagi-Sugeno Model

The **Tsukagi-Sugeno fuzzy models** (also known as *TS* fuzzy models) are characterized by the fact that their *consequents* are linear functions of the *antecedent* variables instead of fuzzy sets.

The rule base in MISO TS models have the following structure,

$$R^k : \textit{If } x_1 \textit{is } A_1^k \textit{and } x_2 \textit{is } A_2^k \textit{ and ... and } x_n \textit{is } A_n^k \textit{ then}$$

$$y^k = \sum_{j=1}^{n} a_j^k x_j + b^k$$

where R^k is the kth rule in the rule-base, x_1, ..., x_n are the premise variables, y^k is the output of the kth rule and A_1^k, ..., A_n^k are the fuzzy sets defined over their respective universes of discourse.

Since each rule has a crisp output, the overall output is obtained via **weighted sum** of each of the rule consequents, given by

$$y^* = \frac{\sum_{k=1}^{K} y^k \beta^k}{\sum_{k=1}^{K} \beta^k} \tag{1}$$

where K is the total number of rules, β^k is the non-normalized degree of fulfilment of the kth rule premise and y^k is the output of rule k.[1]

3.2. Clustering Methods

Clustering is an unsupervised learning task that aims at decomposing a given set of objects into subgroups

or clusters based on similarity. It is primarily a tool for discovering previously hidden structure in a set of unordered objects, which implies one assumes that a 'true' or natural grouping exists in the data [3].

The number of clusters determines the number of rules in the obtained fuzzy model. Thus this parameter heavily influences the accuracy and transparency of the fuzzy models [1].

Since the Gustafson–Kessel Algorithm is usually preferred when clustering is applied for the generation of fuzzy rule systems [3], this was the algorithm used throughout the analysis.

1) Gustafson-Kessel Algorithm: The Gustafson–Kessel algorithm extends the Fuzzy c-means, replacing the Euclidean distance by a cluster-specific Mahalanobis distance (Eq. 3), so as to adapt to various sizes and forms of the clusters.[3]

The algorithm can be expressed as follows

$$J(X,U,V,A) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{i,k})^m D^2_{i,kA_i} \tag{2}$$

where

$$D^2_{i,kA_i} = (x_k - v_i)^T A_i (x_k - v_i), \tag{3}$$

$$1 \leq i \leq c, \tag{4}$$

$$1 \leq k \leq N \tag{5}$$

the matrices A_i are used as optimization variables, which allow each cluster to adapt the distance norm to the local topological structure of the data.

In order to minimize J, Ai has to be made less positive definite. Allowing the matrix Ai to vary with its determinant fixed corresponds to optimizing the cluster's shape while its volume remains constant.

$$A_i = |F_i|^{\frac{1}{(n+1)}} F_i^{-1}, \tag{6}$$

$$F_i = \frac{\sum_{k=1}^{N}(\mu_{i,k})^m (x_k - v_i)(x - k - v_i)}{\sum_{k=1}^{N}(\mu_{i,k})^m} \tag{7}$$

where F_i represents the fuzzy covariance matrix. [6]

3.3. Parameter Estimation

Clustering algorithms always fit the clusters to the data, even if the cluster structure is not adequate for the problem [3] .

Unfortunately, fuzzy clustering algorithms do not give any indication of the correct number of clusters needed. For this reason the conventional approach to determine a correct number of clusters in cluster analysis is based on validity measures.[1]

1) Validity Measures: From [5**?**] 4 validity measures were used and are presented in the following paragraphs.

1) Partition Index (SC): The SC (Eq. 8) is the ratio of the sum of compactness and separation of clusters. It is useful when comparing different cluster partitions which have equal number of clusters. A better partition is given by a smaller value of SC.

$$SC(c) = \sum_{i=1}^{c} \frac{\sum_{k=1}^{n} U_{ik}^m \|x_k - v_i\|^2}{n_i \sum_{q=1}^{c} \|v_q - v_i\|^2} \tag{8}$$

2) Separation Index (S): The S (Eq. 9) uses a minimum distance separation for partition validity, contrary to SC. A better partition is as well given by a smaller value of S.

$$S(c) = \sum_{i=1}^{c} \frac{\sum_{k=1}^{n} U_{ik}^m \|x_k - v_i\|^2}{n \quad min_i q \|v_q - v_i\|^2} \tag{9}$$

3) Xie and Beni's Index (XB): The XB (Eq. 10) intend to quantify the ratio of the total variation within clusters and their separation. The optimal number of clusters is given by the resulting smaller value.

$$XB(c) = \sum_{i=1}^{c} \frac{\sum_{k=1}^{n} U_{ik}^m \|x_k - v_i\|^2}{n \quad min_i k \|x_k - v_i\|^2} \tag{10}$$

4) Variance Accounted For (VAF): Variance accounted for (VAF) (Eq. 11) determines the percentile variance measured between two signals and is given by

$$VAF = 100 \times \frac{Cov(y - y^*)}{Cov(y)} \tag{11}$$

where y is the measured data and y^* the model output.

4. Methodology

To determine a suitable number of clusters (c) for which to optimize the model, a large amount of simulations was done varying c, and m, to check how its performance behaved.

The methodology adopted consisted on varying the values of the afore mentioned parameters from 2 up to 20, followed by a selection based on the mean value of VAF. The methodology is schematically represented in figure 2.

5. Results

5.1. Clustering

Figures 3 and 4 present the results obtained using the different validity measures.

It is clear the model's performance gets better as the number of clusters increases. Although this result was to be expected, one of the great advantages of working with fuzzy models is the balance between performance and transparency. Thus, noting that from figures 3 and 4 the best choice was to set $c = 5$ and $m = 2$.

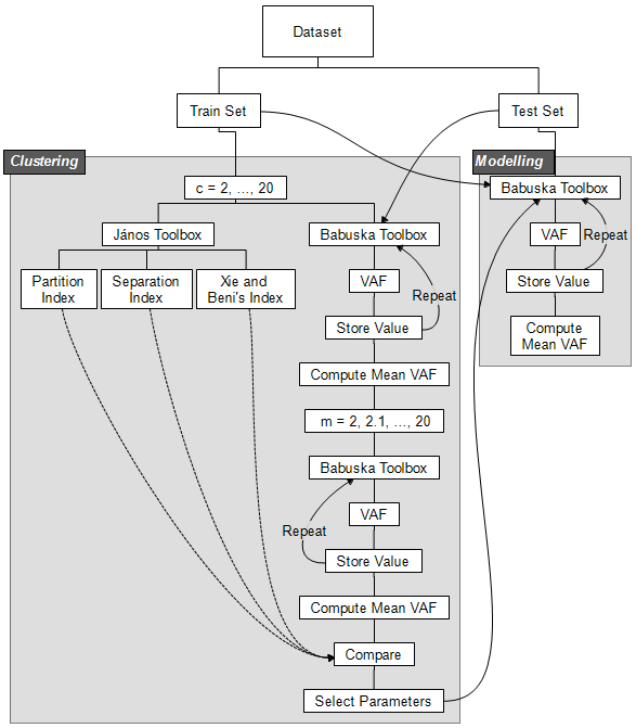

Figure 2. Methodology Employed

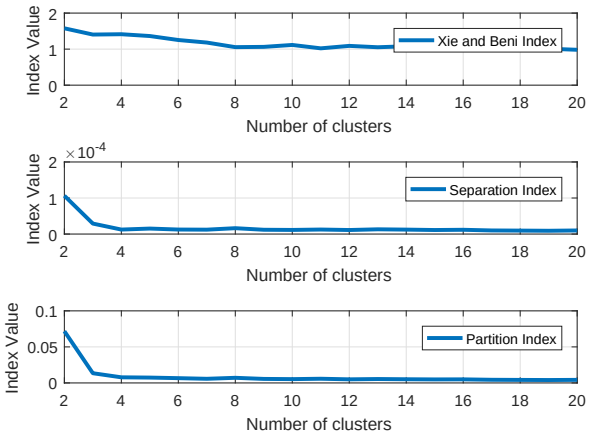

Figure 3. Results for different Validity Indexes

5.2. *Fuzzy Model*

As can be seen from figure 5 there are clearly 5 main groups of data, with some variability in size and compactness.

Looking now at the membership functions presented in figure 5.2 it is not only clear the reduction in complexity but also the relationship between it has with figure 5.

Finally figure 5.2 shows the comparison between the process and the model's behaviour.

Making use of the function *fm2tex* it was possible to

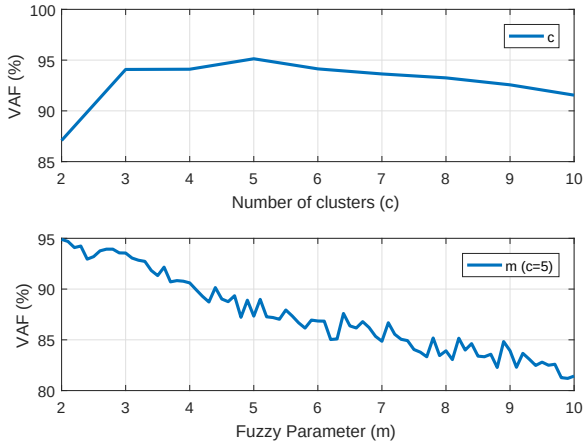

Figure 4. Number of clusters vs Performance

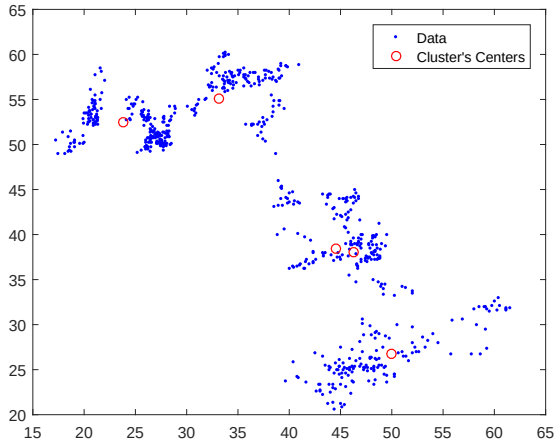

Figure 5. Clusters configuration for $c = 5$

have much more information about the model developed. For the sake of brevity, those informations can be consulted in Appendix 6.

Although, it is worth noting that from a batch of 200 simulations, the average VAF registered was 94.84%, which is a very good result taking into account the process to be identified is somewhat non-causal, since the system's output does not depend exclusively on the inputs.

References

[1] Joao M. C. Sousa, Uzay Kaymak, *Fuzzy Decision Making in Modelling and Control*. 2002 World Scientific and Imperial College Editors, 2002.

[2] J.-S. Jang, C.-T., Sun and E. Mizutani *A Computational Approach to Learning and Machine Intelligence*. 1997 Prentice Hall, New Jersey

[3] J. Valente de Oliveira, W. Pedrycz, *Advances in Fuzzy Clustering and its Applications*. John Wiley & Sons Ltd

[4] Robert Babuska, *Fuzzy Modeling for Control*. 1998 Kluwer Academic Publishers

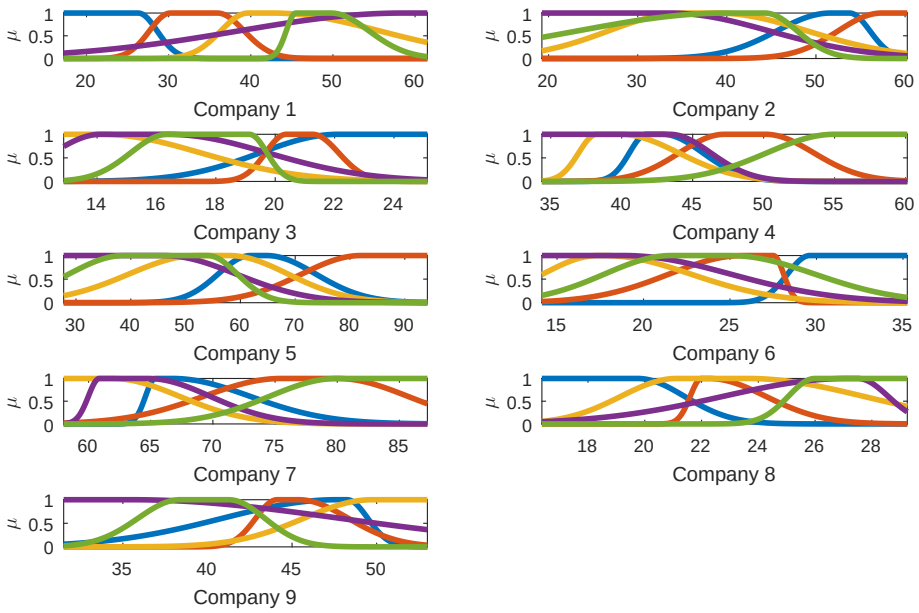

Figure 6. Membership Functions

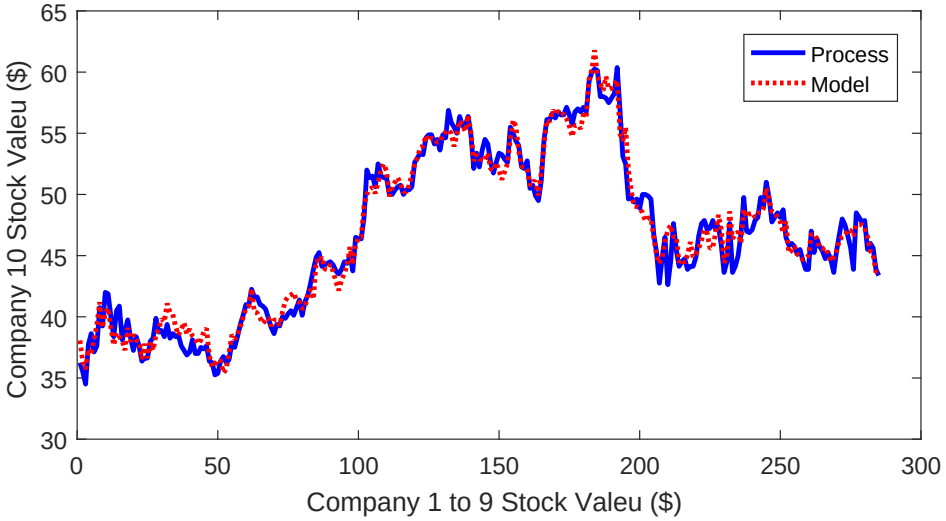

Figure 7. Process Vs Model

[5] Fuzzy Identification Toolbox for MATLAB (ver. 4.0), `http://www.dcsc.tudelft.nl/%7Erbabuska/`

[6] János Abonyi, Balázs Feil, *Cluster Analysis for Data Mining and System Identification*. 2007 Birkhäuser Verlag AG

6. Appendix - fm2tex Results

The output-specific parameters are given in the following table.

Table 1
Model parameters.

output	antecedent	c	m	n_y	n_u
1	2	5	2.2	{{ [ ] }, { [ ] }}	{{ [0] [0] [0] [0] [0] [0] [0] [0] [0] }, { [0] [0] [0] [0] [0] [0] [0] [0] [0] }}

In the following, the output-specific information is shown for each output.
Output1:
Rules:

1.u_1A_{11} & u_2A_{12} & u_3A_{13} & u_4A_{14} & u_5A_{15} & u_6A_{16} & u_7A_{17} & u_8A_{18} & u_9A_{19}
$y(k) = 3.9 \cdot 10^{e-1}u_1 + 4.6 \cdot 10^{e-1}u_2 - 5.8 \cdot 10^{e-1}u_3 + 2.8 \cdot 10^{e-1}u_4 + 9.9 \cdot 10^{e-2}u_5 - 4.9 \cdot 10^{e-1}u_6 + 4.5 \cdot 10^{e-1}u_7 - 1.8 \cdot 10^{e-1}u_8 + 1.1 \cdot 10^{e-1}u_9 - 1.8 \cdot 10^{e+1}$

2.u_1A_{21} & u_2A_{22} & u_3A_{23} & u_4A_{24} & u_5A_{25} & u_6A_{26} & u_7A_{27} & u_8A_{28} & u_9A_{29}
$y(k) = 6.1 \cdot 10^{e-1}u_1 + 2.2 \cdot 10^{e-1}u_2 + 1.8 \cdot 10^{e-1}u_3 + 3.3 \cdot 10^{e-1}u_4 + 2.4 \cdot 10^{e-1}u_5 - 6.2 \cdot 10^{e-1}u_6 + 1.4 \cdot 10^{e-1}u_7 - 3.5 \cdot 10^{e-1}u_8 + 3.8 \cdot 10^{e-1}u_9 - 2.3 \cdot 10^{e+1}$

3.u_1A_{31} & u_2A_{32} & u_3A_{33} & u_4A_{34} & u_5A_{35} & u_6A_{36} & u_7A_{37} & u_8A_{38} & u_9A_{39}
$y(k) = 4.4 \cdot 10^{e-1}u_1 + 3.1 \cdot 10^{e-1}u_2 + 1.0 \cdot 10^{e-2}u_3 + 1.0 \cdot 10^{e-1}u_4 - 1.9 \cdot 10^{e-1}u_5 - 6.8 \cdot 10^{e-2}u_6 - 2.1 \cdot 10^{e-1}u_7 - 2.8 \cdot 10^{e-2}u_8 + 7.9 \cdot 10^{e-1}u_9 + 9.5 \cdot 10^{e+}$

4.u_1A_{41} & u_2A_{42} & u_3A_{43} & u_4A_{44} & u_5A_{45} & u_6A_{46} & u_7A_{47} & u_8A_{48} & u_9A_{49}
$y(k) = 1.3 \cdot 10^{e-1}u_1 - 1.5 \cdot 10^{e-1}u_2 + 5.1 \cdot 10^{e-2}u_3 - 2.0 \cdot 10^{e-1}u_4 - 3.8 \cdot 10^{e-2}u_5 + 2.1 \cdot 10^{e-1}u_6 + 1.9 \cdot 10^{e-1}u_7 + 3.3 \cdot 10^{e-1}u_8 + 2.3 \cdot 10^{e-1}u_9 + 1.7 \cdot 10^{e+1}$

5.u_1A_{51} & u_2A_{52} & u_3A_{53} & u_4A_{54} & u_5A_{55} & u_6A_{56} & u_7A_{57} & u_8A_{58} & u_9A_{59}
$y(k) = 5.0 \cdot 10^{e-1}u_1 + 5.0 \cdot 10^{e-1}u_2 - 4.1 \cdot 10^{e-1}u_3 + 4.6 \cdot 10^{e-1}u_4 + 1.5 \cdot 10^{e-2}u_5 - 2.5 \cdot 10^{e-1}u_6 - 4.1 \cdot 10^{e-1}u_7 - 2.6 \cdot 10^{e-2}u_8 - 8.6 \cdot 10^{e-2}u_9 + 3.4 \cdot 10^{e+1}$

Table 2
Consequent parameters.

rule	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	offset
1	$3.9 \cdot 10^{e-1}$	$4.6 \cdot 10^{e-1}$	$-5.8 \cdot 10^{e-1}$	$2.8 \cdot 10^{e-1}$	$9.9 \cdot 10^{e-2}$	$-4.9 \cdot 10^{e-1}$	$4.5 \cdot 10^{e-1}$	$-1.8 \cdot 10^{e-1}$	$1.1 \cdot 10^{e-1}$	$-1.8 \cdot 10^{e+1}$
2	$6.1 \cdot 10^{e-1}$	$2.2 \cdot 10^{e-1}$	$1.8 \cdot 10^{e-1}$	$3.3 \cdot 10^{e-1}$	$2.4 \cdot 10^{e-1}$	$-6.2 \cdot 10^{e-1}$	$1.4 \cdot 10^{e-1}$	$-3.5 \cdot 10^{e-1}$	$3.8 \cdot 10^{e-1}$	$-2.3 \cdot 10^{e+1}$
3	$4.4 \cdot 10^{e-1}$	$3.1 \cdot 10^{e-1}$	$1.0 \cdot 10^{e-2}$	$1.0 \cdot 10^{e-1}$	$-1.9 \cdot 10^{e-1}$	$-6.8 \cdot 10^{e-2}$	$-2.1 \cdot 10^{e-1}$	$-2.8 \cdot 10^{e-2}$	$7.9 \cdot 10^{e-1}$	$9.5 \cdot 10^{e+}$
4	$1.3 \cdot 10^{e-1}$	$-1.5 \cdot 10^{e-1}$	$5.1 \cdot 10^{e-2}$	$-2.0 \cdot 10^{e-1}$	$-3.8 \cdot 10^{e-2}$	$2.1 \cdot 10^{e-1}$	$1.9 \cdot 10^{e-1}$	$3.3 \cdot 10^{e-1}$	$2.3 \cdot 10^{e-1}$	$1.7 \cdot 10^{e+1}$
5	$5.0 \cdot 10^{e-1}$	$5.0 \cdot 10^{e-1}$	$-4.1 \cdot 10^{e-1}$	$4.6 \cdot 10^{e-1}$	$1.5 \cdot 10^{e-2}$	$-2.5 \cdot 10^{e-1}$	$-4.1 \cdot 10^{e-1}$	$-2.6 \cdot 10^{e-2}$	$-8.6 \cdot 10^{e-2}$	$3.4 \cdot 10^{e+1}$