Proceedings of the World Congress on Engineering 2007 Vol II
WCE 2007, July 2 - 4, 2007, London, U.K.

Model Selection in Functional Networks via Genetic Algorithms

R. E. Pruneda and B. Lacruz *

Abstract—**Several statistical tools and most recently Functional Networks (FN) have been used to solve nonlinear regression problems. One of the tasks associated with all of these methodologies consists of discovering the functional form of the contribution of the explanatory variables to the response variable. In this paper, we tackle this problem using functional network models (FNs). Since these models usually involve from a moderate to high number of parameters, a genetic algorithm (GA) for model selection is proposed. After an introduction of FNs and GAs, the performance of the proposed methodology is assessed using a simulation study as well as a real-life data set.**

Keywords: functional networks, genetic algorithms, nonlinear regression

1 Introduction

Functional networks have been applied to reproduce the relationship between a *response* variable Y and one or more *predictor* variables $X_1, X_2, \ldots, X_k$ in [3], [4] and [5]. In this paper, we consider that the relationships among these variables can be written as

$$f(Y) = h(X_1, X_2, \ldots, X_k) + \epsilon. \tag{1}$$

where ϵ is a random *error* whose expected value is assumed to be 0. Our purpose is to discover the structure of the transformations f and h in (1).

Functional networks can be seen as the graphical representation of a functional equation (an equation where the unknowns are functions), which provides a better understanding of the properties of the model at hand. The principal steps to work with functional networks consist of (1) selecting the topology (guarantying the uniqueness) and a set of basic functions to approximate the unknown functions in the model and (2) learning, which involves to choose a criterion to estimate the parameters and a procedure to select the best model.

*Dpt. of Mathematics, University of Castilla-La Mancha and Dpt. of Statistical Methods, University of Zaragoza (Spain). E-mails: rosa.pruneda@uclm.es, lacruz@unizar.es. The authors are partially supported by the Ministry of Science and Technology of Spain through CICYT/FEDER Project MTM2006-15671, by the Junta de Comunidades de Castilla-La Mancha, through project PAI-05-044 and Gobierno de Aragón (through consolidated research group Stochastic Models).

In this paper, we present two basic functional network topologies: the additive model and the additive general model, which are briefly described in Sections 2.1 and 2.2. The advantage of these two models is that the estimation problem can be solved using the constrained least squares criterion, which leads to solve a linear system of equations. Moreover, the set of basic functions chosen to approximate the unknown functions is the polynomial family of linearly independent functions $\Phi = \{1, t, t^2, \ldots, t^q\}$.

Model selection is tackled via Genetic Algorithms (GAs). They are heuristic search algorithms based on the evolutionary ideas of natural selection and genetics. An introduction can be found in [7]. Previous works in functional networks consider forward-backward or exhaustive searching methods ([3], [4] and [5]). But, because of the computational cost, these procedures are useless when the number of parameters is large.

The rest of the paper is structured as follows. In Section 2 two basic functional network models are introduced. Genetic algorithms are described in Section 3 together with the strategies for its application to our particular problem. The performance of the proposed techniques is showed in Sections 4 and 5, where a simulation study and a real-life data set are presented.

2 Some functional network models

We propose to approximate (1) using the additive and the general additive functional network models. The additive model approximates h in (1) by a sum of functions, one on each explanatory variable, which allows us to analyse the contribution of each predictor separately. The general additive model also includes the interactions among them.

2.1 Additive model

The additive functional equation is

$$f(y) = h_1(x_1) + h_2(x_2) + \ldots + h_k(x_k), \tag{2}$$

which leads to the functional network showed in Figure 1.

To estimate f and $h_1, \ldots, h_k$ in (2) we consider linear

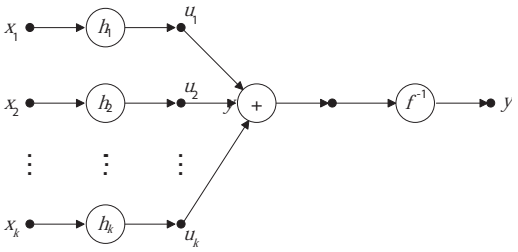

Figure 1: The additive functional network.

combinations of basic functions ϕ, that is,

$$\hat{f}(y) = \sum_{j=1}^{q_0} a_{0j}\phi_{0j}(y) \text{ and } \hat{h}_i(x_i) = \sum_{j=1}^{q_i} a_{ij}\phi_{ij}(x_i) \quad (3)$$

and the problem is reduced to estimate the parameters $a_{ij}, \forall i, j$. If the set of basic functions is the polynomial family, equation (2) can be approximated by

$$\sum_{j=1}^{q_0} a_{0j}y^j = a + \sum_{j=1}^{q_1} a_{1j}x_1^j + \ldots + \sum_{j=1}^{q_k} a_{kj}x_k^j, \quad (4)$$

where a is the constant term, which is included just once for avoiding identifiability problems. Note that the number of parameters in this model is $\sum_{i=0}^{k} q_i + 1$. If $q_i = q$, $\forall i$, then the number of parameters is $q\times(k+1)+1$.

2.2 The General Additive Model

A more general form of approximating h in (1) is considered when we use the general additive functional equation

$$f(y) = \sum_{r_1=1}^{q_1} \cdots \sum_{r_k=1}^{q_k} c_{r_1\ldots r_k}\phi_{r_1}(x_1)\ldots\phi_{r_k}(x_k), \quad (5)$$

where $c_{r_1\ldots r_k}$ are unknown parameters and each ϕ belongs to the family of basic functions. An example of its corresponding functional network model, for $k = 2$ and $q_1 = q_2 = q$, is shown in Figure 2.

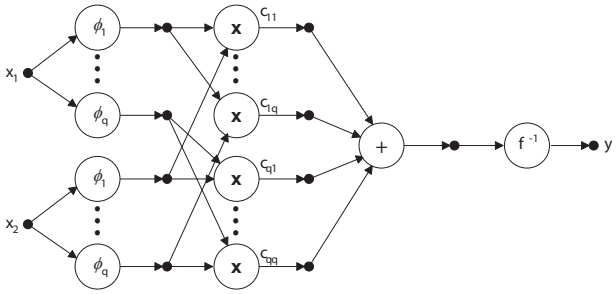

Figure 2: The general additive functional network for $k = 2$ and $q_1 = q_2 = q$.

To estimate the unknown function in (5) we just need to approximate f as in (3). Then, the problem is reduced to estimate the parameters a_{0j} and $c_{r_1\ldots r_k}$, $\forall j, r_1, \ldots, r_k$. If the set of basic functions is the polynomial family, equation (5) can be approximated by

$$\sum_{j=1}^{q_0} a_{0j}y^j = \sum_{r_1=0}^{q_1} \cdots \sum_{r_k=0}^{q_k} c_{r_1\ldots r_k}x_1^{r_1}\ldots x_k^{r_k}. \quad (6)$$

Note that the number of parameters in this model is $q_0 + \prod_{i=1}^{k}(q_i+1)$. When $q_i = q$, $\forall i$, the number of parameters is $q + (q+1)^k$.

3 Genetic Algorithms

The number of parameters in a functional network model depends on the number of explanatory variables and the number of elements in the set of basic functions. In the additive model, the dimensionality of the problem grows linearly with the number of explanatory variables, but when the general additive model is considered, the dimensionality grows exponentially with the number of explanatory variables. Then, when the number of parameters is from moderate to high, an heuristic search method must be implemented for model selection. In this paper, a genetic algorithm is considered. Its purpose is to select the optimal subset of parameters to compose a model which provides a good approximation to (1).

A genetic algorithm starts with a random set of initial models. Each model is represented by a string of binary characters, called chromosomes. As an example, let us consider the model selection problem of an additive functional network with two explanatory variables

$$f(y) = h_1(x_1) + h_2(x_2),$$

where f, h_1 and h_2 are approximated by polynomials of degree 3. Each model is then represented by a chromosome of length 10. The chromosome $\mathbf{v} = [1111111111]$ represents the complete model, that is, all the terms are included in the model,

$$\alpha_1 y + \alpha_2 y^2 + \alpha_3 y^3 = \beta_0 + \beta_{11}x_1 + \beta_{12}x_1^2 + \beta_{13}x_1^3 + \beta_{21}x_2 + \beta_{22}x_2^2 + \beta_{23}x_2^3.$$

And, the chromosome $\mathbf{v} = [1001100100]$ represents the linear model, that is, only the linear terms in y, x_1 and x_2 are included in the model, $\alpha_1 y = \beta_0 + \beta_{11}x_1 + \beta_{21}x_2$.

Each model of the initial set is evaluated by the adjusted R-squared criterion (R_a^2), which penalizes models with a high number of parameters:

$$R_a^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2/(n-p)}{\sum_{i=1}^{n}(\hat{f}(y_i) - \overline{\hat{f}(y_i)})^2/(n-1)}, \quad (7)$$

where n is the sample size, p is the number of parameters in the model, e_i is the i-th residual (for example, for the additive model, $e_i = \hat{f}(y_i) - \hat{h}_1(x_{1i}) - \ldots - \hat{h}_k(x_{ki})$, where $\hat{f}$ and $\hat{h}_i$ are obtained as in (3)) and $\overline{\hat{f}(y)} = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(y_i)$.

Next, the best chromosomes of the initial population, in the sense of those with highest R_a^2, are selected. Then, a new population is obtained by applying genetic operations as crossover and mutation. This population is again evaluated and the process is repeated for some specified number of additional generations or when the evaluation function does not improve any more.

The size of the initial population and the probabilities of mutation and crossover have to be chosen by the user. See [7] for further details.

4 Assessing Performance Using Simulation

To assess the performance of the proposed method we use a set of simulated data from the model

$$Y^2 = X_1^2 + X_2^2 - X_1 * X_2 + \epsilon, \tag{8}$$

where X_1 and X_2 are independent $U[0,2]$, and ϵ is $N[0,0.1]$ and independent of X_1 and X_2.

We have estimated the model by applying the general additive functional network model, described in Section 2.2, with two explanatory variables. All the functions have been approximated by third degree polynomials.

The genetic algorithm described in Section 3, with populations of size 700, and crossover and mutation probabilities equal to 0.4, and 0.1, respectively, has been applied hundredfold. In this example, the models are represented by chromosomes of 19 bits. The chromosome representing the true model (8), including the constant term, is [0101010010100000000]. The mapping of this chromosome and the corresponding terms in the functional network model can be depicted as follows:

y	y^2	y^3	1	x_1	x_1^2	x_1^3	x_2	x_2^2	x_2^3
0	1	0	1	0	1	0	0	1	0
x_1x_2	$x_1x_2^2$	$x_1x_2^3$	$x_1^2x_2$	$x_1^2x_2^2$	$x_1^2x_2^3$	$x_1^3x_2$	$x_1^3x_2^2$	$x_1^3x_2^3$	
1	0	0	0	0	0	0	0	0	

Table 1 shows the terms included in the 17-th best selected models, ordered by the value of the evaluation function R_a^2. They are models whose R_a^2 is greater than 0.79. The constant term does not appear in the table, since it is always included in the model. Model number 5 is the true model with $R_a^2 = 0.8033$. The rest of the models have R_a^2 values very close to that. Note that all the models contain a number of terms less than 6 (including the constant term), far away from the maximum, 19. Moreover, most of them (10/17) have exactly 4 parameters, the same as the true model. We can define a simple measure of the complexity of the model by adding the powers of the terms included on it. It is shown in the last column of Table 1. With the help of this measure, we can choose the model with smallest complexity among those with highest R_a^2. In this case, this model is number 5, the true model.

Table 1: Selected Models to Approximate the Simulated Model in (8).

	y	x_1	x_2	Interactions	R^2	Complexity
1	y^2, y^3	x_1^2	x_2	$x_1^2x_2^3$	0.8124	13
2	y^3	x_1^3	x_2^2	$x_1^2x_2^2$	0.8121	12
3	y^3	x_1^3	x_2	$x_1^2x_2^3$	0.8111	12
4	y^3	x_1^3	x_2^2	$x_1^3x_2^3$	0.8039	11
5	y^2	x_1^2	x_2^2	x_1x_2	0.8033	8
6	y^2, y^3	x_1^2	x_2	$x_1^3x_2^3$	0.8030	14
7	y^2	x_1^2	x_2	$x_1^3x_2^3$	0.8017	11
8	y^2, y^3	x_1^2	x_2^2	$x_1x_2^2$	0.8008	12
9	y^2, y^3	x_1^3	x_2	$x_1^2x_2^2$	0.7998	13
10	y^2	x_1^2	x_2^2	$x_1^2x_2$	0.7971	9
11	y^3	x_1^3	x_2	$x_1^3x_2^2$	0.7970	12
12	y^3	x_1^3	x_2^2	$x_1x_2^2$	0.7965	11
13	y^2	x_1^2	x_2	$x_1^2x_2$	0.7945	8
14	y^2, y^3	x_1, x_1^2	x_2^2	$x_1^2x_2^3$	0.7924	15
15	y^2, y^3	x_1^2	x_2, x_2^2	$x_1^2x_2$	0.7922	13
16	y^3	x_1^3	x_2^3	$x_1x_2, x_1^2x_2$	0.7912	14
17	y^3	x_1	x_2	$x_1^2x_2^3$	0.7911	10

5 Assessing Performance Using Real data

Boston Housing data set contains 506 observations of 13 continuous variables and 1 binary valued variable related with housing values in suburbs of Boston. The purpose is to find the best fitting functional form and, in particular, to determine the pattern of the influence of air pollution on housing values as measured by x_5. The variables are:

y: Median value of owner-occupied homes in dollar 1000's,
x_1: Per capita crime rate by town,
x_2: Percentage of residential land zoned for lots over 25,000 sq.ft.,
x_3: Percentage of non-retail business acres per town,
x_4: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise),
x_5: Nitric oxides concentration (parts per 100 millions),
x_6: Average number of rooms per dwelling,
x_7: Percentage of owner-occupied units built prior to 1940,

x_8: Weighted distances to five Boston employment centres,
x_9: Index of accessibility to radial highways,
x_{10}: Full-value property-tax rate per dollar 10,000,
x_{11}: Pupil-teacher ratio by town,
x_{12}: $(Bk - 0.63)^2$ where Bk is the proportion of blacks by town,
x_{13}: Proportion of lower status of the population.

This data set was created by Harrison and Rubinfeld, [6], and it is analyzed in [1] and [2], among others. In [6] and [1] a linear model is proposed where y, x_8, x_9 and x_{13} are transformed by logarithms and x_5 and x_6 are squared. In [2] the ACE algorithm is applied to the transformed variables suggested in [6]. They conclude that the best model only need x_6, x_{10}, x_{11} and x_{13}, as predictors, with a milder transformation for y (different than logarithmic), a transformation for x_6 different from squared one and some transformation for x_{10}.

We apply an additive functional network model. Each function is approximated by third degree polynomials. The GA proposed in Section 3 is applied. The models are represented by chromosomes of 42 bits. Table 2 shows the 11 best models ($R_a^2 > 0.68$) obtained by repeating hundredfold the GA. All these models include a complete transformation in y and the constant term. Note that any model contain x_6, x_{10} or x_{12}. Most of the models suggest a complete transformation of x_1, x_2, x_3 and x_{13} and include just one term of x_5, x_7, x_9 and x_{11}. Note that x_5 appears in all the models squared or cubic, as it was found by Harrison and Rubinfeld.

Attending to the complexity measure introduced in Section 4, the best model is number 4 followed by number 9. Both give complete transformations of x_1 and x_{13}, do not transform x_7 and include x_5^2 and x_9^3. They just differ in the transformations suggested for x_3 and x_8.

6 Conclusions and Future Work

A genetic algorithm is presented as a powerful tool to select the terms involved in a functional network model. The GAs solve the computational problems which appear in model selection with a moderate to high number of parameters. The obtained models are simple and provide satisfactory approximations.

References

[1] Belsley, D. A., Kuh, E. and Welsch, R. E. *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, 1980.

[2] Breiman, L., and Friedman, J.H., "Estimating Optimal Transformations for Multiple Regression and Correlation (with discussions)," *Journal of the American Statistical Association*, V80, N391, pp. 580-619, 1985.

[3] Castillo, E., Cobo, A., Gutiérrez, J.M. and Pruneda, R.E., *An Introduction to Functional Networks with Applications*, Kluwer Academic Publishers: New York, 1998.

[4] Castillo, E., Hadi, A. S, Lacruz, B., "Optimal Transformations in Multiple Linear Regression Using Functional Networks," *Lecture Notes in Computer Science* V2084, Part I, pp. 316324.

[5] Castillo, E., Hadi, A. S, Lacruz, B. and Gutiérrez, J.M., "Some Applications of Functional Networks in Statistics and Engineering," *Technometrics*, V43, pp. 1024, 2001.

[6] Harrison, D. and Rubinfeld, D. L., "Hedonic Housing Prices and the Demand for Clean Air," *Journal of the Environmental Economics and Management*, N5 pp. 81-102, 1978.

[7] Michalewicz, Z., *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd Edition, Springer, 1999.

[8] UCI ML Repository Database. http://www.ics.uci.edu/∼mlearn/MLSummary.html.

Table 2: Selected Models GA.

	Terms in the selected model										R2	Complexity
1	x_1, x_1^3		x_3, x_3^2, x_3^3	x_4	x_5^3	x_7^3	x_8^2, x_8^3	x_9^2, x_9^3	x_{11}^3	x_{13}, x_{13}^3	0.7057	40
2	x_1, x_1^2	$x_2, x_2^2 x_2^3$	x_3, x_3^2, x_3^3	x_4	x_5^3	x_7^3	x_8^2, x_8^3	x_9^2, x_9^3	x_{11}^3	x_{13}, x_{13}^3	0.7002	45
3	x_1, x_1^2		x_3, x_3^3	x_4	x_5^2	x_7	x_8^3	x_9^3	x_{11}^2	$x_{13}, x_{13}^2, x_{13}^3$	0.6909	31
4	x_1, x_1^2		x_3, x_3^3	x_4	x_5^2	x_7	x_8^3	x_9^3	x_{11}^2	x_{13}^2, x_{13}^3	0.6908	30
5	x_1^2, x_1^3		x_3, x_3^2		x_5^3	x_7^3	x_8^2	x_9^3	x_{11}^3	x_{13}, x_{13}^3	0.6845	31
6	x_1, x_1^3	x_2^2, x_2^3	x_3^2, x_3^3		x_5^3	x_7	x_8^2, x_8^3	x_9^3	x_{11}^3	x_{13}^2, x_{13}^3	0.6835	40
7	x_1, x_1^3	x_2, x_2^3	x_3, x_3^2, x_3^3		x_5^3	x_7^2	x_8^3	x_9^3	x_{11}^3	x_{13}^2, x_{13}^3	0.6821	39
8	x_1^2, x_1^3	x_2^2	x_3, x_3^2, x_3^3		x_5^3		x_8	x_9^2	x_{11}^2	x_{13}^2, x_{13}^3	0.6819	32
9	x_1, x_1^2		x_3^3	x_4	x_5^2	x_7	x_8^2, x_8^3	x_9^3	x_{11}	x_{13}^2, x_{13}^3	0.6817	30
10	x_1, x_1^3	x_2^3	x_3^2, x_3^3		x_5^3	x_7, x_7^3	x_8^3	x_9	x_{11}^3	x_{13}^2, x_{13}^3	0.6817	37
11	x_1, x_1^2, x_1^3	x_2, x_2^3	x_3, x_3^2, x_3^3		x_5^2	x_7^3	x_8^3	x_9^2	x_{11}^3	x_{13}, x_{13}^3	0.6802	39